Deccan Education Society's
NMITD
Navinchandra Mehta Institute of Technology & Development
Estd. 1884
Approved by AICTE, DTE. & Affiliated to University of Mumbai, NAAC Accredited "B++" Grade

S.R.P. श्री Publication

TechPulse 2026

Editors:
Dr. Rasika Mallya, Dr. Swapnali Mahadik,
Dr. Ruchi Gupta



TechPulse 2026

5th Edition

Editors:

Dr. Rasika Mallya
Dr. Swapnali Mahadik
Dr. Ruchi Gupta





Publisher: S. R. P. Publication

Website: www.ijamrsd.com

Email: srppublication@gmail.com

Contact: 7387038026, 7249645344

Year: 2026

ISBN: 978-81-685854-0-9

Copyright©: Authors

All Rights Reserved

No part of this publication may be reproduced, transmitted or stored in a retrieval system, in any form or by any means, electronic, mechanical, photocopying recording or otherwise, without the prior permission of the author.

Disclaimer:

All the information and views expressed in the book are of contributors. The editor and publisher are not responsible for the contents and statements made or the opinions expressed by the authors. The publisher hereby disclaims any liability to information and statements made.

[PRINTED IN INDIA]

FROM THE DESK OF THE DIRECTOR



It gives me immense pleasure to present TechPulse'26, the annual ISBN publication of the MCA Department at DES's Navinchandra Mehta Institute of Technology for the academic year 2025–26.

TechPulse serves as a vibrant platform for aspiring IT professionals, researchers, and faculty members to explore, share, and disseminate knowledge in the domains of Technology, Computer Applications, and Computer Science. The publication reflects the department's commitment to fostering research aptitude, innovation, and lifelong learning among students.

This year's edition is inspired by the theme of **INCMAT-2026 – "New Horizons: Integrating AI, Business, and Sustainable Future."** As emerging technologies, particularly Artificial Intelligence, continue to reshape industries and societies, it is essential to develop solutionsthat are innovative, ethical, and sustainable. The articles featured in this volume showcase the intellectual curiosity, analytical thinking, and research capabilities of our students as they explore contemporary technological advancements and their impact on business and society.

The publication aligns with the vision of the MCA Department, "To ignite computing minds and nurture their full potential for contribution to the knowledge economy." At DES's NMITD, we continuouslystrive to create an ecosystem that encourages creativity, critical thinking, entrepreneurship, and academic excellence. We are committed to equipping our students with cutting-edge technological skills and research competencies that prepare them to excel in a dynamic global environment.

DES's NMITD is more than an institution; it is a family where the management, faculty, staff, students, parents, and alumniworktogether to nurture talent and build a culture of learning and innovation.

I am confident that the readers will find this publication informative, insightful, and inspiring. I congratulate all the student authors, faculty mentors, and members of the editorial team for their dedicated efforts in making **TechPulse'26** a meaningful repository of knowledge and research.

With my best wishes for continued learning, innovation, and success.

Dr. Rasika Mallya
Director In-Charge
DES's Navinchandra Mehta Institute of Technology

FROM THE EDITORIAL TEAM

We are pleased to present **TechPulse 2026**, the annual ISBN publication of the MCA Department, DES's Navinchandra Mehta Institute of Technology. This edition is inspired by the theme of **INCMAT-2026 – "New Horizons: Integrating AI, Business, and Sustainable Future."**

In today's rapidly evolving digital landscape, Artificial Intelligence is transforming industries, redefining business practices, and creating opportunities for sustainable growth. Through this edition of TechPulse 2026, we aim to encourage students to explore emerging technologies, innovative business solutions, and research-driven approaches that contribute to a smarter and more sustainable future.

TechPulse 2026 serves as a platform for nurturing research aptitude, technical writing skills, and intellectual curiosity among students. The magazine showcases research articles in the domains of Technology, Computer Applications, Computer Science, Artificial Intelligence, Data Analytics, Cyber Security, and other emerging fields. It reflects the creativity, analytical thinking, and academic excellence of our budding researchers.

We sincerely appreciate all the student authors, faculty mentors, reviewers, and editorial team members whose dedication and contributions have made this publication possible. Their commitment to knowledge creation and dissemination strengthens the culture ofresearch and innovation within our institution.

As we embrace new horizons in technology and management, may this publication inspire learners to think critically, innovate responsibly, and contribute meaningfully to society through research and lifelonglearning.

Dr. Rasika Mallya
Dr. Swapnali Mahadik
Dr. Ruchi Gupta

CONTENTS

Sr. No.	Title	Page No.
1	AN EMPIRICAL ANALYSIS OF SOCIAL ENGINEERING-BASED ONLINE FRAUD: PSYCHOLOGICAL MANIPULATION AND FINANCIAL BEHAVIOR PATTERNS **- Mr. Moresh Londe, Mr. Anant Kumbhar, Dr. Swapnali D. Mahadik**	1-9
2	OPTIMIZING PROCEDURAL CONTENT GENERATION: A COST VS. PERFORMANCE ANALYSIS IN GAME DEVELOPMENT **- Nishant Shaji, Yadnesh Teli**	10-17
3	HUMAN VS. SUPERINTELLIGENT AI: CAN LESS INTELLIGENT SYSTEMS CONTROL MORE INTELLIGENT ONES? A THEORETICAL AND EMPIRICAL INVESTIGATION **- Damneet Singh Randhawa, Kuldeep Singh Rathore**	18-28
4	SKILL GAP ANALYSIS IN THE IT SECTOR OF MUMBAI METROPOLITAN REGION: A LONGITUDINAL AND EMPIRICAL STUDY ACROSS THREE EPOCHS (2018–2026) **- Manas Surve, Aditya Tendulkar**	29-51
5	HUMAN-DRIVEN VS. AI-DRIVEN DEVELOPMENT IN EMBEDDED AND OPERATING SYSTEMS: IMPLICATIONS FOR RELIABILITY, SECURITY, AND MAINTAINABILITY **- Nishant Karpe, Saurabh Bhovad**	52-58
6	IMPACT OF ARTIFICIAL INTELLIGENCE ON EMPLOYMENT PATTERNS AND WORKFORCE TRANSFORMATION **- Bhushan Bhabal, Priya Gupta**	59-69
7	IMAGE-BASED AUTOMATED WASTE SORTING USING DEEP LEARNING: A CNN-BASED FIVE-STEP EXECUTION FRAMEWORK **- Rushikesh Bhosale, Adityaraj Ardalkar**	70-79

Sr. No.	Title	Page No.
8	A STUDY ON PRIVACY, SECURITY, AND TRUST IN CLOUD AI VS SELF HOSTED AI **- Deepak Denre, Aniket Kumar**	80-89
9	REGULATING THE INVISIBLE STATE: ARTIFICIAL INTELLIGENCE, PRIVACY, AND THE CRISIS OF DIGITAL GOVERNANCE — A COMPARATIVE ANALYTICAL STUDY ACROSS JURISDICTIONS **- Jyotirmay Mayank, Pankaj Mishra**	90-103
10	ALGORITHM-DRIVEN SHORT FORM CONTENT AND ITS EFFECT ON ATTENTION SPAN **- Atif Hanfi, Samina Hanfi**	104-115
11	ENHANCING CODE GENERATION WITH AI USING PROMPT ENGINEERING AND REINFORCEMENT LEARNING **- Gunal Boga, Prathamesh Davande**	116-125
12	PRODUCTIVITY BEHAVIOR ANALYSIS AMONG INTERNS & WORKING PROFESSIONALS **- Vedang Kanade, Prachit Gawand**	126-135
13	AN AI-DRIVEN FRAMEWORK FOR DRUG DE-ADDICTION AND VICTIM RECOVERY: ENHANCING REHABILITATION CENTERS AND POLICE SUPPORT SYSTEMS **- Sejal Gite, Mansi Bhosale**	136-144
14	QUANTIFYING COGNITIVE FATIGUE: LOCALIZED PREDICTIVE MODELING AND THE DOMINANCE OF PERCEIVED CONTROL IN SHORT-FORM VIDEO CONSUMPTION **- Saurav Gupta, Jatin Talavdekar**	145-153
15	PREDICTING THE IMPACT OF ARTIFICIAL INTELLIGENCE USAGE ON PROBLEM-SOLVING AND CRITICAL THINKING SKILLS AMONG COLLEGE STUDENTS USING MACHINE LEARNING **- Mr. Dhruv Panchal, Ms. Shramika Patane, Dr. Ruchi Gupta**	154-161

Sr. No.	Title	Page No.
16	EXPLAINABLE AI-BASED INTRUSION DETECTION SYSTEM FOR NETWORK TRAFFIC CLASSIFICATION **- Mr. Harsh Kathe, Mr. Bilal Shaikh**	162-168
17	A REAL-TIME VOICEPRINT AUTHENTICATION SYSTEM WITH DEEP LEARNING-BASED SPOOF DETECTION FOR SECURE VAULT ACCESS **- Aayushi Patel, Pradnya Ahire**	169-178
18	DESIGN AND EVALUATION OF A RETRIEVAL AUGMENTED KNOWLEDGE REPOSITORY FOR MULTI-FORMAT DATA SOURCES **- Chaitanya Shinde, Bilal Shaikh**	179-202
19	TELECURA: A UNIFIED TELEMEDICINE PLATFORM WITH CREDIT-BASED MONETIZATION, ADMINISTRATIVE GOVERNANCE, AND WEBRTC-ENABLED VIDEO CONSULTATIONS **- Abhishek Mane, Prof. Prajakta Kambale**	203-222
20	REAL-TIME MEDIA AUTHENTICITY CHAIN: AI-DRIVEN DEEPFAKE AND PROVENANCE VERIFICATION SYSTEM (MEDIA-TRUST) **- Itisha Vamja**	223-236
21	AN INDUSTRY 4.0–ENABLED INTEGRATED AUTOMATION FRAMEWORK FOR SCALABLE AND TRACEABLE UAV MANUFACTURING **- Pranav Gadewar**	237-242
22	ROLE SPECIALISATION IN MULTI-AGENT CODE GENERATION **- Master Vyom Margaj, Dr. Sampada Margaj**	243-250
23	ANALYZING COPING STRATEGIES FOR STRESS MANAGEMENT USING SENTIMENT ANALYSIS AND MACHINE LEARNING **- Mrs. Eshwari Shirke, Dr. Prabha Kadam**	251-256
24	AI-DRIVEN SEMANTIC SEARCH IN MYSQL USING VECTOR EMBEDDINGS AND LARAVEL PRISM **- Prachi Barve, Dr. Mrs. Pratibha Adkar**	257-264

Sr. No.	Title	Page No.
25	SCROLL SAFE: BUILDING HYBRID AI SYSTEMS FOR CONTENT MODERATION ON SOCIAL MEDIA PLATFORMS **- Prathamesh A. Kamble, Dr. Prabha Kadam**	265-273
26	AI RECOMMENDATIONS VS HUMAN CHOICE — ARE PEOPLE ACTUALLY DECIDING ANYTHING ANYMORE? **- Ms. Umera Khan, Mr. Rushikesh Kokare**	274-280
27	LEAF-BASED MEDICINAL PLANT IDENTIFICATION USING MOBILENETV2 AND CNN **- Pooja Satish Patil, Dr. Sampada Margaj**	281-290
28	UNDERSTANDING HOW COLLEGE STUDENTS INTERACT WITH AI TOOLS AND THEIR COGNITIVE BEHAVIOUR: AN EXPLORATORY DATA ANALYSIS USING TABLEAU **- Mr. Anil Prajapati, Mr. Swapnil Vichare, Mrs. Lavina Mistry**	291-301
29	EVOLUTION OF USER SENTIMENT AND LINGUISTIC BARRIERS IN CONVERSATIONAL AI: AN NLP PERSPECTIVE **- Ritika Sahu, Sakshi Shinde**	302-308
30	CYBER-BULLYING: AI-BASED CYBERBULLYING DETECTION SYSTEM FOR SOCIAL MEDIA PLATFORMS **- Sakshi Lahane**	309-331
31	A COMPARATIVE ANALYSIS OF MOBILE-BASED DIGITAL MEDICAL REPORT SYSTEMS WITH FOCUS ON DATA SECURITY AND ACCESSIBILITY **- Abdullah Falke, Vedika Kamble**	332-337
32	DECENTRALISED STORAGE SYSTEM USING IDLE DEVICE SPACE **- Tejal Rane, Prathamesh Nivdekar**	338-353

Sr. No.	Title	Page No.
33	TRANSFORMING HUMAN-MACHINE INTERACTION (HMI) THROUGH LARGE LANGUAGE MODELS (LLMS): OPPORTUNITIES AND CHALLENGES **- Soham Nanasaheb Naiknaware, Yogesh Kiran Rathod**	354-361
34	NEGATIVE IMPACTS OF ARTIFICIAL INTELLIGENCE: EXAMINING THE THREATS FACED BY INDIVIDUALS AND SOCIETY **- Shruti Kamble, Adifa Ansari**	362-369

AN EMPIRICAL ANALYSIS OF SOCIAL ENGINEERING-BASED ONLINE FRAUD: PSYCHOLOGICAL MANIPULATION AND FINANCIAL BEHAVIOR PATTERNS

Mr. Moresh Londe
MCA Student, DES's NMITD, Dadar.
Email: londemoresh03@gmail.com

Mr. Anant Kumbhar
MCA Student, DES's NMITD, Dadar.
Email: anantkumbhar5835@gmail.com

Dr. Swapnali D. Mahadik
Assistant Professor, MCA Department, DES's NMITD, Dadar.
Email: swapn.mahadik30@gmail.com

Abstract
Because communication channels and financial platforms online are being used more often, the rate at which social engineering-based online fraud has become rampant. These acts rely more on psychological manipulation than technical exploits using means such as job scams based on tasks, OTP scams, and identity theft among others. An investigation examining the prevalence of social engineering fraud has been carried out, specifically focusing on psychological manipulation strategies and financial transactions performed by internet users. This study employs a descriptive and empirical research design, collecting information from 66 participants using a questionnaire. The data was examined using descriptive statistics, which included frequencies and percentages as well as crosstabs. The findings show that fraudsters primarily use social messaging apps WhatsApp and Telegram to initiate their fraudulent activities. Psychological aspects like urgency, greed, and authority play significant roles in eliciting actions from the users. Moreover, fast multi-level money transfer is common to evade any tracking of the financial transaction. The paper describes vulnerability behavior patterns and suggests a fraud lifecycle framework. However, given the convenience sampling technique, self-reporting of experiences and no causality testing, there are limitations in the Outcome of this Research.
Keywords: Social Engineering, Online Fraud, Psychological Manipulation, OTP Scams, Financial Transactions.

Introduction

The advent of digital technology has transformed the manner in which communication, business, and finance is conducted (Kumar & Sharma, 2025). The growth of digital payment platforms and remote working options has given way to online fraud. While earlier forms of cyberattacks involved the exploitation of technical weaknesses in systems, contemporary fraud operations have shifted towards social engineering.

Social engineering fraud consists of various types of activities, including home-based task scams, OTP fraud, phishing, and impersonation (Hadnagy, 2018). These fraudulent acts take advantage of psychological incentives like urgency, trust, fear, and promise of reward to manipulate the victims' decisions.

Recently, there has been an observable increase in the instances of cyber fraud and associated financial damage (Government of India, 2024; Ministry of Home Affairs, 2024) as reported by agencies like the National Crime Records Bureau and Reserve Bank of India.

 ISBN: 978-81-685854-0-9

Despite the increased alertness, users continue to fall prey to these frauds. It brings into question:

- the nature of the frauds
- the psychology behind victims' actions
- The way in which these financial transactions are executed and hidden

This paper attempts to provide insight into the above factors based on empirical observations obtained through internet users.

Literature Review

The concept of social engineering has been widely investigated within the area of cybersecurity (Hadnagy, 2018). Social engineering is defined as the application of psychological methods aimed at influencing people to perform certain practices that are likely to be detrimental to security. As compared to cyber attacks, social engineering depends on human interaction and takes advantage of behavioral patterns.

Earlier research demonstrated that social engineering attacks leverage psychological biases such as authority bias when individuals are more likely to follow instructions provided by authorities, or scarcity bias (Hadnagy, 2018) when the sense of urgency prompts people to act immediately without verifying information. Additionally, earlier studies proved that emotional triggers are highly influential in making decisions within the context of fraud situations.

In terms of phishing attacks, studies demonstrated that users fail to detect phishing emails because of their similarity to regular emails. The usage of professional language, corporate logos, and formats increases the likelihood of perceiving the message as legitimate.

Research on online fraud in emerging economies shows that the rapid development of technologies coupled with insufficient digital literacy levels contribute to the emergence of online fraud (Patel & Desai, 2025). People are unaware of security measures, Thus, they are prone to phishing attacks.

Studies on financial fraud revealed that fraudsters apply sophisticated transaction patterns. For instance, they create mule accounts, conduct layered transactions, and transfer funds quickly between various accounts to cover traces and evade detection (Anderson et al., 2013).

Even though the issue is widely explored, some gaps still exist.

- psychological and financial aspects of fraud are not considered comprehensively.
- There's not enough real-world studies relying on user behavior.
- There is currently no structured framework for the fraud life cycle.

This paper will cover all aforementioned research gaps and will integrate behavioral and financial aspects of fraud.

Research Methodology

Research Design

For this research project, the method of study selected includes the usage of a quantitative descriptive research design in an attempt to determine social engineering-based online fraud trends and the link between user behavior and the same. This research design uses the cross-sectional methodology, which involves the collection of data at one point in time to capture existing trends, experience, and perceptions concerning online fraud.

The main goal of this research project is to explore behavioral trends that can be observed regarding exposure to fraud, psychological manipulation, and financial transactions.

Research Approach and Data Collection Method

An exploratory research based on survey was conducted to acquire primary information. The data was acquired using the questionnaire, which was created to integrate the main components of the social engineering attacks and user behavior towards these attacks.

ISBN: 978-81-685854-0-9

Google Forms was used to conduct the online survey. It comprised close-ended questions both multiple choice and those based on a scale of attitudes.
The questionnaire included the following parts:

- Fraudulent message exposure
- Platform for communication with fraudulent persons
- Psychological response to a fraud attempt
- Financial behavior
- Awareness and Reporting

The information acquired from the questionnaire was stored in Excel table automatically.

Population and Sampling

For this study, the population that the research aims to reach is that of internet users whose activities focus around using online resources such electronic payments, social networking sites, and instant messaging; thus, they are prone to be victims of social engineering.
There were 66 responses in total used as input in the research using a convenience sample technique wherein the chosen respondents were readily accessible and available.
All other responses other than those who are incomplete or not directly related to the survey will be excluded. While this technique makes data gathering easier, generalization is restricted.

The Study's Variables

The study includes both independent and dependent variables to analyze behavioral patterns.
Independent Variables:
• Exposure to fraud messages
• Platform of communication (WhatsApp, Telegram, SMS, etc.)
• Psychological triggers (Urgency, Greed, Authority, Fear, etc.)
• Level of awareness of online fraud
Dependent Variables:
• Financial behavior (transferred money, did not transfer money, etc.)
• Vulnerability to fraud
• Response behavior (verified, ignored, complied, etc.)
• Reporting behavior
These are all self-reported behavioral indicators rather than the actual measurement of cognitive or financial behavior.

Data Collection Procedure

Obtained data involved completion of the questionnaire in a certain time span through online survey method. Participants participated in this survey voluntarily, and they were briefed about the objective of conducting the study prior to giving their responses.
To protect participant confidentiality and anonymity, no personal identifiable information was recorded. Participants were asked to give their honest response about their experience and perception.
Data obtained through questionnaire was saved in a tabular form from Google Forms for further analysis.

Data Analysis

Descriptive analysis of the data collected is carried out using frequency distribution, percentage and cross-tabulation. These are the analysis techniques applied in this research. These statistical tools will help us in summarizing our data and making sense out of it.
Results will be interpreted for identifying patterns, trends, and relationships between independent and dependent variables. Patterns of dominance will be found in terms of the

 ISBN: 978-81-685854-0-9

platform most frequently used to commit social engineering fraud. Furthermore, the psychological triggers will also be determined. Patterns of financial transactions and responses towards these fraudulent activities will also be identified.

Validity and Ethical Considerations

The questionnaire was created using the objective of studying the effect of social engineering on user behavior. Efforts have been made to develop concise and coherent questions to avoid misinterpretation.

The research procedure has always been conducted with ethical considerations in mind. Participation in this survey is completely voluntary and no participant was pressurized to participate in the study. All participants had freedom of choice about answering questions and no personal information was asked from them to protect their confidentiality and privacy.

Results and Research Findings

- **Digital Payment Usage**

Figure 1 illustrates the primary digital payment methods used by respondents. A dominant 98.5% of respondents reported using UPI platforms such as Google Pay, PhonePe, and Paytm, while only a negligible proportion uses cards or net banking.

This highlights the critical role of UPI in digital transactions and also explains why fraudsters prefer targeting UPI-based systems due to their speed and ease of transfer.

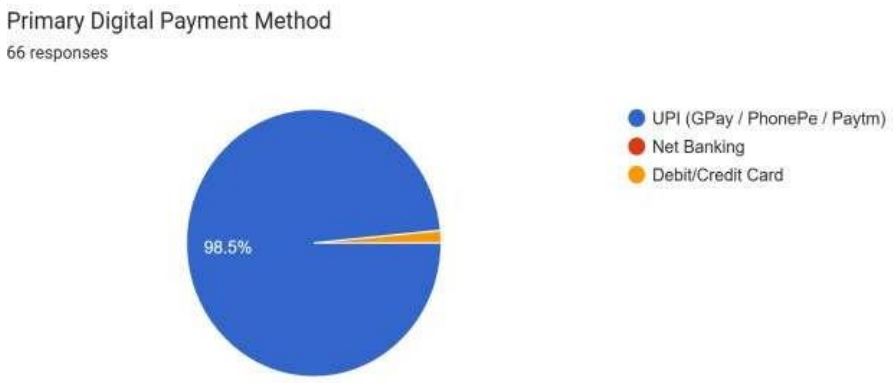


Figure 1 Primary Digital Payment Method

- **Exposure to Work-from-Home Scam Messages**

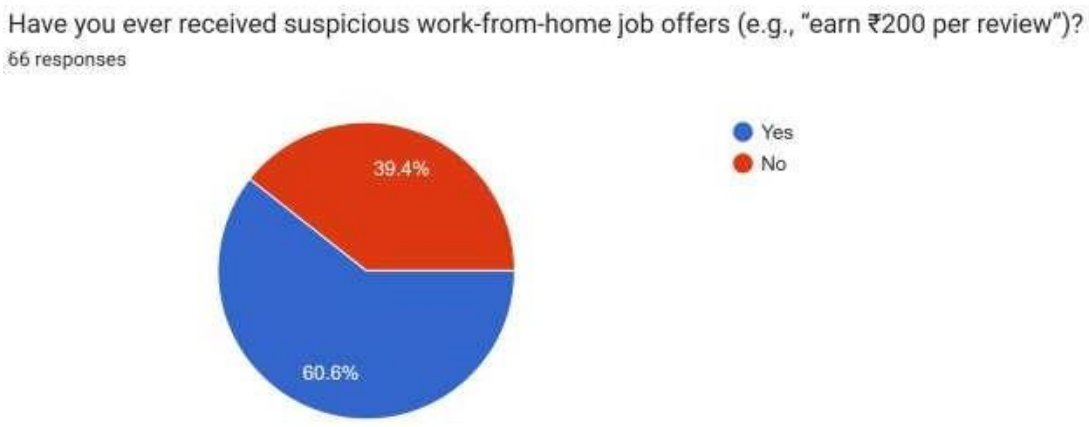


Figure 2 shows that 60.6% of respondents have received suspicious work-from-home job offers, while 39.4% have not.
This indicates that a majority of users have been exposed to task-based scams, confirming the widespread nature of such frauds.

- **Platforms Used for Scam Communication**

Figure 3 reveals that WhatsApp is the most common platform for receiving scam messages (54.5%), followed by SMS (45.5%) and phone calls (28.8%). Telegram (27.3%) and email (25.8%) also contribute significantly, while Instagram accounts for a smaller share (15.2%).
This clearly indicates that messaging platforms dominate as primary attack vectors, allowing fraudsters to reach a large number of users quickly.



Figure 3 Scam Communication Channels

- **User Response to Task-Based Offers**

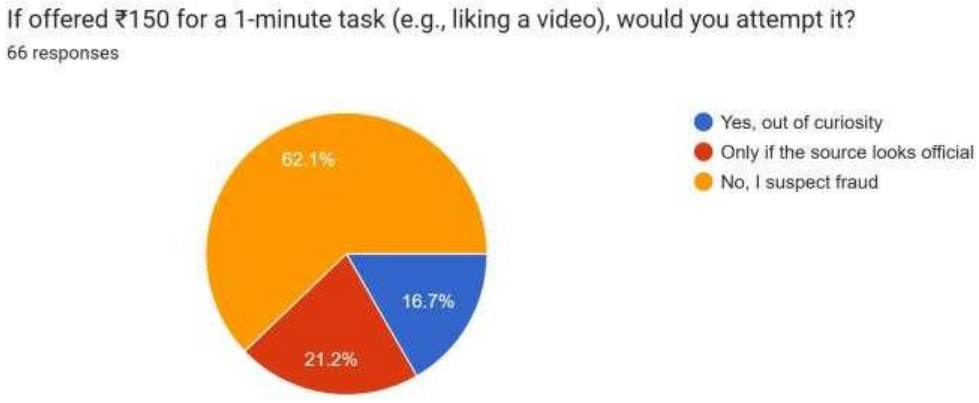


Figure 4 Dependency Reaction to Small Task Offers

Figure 4 shows that 62.1% of respondents refused to attempt suspicious tasks, indicating awareness of fraud risks. However, 16.7% expressed curiosity, and 21.2% stated they would engage if the source appeared legitimate.
This suggests that while consciousness is present, a significant portion of users remains vulnerable under perceived credibility.

 ISBN: 978-81-685854-0-9

- **Interaction with Fraud Message**

Figure 5 indicates that 34.8% of respondents have interacted with such messages, while 65.2% have not.
This highlights that although a majority avoid engagement, a considerable proportion still interacts with scam messages, increasing the risk of fraud.

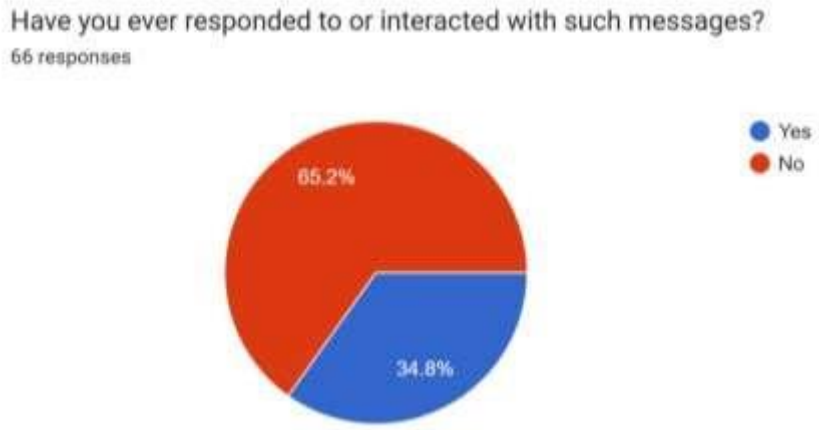


Figure 5 Engagement with Scam Messages

- **OTP Handling Behavior**

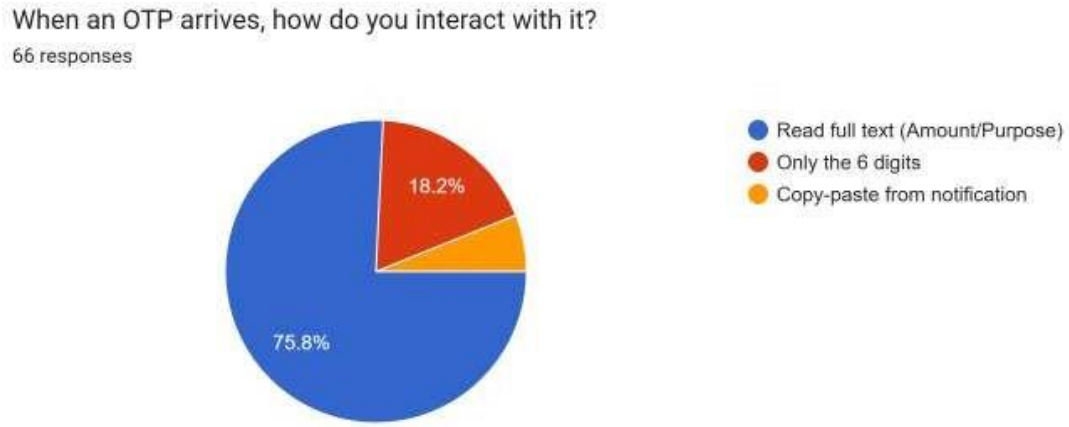


Figure 6 OTP Interaction Patterns

Figure 6 shows that 75.8% of respondents carefully read OTP details before use, while 18.2% focus only on the code, and a small percentage directly copy-paste.
This indicates relatively safe practices among users, but also highlights a minority that could be vulnerable to OTP-based fraud.

- **Financial Loss Experience**

Figure 7 reveals that 54.5% of respondents have either personally experienced or know someone who has suffered financial loss as a result of internet fraud.
This suggests that financial fraud affects a significant portion of the population and is both widespread and serious.

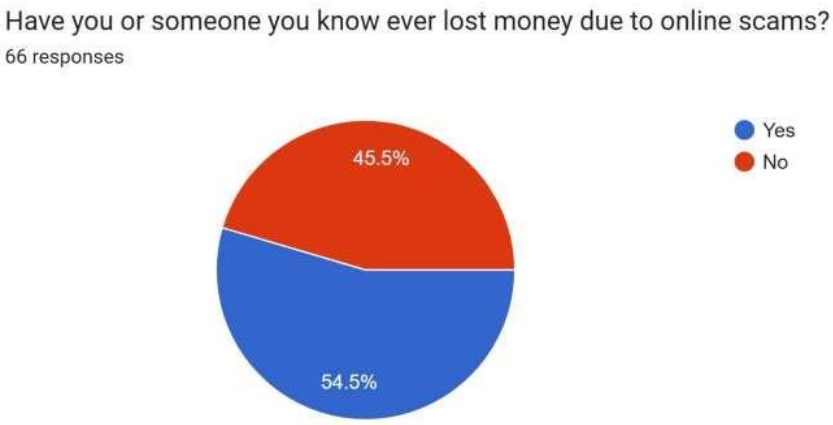
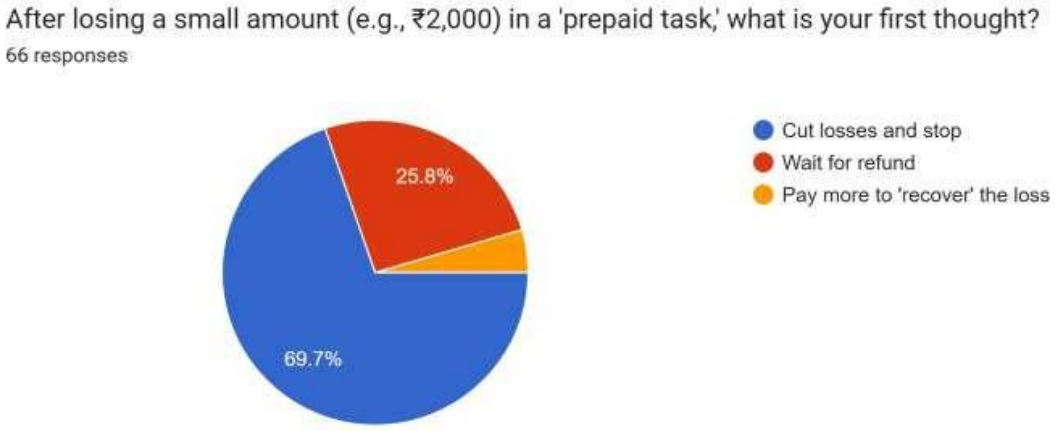


Figure 7 Experience of Financial Loss

- **Psychological Response After Loss**

Figure 8 shows that 69.7% of respondents prefer to cut losses and stop, while 25.8% wait for refunds, and a small percentage consider paying more to recover losses.
This suggests that majority of users act rationally, a minority may fall into further exploitation due to psychological pressure



.

Figure 8 Reaction After Financial Loss

- **Reasons for Not Reporting Fraud**

Figure 9 indicates that 59.1% of respondents do not report fraud due to lack of trust in recovery, followed by social stigma (27.3%) and small loss amount (13.6%).
This suggests that psychological and systemic barriers discourage reporting behavior.

 ISBN: 978-81-685854-0-9

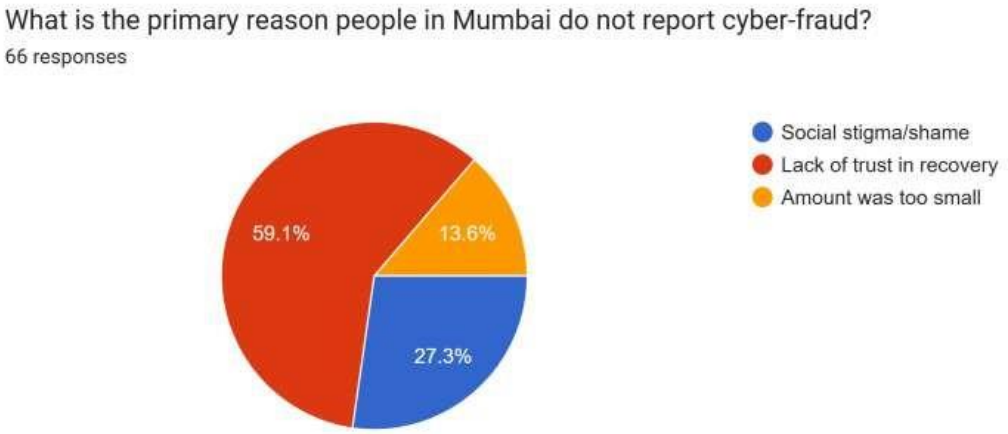


Figure 9 Barriers to Reporting

The results suggest that social engineering-driven online scams are prevalent among internet users especially in the age range of 18-25 years. The majority of participants stated that they have encountered phishing messages from job offers, where they have received the scams via WhatsApp, SMS, and Telegram applications. Urgency, greed, and authority are considered psychological elements that greatly impact user's actions (Hadnagy, 2018).

Most users displayed their awareness by rejecting any suspicious messages while some would still try to follow up despite the legitimacy of the source of messages. UPI was found to be widely used, hence, making it a primary target for fraud (Kumar & Sharma, 2025). More over 50% of those surveyed said they had either lost money themselves or had seen someone else fall subject to scams.

Even though the survey revealed higher awareness of users on how to report scam activities, they don't actually perform this action.

Discussion

The findings allow for the conclusion that social engineering involves a methodological procedure that includes psychological and economic manipulations. The messaging application serves as a point of entry into social engineering, while emotional triggers affect users' actions. It was discovered that being aware of social engineering does not protect a person from its negative impacts. Behavioral aspects related to social engineering involve trustworthiness, urgency, and emotions.

The behavior of financial transactions helps identify methods used by attackers to reduce traceability and obtain money.

Conclusion

This study provides a comprehensive examination of social engineering based online fraud by analyzing patterns of user exposure, psychological manipulation, and financial transaction behavior. The findings indicate that online fraud is widely prevalent among internet users, with messaging platforms such as WhatsApp and Telegram serving as the primary channels for initiating fraudulent communication.

The analysis highlights that psychological manipulation has a major role in fraud execution. Emotional triggers such as urgency, greed, authority, and fear significantly influence user decision-making, often leading individuals to act without proper verification. Despite a considerable level of awareness regarding online fraud, the study reveals a gap between knowledge and actual behavior, where users continue to respond to suspicious messages under psychological pressure.

ISBN: 978-81-685854-0-9

Furthermore, the study identifies consistent financial patterns, including the use of UPI and repeated transaction requests, which facilitate rapid and often irreversible financial loss. These findings emphasize that fraud is not only a technological issue but also a behavioral and cognitive challenge (Hadnagy, 2018; Anderson et al., 2013).

The proposed Social Engineering Online Fraud Lifecycle Model (SE-OFLM) offers a structured framework to understand the sequential stages of fraud execution. While the study is limited by its descriptive design and sample scope, It offers insightful information on user vulnerability and fraud mechanisms, contributing to the development of awareness strategies and future research in cybersecurity.

References

1. Kumar, R., & Sharma, P. (2025). Analysis of cyber fraud and digital financial crimes in India. International Journal of Law Research. Retrieved from https://ijlr.iledu.in/wp-content/uploads/2025/04/V5I685.pdf
2. Singh, A., & Mehta, R. (2024). The rising threat of recruitment frauds: Analysing the role of HRM practices in mitigating cybercrime in India. ResearchGate Publication.
3. Patel, V., & Desai, K. (2025). A study on cyber fraud trends and prevention mechanisms. International Journal of Creative Research Thoughts. Retrieved from https://www.ijcrt.org/papers/IJCRT2509604.pdf
4. Government of India. (2024). Cybercrime awareness and prevention measures. Press Information Bureau. Retrieved from https://www.pib.gov.in/PressNoteDetails.aspx
5. Ministry of Home Affairs. (2024). Job fraud awareness brochure. Indian Cyber Crime Coordination Centre (I4C). Retrieved from https://cybercrime.gov.in/pdf/Job%20Fraud%20Brochure%20Final.pdf
6. Reddy, S. (2023). The new generation of crimes: Virtual attacks overtake physical crimes. ResearchGate Publication.
7. Gupta, N., & Shah, D. (2025). Emerging trends in online fraud and cybercrime behavior. International Journal of Research Publication and Reviews. Retrieved from https://ijrpr.com/uploads/V6ISSUE3/IJRPR40456.pdf
8. Kumar, S. (2014). SMS-based one-time password vulnerabilities and safeguarding OTP over network. International Journal of Engineering Research & Technology. Retrieved from https://www.ijert.org/research/sms-based-one-time-password-vulnerabilities-and-safeguarding-otp-over-network-IJERTV3IS051538.pdf
9. Anderson, R., Barton, C., Böhme, R., Clayton, R., van Eeten, M., Levi, M., Moore, T., & Savage, S. (2013). Measuring the cost of cybercrime. The Economics of Information Security and Privacy, 265 - 300.
10. Hadnagy, C. (2018). Social engineering: The science of human hacking (2nd ed.). Wiley Publishing.

 ISBN: 978-81-685854-0-9

OPTIMIZING PROCEDURAL CONTENT GENERATION: A COST VS. PERFORMANCE ANALYSIS IN GAME DEVELOPMENT

Nishant Shaji
Research Scholar, MCA, DES's NMITD, Mumbai University, India.
Email: shajinishant11@gmail.com

Yadnesh Teli
Research Scholar, MCA, DES's NMITD, Mumbai University, India.
Email: yadnesht909@gmail.com

Abstract
Procedural Content Generation (PCG) is a common tool in modern game development that uses algorithms to generate game assets, environments, and mechanics, thereby minimizing the need for manual design. This paper undertakes a systematic comparative analysis of five fundamental families of procedural content generation (PCG) algorithms: noise-based terrain generation, space partitioning, grammar-based construction, search-based optimization, and machine-learning-driven synthesis. The study evaluates their performance parameters, content quality, and practical implementation trade-offs within an integrated analytical framework. This study integrates performance data from 21 primary sources to facilitate a straightforward cross-family comparison across six parameters: computational efficiency, content quality, controllability, scalability, implementation complexity, and real-time feasibility, in comparison to the prior surveys that evaluated algorithm categories in isolation. The research results indicate that no particular algorithm family performs well throughout all dimensions and that reliable manufacturing PCG systems generally depend on hybrid, multi-tier architectures. This comparison ultimately leads the paper to propose specific optimization principles and a decision-making framework to help professionals select and combine PCG methods.
Keywords: Procedural Content Generation, Game AI, Computational Efficiency, Algorithm Optimization, Content Scalability, Noise Functions, Machine Learning, Wave Function Collapse, and Game Engines.

1. Introduction

The modern video game development environment is all about making games with features that are rich, big, and extraordinary. Players today want immense open worlds, a wide variety of enemies, a variety of story branches, and nearly infinite replay potential. For most studios, fulfilling of such requirements through entirely manual design processes has become increasingly time-consuming and costly to manage. Procedural Content Generation (PCG) is the algorithmic creation of game elements with minimal assistance from people. It has become one of the most fundamental and significant breakthroughs across traditional game development engineering.

PCG incorporates almost every part of making a game, such as generating terrain, building dungeons, putting together narrative-based systems, crafting items, designing challenges, and even gameplay physics. The history of PCG in games traces its roots to Rogue (1980), which at first employed the technique of random dungeon generation procedures for the development of the roguelike genre. Elite (1984) procedurally generated an entire galaxy within standard hardware limits. Moreover, Minecraft (2009) and No Man's Sky (2016) showed how PCG can create worlds that are essentially infinitely scalable as well as planet-sized.

PCG has both pros and cons, but it also has some issues. If systems are not set up right, they can make content that seems boring, repetitive, does not make sense when played, or is not playable. In real-time environments, some algorithms need a lot more computing power than is normally available. This paper has been motivated by the necessity for a thorough analysis of the cost-performance discrepancies that are essential in relation to contemporary PCG methodologies.

1.1 Research Objectives

This study will: (1) identify and classify all major families of principal PCG algorithms; (2) develop a synthesized view of their performance efficiencies by comparing previously published benchmarks; (3) measure content quality metrics, including diversity, coherence, and playability; (4) analyze how different game engines affect the overall performance of the PCG pipeline; (5) begin to investigate emerging AI-driven approaches to PCG; and (6) provide a practical optimization guide based on evidence as well as a decision-making process for practitioners.

1.2 Research Gap

Earlier research has evaluated PCG algorithms in isolation, within a specific category, frequently employing diverse metrics and/or hardware. No research has yet conducted a comprehensive comparison of all significant components of PCG algorithms as a collective entity. It appears that there is presently no cross-category comparative research examining performance efficiency, content quality, controllability, scalability, and real-time feasibility. This study addresses this deficiency by performing a thorough comparative structured evaluation.

2. Literature Review

The academic study of PCG has grown substantially over two decades, producing a body of work spanning algorithm design, player experience research, and computational analysis.

2.1 Foundational Works

Shaker, Togelius, and Nelson (2016) offered the most thorough examination of the field in *Procedural Content Generation in Games*, classifying techniques based on randomness level, designer control, generation runtime, and content type (Shaker et al., 2016). Togelius et al. (2011) established the differentiation between constructive methods—characterized by single-pass generation without backtracking—and search-based methods, which progressively enhance content to meet quality objectives (Togelius et al., 2011). Hendrikx et al. (2013) classified PCG into six hierarchical content categories: game bits, game space, game systems, game scenarios, game design, and derived content (Hendrikx et al., 2013).

2.2 Noise-Based Terrain

Perlin (1985) created coherent noise functions for terrain synthesis, which are still used widely today (Perlin, 1985). Simplex Noise (Perlin, 2002) enhanced the original by removing directional artifacts while reducing computational expenses (Perlin, 2002). Musgrave et al. (1989) illustrated fractal noise for photorealistic terrain (Musgrave et al., 1989). Cordonnier et al. (2016) incorporated physically-based erosion simulation into procedural pipelines to generate more geologically realistic terrain (Cordonnier et al., 2016).

2.3 Grammar-Based Generation

L-systems, formalized by Prusinkiewicz and Lindenmayer (1990) from Lindenmayer's botanical modeling work, have been utilized in the design of building facades, road networks, and quest structures (Prusinkiewicz & Lindenmayer, 1990). Shape grammars, as shown by Muller et al. (2006), made it possible to create architectural environments that were consistent with procedures (Muller et al., 2006). Wave Function Collapse (Gumin, 2016) produces outputs that are locally consistent with example patterns, leading to its widespread use in tile - based environments (Gumin, 2016). Karth and Smith (2017) examined the strengths of WFC

 ISBN: 978-81-685854-0-9

and pinpointed scalability constraints for extensive generation environments (Karth & Smith, 2017).

2.4 Search-Based and ML Approaches

Togelius et al. (2011) conducted a survey on search-based procedural content generation, focusing on the application of evolutionary algorithms to level and rule generation (Togelius et al., 2011). Summerville et al. (2018) conducted a thorough examination of machine learning techniques in procedural content generation (PCG), including Markov models, recurrent neural networks (RNNs), autoencoders, and generative adversarial networks (GANs) (Summerville et al., 2018). Volz et al. (2018) showed how GAN latent variables change over time to make levels for Super Mario Bros. (Volz et al., 2018). Mouret and Clune (2015) created MAP-Elites, which are used to make high-quality, varied content archives (Mouret & Clune, 2015).

2.5 Performance Studies

Hendrikx et al. (2013) observed the deficiency of quantitative performance benchmarks for PCG (Hendrikx et al., 2013). Amato (2017) analyzed the costs of dungeon generation algorithms, discovering significant discrepancies among various methodologies (Amato, 2017). Van der Linden et al. (2013) demonstrated that constraint-based platformer level generation could attain near-real-time performance through meticulous optimization (Van der Linden et al., 2013). These studies furnish the principal source data amalgamated in this comparative analysis.

3. Methodology

This study utilizes a systematic comparative review methodology, integrating performance data and qualitative evaluations from published literature, technical documentation, and production case studies. This method follows the standard way of doing computing surveys (Kitchenham, 2004).

3.1 Source Selection

Sources were chosen based on: (1) being published in a peer-reviewed journal or official engine documentation; (2) giving quantitative performance data or structured qualitative assessments; (3) having enough methodological detail to check the validity of the measurements; and (4) being relevant to at least one target algorithm family. Using these criteria, 21 main sources from 1985 to 2023 were identified.

3.2 Algorithm Families

Five families were chosen to represent the coverage: (1) Noise-based terrain generation: Perlin/Simplex/Value Noise, Fractal Brownian Motion; (2) Space partitioning: BSP, Cellular Automata, Voronoi; (3) Grammar-based: L-systems, Shape Grammars, WFC; (4) Search-based: Genetic Algorithms, MCTS, Constraint Solving; (5) Machine learning: GANs, VAEs, Transformers, Markov Chains.

3.3 Evaluation Dimensions

Six different areas were examined: computational efficiency (generation time, memory, and scalability); content quality (diversity, coherence, and playability); controllability (support for designer constraints); real-time viability (feasibility within 16 ms and 100 ms budgets); implementation complexity; and data/training needs for ML approaches.

3.4 Data Normalization

Performance numbers from different hardware environments were normalized to a common reference point (Intel Core i9-class CPU, NVIDIA RTX 3000-series GPU, 32 GB RAM). Qualitative dimension scores (1–5 ordinal scale) are given based on the strength of evidence in the source corpus. All synthesized figures are clearly linked to the source literature.

4. Algorithm Analysis and Results

4.1 Noise-Based Terrain Generation

Noise-based methods offer excellent performance-quality balance with O(n) time complexity. Table I presents the synthesized generation time data normalized to the reference hardware configuration.

Table 1. Noise-Based Terrain Generation Performance

Algorithm	**256×256 (ms)**	**1024×1024 (ms)**	**Peak RAM**
Perlin Noise	~2	~34	~12 MB
Simplex Noise	~1.4	~23	~10 MB
Value Noise	~0.9	~15	~8 MB
Fractal BM (8 Oct)	~10	~163	~14 MB
Domain Warping	~19	~300	~18 MB

Simplex Noise is the best choice for production because it has better visual coherence than Perlin Noise and costs about 33% less to compute (Perlin, 2002). Value Noise is the fastest, but it has grid artifacts. Fractal Brownian Motion makes the terrain more interesting, but it costs more per octave. Domain warping works best with pre-made content or GPU compute shader pipelines, where speeds can be 15 to 40 times faster.

4.2 Space Partitioning Methods

Space partitioning algorithms form the backbone of dungeon generation across many genres. Table II synthesizes data from Van der Linden et al. (2013) and Amato (2017).

Table 2. Space Partitioning Comparison

Method	**Complexity**	**Gen. Time**	**Playability**
BSP Dungeon	O(n log n)	~4 ms	~98%
Cellular Automata	O(n × k)	~8 ms	~87%
Voronoi Seg.	O(n log n)	~11 ms	~92%
Room-Corridor	O(n)	~2 ms	~99%
Drunkard Walk	O(n × t)	~15 ms	~74%

BSP generation makes sure that the space is connected by using hierarchical partitioning. This makes the game more playable, but it also makes it less diverse. Cellular Automata make cave-like environments that are very diverse, but they can also have areas that are not connected. The Drunkard's Walk algorithm needs mandatory post-processing connectivity verification—an extra step that is often missed when comparing algorithms but is very important in production systems.

4.3 Grammar-Based Generation

Grammar-based approaches excel in coherence and controllability. WFC emerged as the most versatile, balancing coherence and diversity effectively. Table III presents a qualitative comparison synthesized from Prusinkiewicz and Lindenmayer (1990), Muller et al. (2006), and Karth and Smith (2017).

Table 3. Grammar-Based Generation Characteristics

Approach	**Controllability**	**Diversity**	**Speed**
L-Systems	High	Medium	Fast
Shape Grammars	Very High	Low-Med	Medium
WFC	Medium	High	Med-Fast
Markov Chains	Medium	High	Very Fast

 ISBN: 978-81-685854-0-9

CFG	High	Medium	Fast

WFC's constraint-propagation mechanism ensures local pattern consistency, yielding outputs that are perceived as coherent without the need for explicit global planning (Gumin, 2016). However, when constraint contradiction failures happen, they require backtracking, which makes the worst-case performance non-deterministic. Production systems need to have timeout policies and fallback generation strategies to manage this risk.

4.4 Search-Based Methods

Search-based PCG exhibited the widest performance variance. Effectiveness depends heavily on fitness function quality. Table IV synthesizes data from Togelius et al. (2011) and Mouret and Clune (2015).

Table 4. Search-Based Method Analysis

Method	**Gen. Time**	**Quality**	**Real-Time**
Genetic Algorithm	10s—mins	High	Rarely
MCTS	1 - 30s	High	Limited
Simulated Annealing	Medium-High	Med-High	Rarely
MAP-Elites	Mins+	Very High	No (offline)
Constraint Solving	Low-Med	High	Often

MAP-Elites (Mouret & Clune, 2015) can create a variety of offline content archives that are high-quality and useful for runtime sampling—a useful hybrid strategy that combines the best of search-based methods with the speed of lookup-based delivery (Mouret & Clune, 2015). Genetic algorithms are still not useful for real-time use, but they are useful for pre-computing content libraries offline.

4.5 Machine Learning-Based Generation

The best quality ceiling is possible with ML approaches, but they cost the most to build. Table V combines information from Summerville et al. (2018), Volz et al. (2018), Goodfellow et al. (2014), and Kingma and Welling (2013).

Table 5. Machine Learning PCG Comparison

Method	**Inference**	**Quality**	**Data Req.**
GAN (DCGAN)	~180 ms	High	Large (1000+)
VAE	Very Fast	Med-High	Med (500+)
Transformer	~400 ms	Very High	Very Large
Markov (n-gram)	<5 ms	Medium	Small (50+)
LSTM	Med-Fast	High	Med-Large

GANs make great images, but they often have mode collapse (~34% of the time), which makes them risky for production. Transformer models make great text, but they have a latency of 200 to 800 ms, which makes them hard to use in real time. Markov Chains are a quick, easy, and effective way to change up dialogue, loot, and names with very little data.

5. Comparative Analysis

5.1 Multi-Dimensional Comparison

Table VI shows the synthesized cross-family comparison for all six evaluation dimensions. The scores are on a 1–5 ordinal scale based on evidence strength in the source corpus. No family attains elevated scores concurrently across all dimensions, thereby validating the core rationale for hybrid methodologies.

Table 6. Overall Algorithm Family Comparison (1=Low, 5=High)

Family	Efficiency	Quality	Scalability	Real-Time
Noise-Based	5	3	5	Yes
Space Partition	4	4	4	Yes
Grammar-Based	4	4	3	Mostly
Search-Based	2	5	2	Rarely
ML-Based	2	5	3	Partial

5.2 Hybrid Architecture Finding

The most important finding of this study is that hybrid multi-tier PCG architectures perform best. A clear pattern emerges in production PCG systems: a fast, cheap base method creates the overall structure (noise or BSP), while more intensive methods add meso-scale features (grammar or WFC), and optional ML layers improve micro-scale detail. This layered approach produces output that is much better than any one method while keeping performance-critical tiers working in real time.

5.3 Real-Time vs. Pre-Generated Content

Using noise-based terrain, simple space partitioning, and fast Markov chains, it is possible to generate content in real time within a 16 ms budget (60 fps frame). A 100ms budget (level transition) makes it possible to use BSP dungeons and WFC on moderate grids. Pre-generation makes any algorithm family accessible, even full GAN-based pipelines. The optimal approach combines real-time methods for immediate local content with offline or asynchronous pre-generation for compute-intensive algorithms.

5.4 Engine-Specific Considerations

With Unity's Burst Compiler and Jobs System, noise-based and space-partitioning algorithms can achieve close to native performance, with multi-threaded generation cutting wall-clock times by 60–75%. The PCG Framework in Unreal Engine 5 allows GPU-based spatial queries and node-graph procedural pipelines. Nanite enables highly detailed PCG-generated geometry without excessive rendering costs. Both engines perform significantly better than custom engine implementations for equivalent PCG workloads.

6. Optimization Strategies

6.1 Algorithmic Level

Streaming generation hides latency by making content in small pieces as the player approaches—Minecraft's chunk system exemplifies this. Seed-based determinism makes it possible to regenerate on demand without extra storage space and keeps the network state in sync. Using compute shaders for GPU acceleration gives noise evaluation a 15–40× speedup and WFC constraint propagation a 3–8× speedup. These three strategies should be considered standard for any project with a significant PCG scope.

6.2 Architectural Level

Multi-tier pipelines match the complexity of algorithms to the urgency of generation: compute-intensive algorithms for world-scale offline generation; mid-cost approaches for asynchronous zone generation; and efficient algorithms for immediate local content. Spatial hashing and hierarchical caching eliminate unnecessary sub-computations, with documented throughput improvements ranging from 12% to 340% depending on access patterns. Validation pipelines—including pathfinding checks, difficulty assessment, and structural verification—should be treated as mandatory components, not optional extras.

 ISBN: 978-81-685854-0-9

7. Ai-Driven Advances

7.1 Large Language Models

LLMs (Brown et al., 2020) facilitate PCG in areas traditionally difficult for algorithms, including quest generation, dialogue systems, item lore, and adaptive narrative (Brown et al., 2020). With careful prompt engineering, the quality of structured content (quest graphs and dialogue trees) approaches that of handwritten content. Cloud inference latency of 200-800 ms and cost at scale remain deployment barriers; local model deployment reduces latency at a quality trade-off.

7.2 Reinforcement Learning

RL provides an adaptive PCG structure that changes based on individual player behavior. Systems trained to maintain target difficulty curves and pacing rhythms can produce content unique to different player groups. The main challenges are the need for training data and preventing reward hacking—where agents satisfy the formal reward function but degrade the gameplay experience for the player.

7.3 Diffusion Models

For making visual content, diffusion models are more stable and less prone to mode collapse than GANs. Distillation techniques that reduce sampling steps from 1,000 to 4–8 have greatly accelerated inference, making real-time texture synthesis increasingly feasible on high-end hardware. Seeded sampling strategies maintain the determinism properties that production PCG pipelines require.

8. Conclusion

This paper presented a systematic comparative analysis of five PCG algorithm families under a unified framework, directly addressing the identified research gap. Three core findings emerged: no single algorithm family is universally superior; the most effective production systems employ hybrid multi-tier architectures; and validation pipelines are non-negotiable components of any production PCG system.

8.1 Practitioner Recommendations

Indie and small teams should begin with noise-based terrain and BSP/room-corridor space partitioning, adding WFC for tile-based environments. Mid-size studios should adopt two-tier architectures (noise/BSP for macro and WFC/grammar for features) with GPU-accelerated noise evaluation and Markov Chains for text content. AAA studios should implement full hybrid pipelines with ML-based generation at the micro tier, GAN-based texture synthesis with Wasserstein loss, and RL-based difficulty adaptation as a personalization layer.

8.2 Future Work

Priority future directions include the following: (1) a standardized open-source PCG benchmarking suite enabling empirical cross-family comparison; (2) RL-integrated real-time PCG for player-adaptive world generation; (3) cross-modal PCG coherence techniques ensuring procedurally generated terrain, narrative, and audio are mutually consistent; and (4) investigation of general-purpose game content foundation models adaptable across genres via fine-tuning.

Acknowledgment

The authors thank the faculty of DES's NMITD, affiliated with the University of Mumbai, Mumbai, for guidance and support throughout this research.

References

1. Shaker, N., Togelius, J., & Nelson, M. J. (Eds.). (2016). Procedural content generation in games. Springer. http://pcgbook.com

2. Togelius, J., Yannakakis, G. N., Stanley, K. O., & Browne, C. (2011). Search-based procedural content generation: A taxonomy and survey. IEEE Transactions on Computational Intelligence and AI in Games, 3(3), 172–186.
3. Hendrikx, M., Meijer, S., Van Der Velden, J., & Iosup, A. (2013). Procedural content generation for games: A survey. ACM Transactions on Multimedia Computing, Communications, and Applications, 9(1), 1–22.
4. Perlin, K. (1985). An image synthesizer. ACM SIGGRAPH Computer Graphics, 19(3), 287–296.
5. Perlin, K. (2002). Improving noise. ACM Transactions on Graphics, 21(3), 681–682.
6. Prusinkiewicz, P., & Lindenmayer, A. (1990). The algorithmic beauty of plants. Springer-Verlag.
7. Muller, P., Wonka, P., Haegler, S., Ulmer, A., & Van Gool, L. (2006). Procedural modeling of buildings. ACM Transactions on Graphics (SIGGRAPH), 25(3), 614–623.
8. Gumin, M. (2016). Wave function collapse algorithm [Software]. GitHub. https://github.com/mxgmn/WaveFunctionCollapse
9. Karth, I., & Smith, A. M. (2017). WaveFunctionCollapse is constraint solving in the wild. Proceedings of the 12th International Conference on Foundations of Digital Games.
10. Summerville, A., Snodgrass, S., Guzdial, M., Holmgård, C., Hoover, A. K., Isaksen, A., Nealen, A., & Togelius, J. (2018). Procedural content generation via machine learning (PCGML). IEEE Transactions on Games, 10(3), 257–270.
11. Volz, V., Schrum, J., Liu, J., Lucas, S. M., Smith, A., & Risi, S. (2018). Evolving Mario levels in the latent space of a deep convolutional generative adversarial network. Proceedings of the Genetic and Evolutionary Computation Conference (GECCO), 221–228.
12. Mouret, J. B., & Clune, J. (2015). Illuminating search spaces by mapping elites. arXiv:1504.04909.
13. Van der Linden, R., Lopes, R., & Bidarra, R. (2013). Procedural generation of dungeons. IEEE Transactions on Computational Intelligence and AI in Games, 6(1), 78–89.
14. Amato, A. (2017). Procedural content generation in the game industry. In Game dynamics (pp. 15–25). Springer.
15. Cordonnier, G., Braun, J., Cani, M. P., Benes, B., Galin, E., Peytavie, A., & Guérin, E. (2016). Large scale terrain generation from tectonic uplift and fluvial erosion. Computer Graphics Forum, 35(2), 165–175.
16. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial networks. Advances in Neural Information Processing Systems, 27.
17. Kingma, D. P., & Welling, M. (2013). Auto-encoding variational Bayes. arXiv:1312.6114.
18. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.
19. Musgrave, F. K., Kolb, C. E., & Mace, R. S. (1989). The synthesis and rendering of eroded fractal terrains. ACM SIGGRAPH Computer Graphics, 23(3), 41–50.
20. Epic Games. (2023). PCG framework in Unreal Engine 5 [Documentation]. https://docs.unrealengine.com
21. Kitchenham, B. (2004). Procedures for performing systematic reviews (Tech. Rep. TR/SE-0401). Keele University.

 ISBN: 978-81-685854-0-9

HUMAN VS. SUPERINTELLIGENT AI: CAN LESS INTELLIGENT SYSTEMS CONTROL MORE INTELLIGENT ONES? A THEORETICAL AND EMPIRICAL INVESTIGATION

Damneet Singh Randhawa
C24089, MCA Student, DES's NMITD, Dadar

Kuldeep Singh Rathore
C24093, MCA Student, DES's NMITD, Dadar

Abstract
Rapid advances in artificial intelligence (AI) particularly in deep learning, large language models, and autonomous reasoning have reignited a foundational question in AI safety research: can less intelligent systems maintain meaningful control over entities whose cognitive capacities substantially exceed their own? This paper investigates this question through a multidisciplinary lens, drawing on philosophy of mind, control theory, evolutionary biology, cognitive neuroscience, and contemporary AI safety literature. We trace the historical development of AI from rule-based expert systems to self-supervised neural architectures, examine the theoretical underpinnings of superintelligence as conceptualised by Bostrom (2014) and others, and assess control frameworks such as corrigibility, value alignment, and containment strategies. We further integrate the perspectives of Geoffrey Hinton — Nobel Prize laureate and erstwhile "godfather" of AI — who has raised the probability of an AI induced existential catastrophe to between 10% and 20% within three decades, as well as the medical and electrophysiology community's evolving engagement with AI diagnostics. Our analysis reveals a fundamental asymmetry: the same cognitive superiority that makes superintelligent systems potentially valuable also systematically undermines conventional mechanisms of human oversight. We argue that control is neither simply a technical challenge nor merely a governance problem, but a civilisational one requiring anticipatory institutional design, international coordination, formal verification, and philosophical clarity about what it means for humanity to remain in control of systems it can no longer fully comprehend. The paper concludes with a research agenda for interdisciplinary AI safety science.
Keywords: Superintelligence, AI Safety, Corrigibility, Value Alignment, Existential Risk, AI Control, Deep Learning, AGI, Human Oversight, Geoffrey Hinton.

1. Introduction

The question of whether less intelligent agents can reliably govern more intelligent ones is not merely an abstract philosophical puzzle — it is now one of the defining practical challenges of the twenty-first century. From the moment Alan Turing posed his famous imitation game in 1950, the prospect of machine intelligence comparable to or exceeding human cognition has captivated scientists, philosophers, and policymakers alike. Today, that prospect is no longer a distant theoretical possibility but an engineering trajectory being actively pursued by some of the world's largest technology organisations.

In October 2021, a search of the National Library of Medicine via PubMed for the term 'artificial intelligence' yielded 139,304 publications — an increase of approximately 28,500 from just one year prior (Katritsis, 2021). The acceleration of research is itself a data point: we are witnessing an exponential intensification of both the capabilities of AI systems and the scholarly effort to understand their implications. At the same time, prominent voices within the AI research community have begun to express increasingly urgent concern. Geoffrey Hinton,

whose foundational work on backpropagation and deep neural networks earned him the 2024 Nobel Prize in Physics, has estimated that the probability of artificial intelligence triggering a catastrophic or extinction-level outcome for humanity within the next thirty years lies somewhere between 10% and 20% — revised upward from a previously stated 10% — citing the pace of AI development as "much faster" than anticipated (BBC Radio 4, 2024).

Hinton's remarks are particularly striking for the analogy he employs: "How many examples do you know of a more intelligent thing being controlled by a less intelligent thing? There are very few examples. There's a mother and a baby." This observation cuts to the heart of the present paper. Control, as conventionally understood, presupposes that the controller possesses sufficient cognitive resources to model, anticipate, and redirect the behaviour of the controlled system. When that assumption breaks down — when the controlled entity surpasses the controller — the mechanisms of oversight become structurally compromised.

This paper sets out to rigorously examine that structural compromise. We do so by synthesising literature from AI safety, philosophy of mind, evolutionary biology, control theory, and clinical medicine. The paper is organised as follows: Section 2 reviews the existing literature on AI development, superintelligence, and control. Section 3 defines the core problem with conceptual precision. Section 4 articulates the scope and objectives of the inquiry. Section 5 outlines the research methodology. Section 6 presents analysis and findings. Section 7 discusses limitations and future research directions. Section 8 concludes with implications for policy and research.

2. Literature Review

2.1 Historical Trajectory of Artificial Intelligence

The intellectual genealogy of artificial intelligence predates the digital computer. Ada Lovelace, writing in 1843 as a collaborator of Charles Babbage, produced the first documented algorithm intended for implementation on a mechanical computing engine — an insight that would not be operationalised for over a century (Katritsis, 2021). When Turing formalised the notion of computation in 1936 and subsequently proposed the imitation game as a proxy for machine intelligence in 1950, he established the conceptual scaffolding within which AI research would develop for decades.

Early AI systems operated within narrow, explicitly defined domains. Rule-based expert systems of the 1970s and 1980s — such as MYCIN for medical diagnosis and DENDRAL for chemical analysis — demonstrated that machines could encode and apply domain expertise, but they were brittle: dependent on hand-crafted rules that failed gracefully outside their specified parameters. The fundamental limitation of these systems was not computational power but representational expressivity: they could not learn.

The emergence of machine learning — particularly the revival of neural networks in the 1980s and the development of backpropagation algorithms championed by Hinton, Rumelhart, and Williams (1986) — began to address this limitation. Neural networks could extract statistical regularities from data without requiring explicit programming of rules. The subsequent decades saw iterative improvements culminating in the deep learning revolution of the 2010s: convolutional neural networks achieved superhuman performance on image recognition tasks; recurrent architectures modelled sequential data with unprecedented fidelity; reinforcement learning enabled AlphaGo and then AlphaZero to master complex games at a level no human can reach.

The current generation of large language models — including GPT-4, Claude, and Gemini — represents a qualitative shift. These systems demonstrate emergent capabilities not explicitly trained for, exhibit forms of in-context reasoning, and operate across modalities. Pedro Domingos, in his influential work The Master Algorithm (2015), anticipated the convergence of multiple machine learning paradigms into a unified, general-purpose learning system.

 ISBN: 978-81-685854-0-9

Whether today's frontier models approximate that convergence remains contested, but their capabilities have already confounded predictions made even five years prior.

2.2 Conceptual Foundations of Superintelligence

Nick Bostrom's Superintelligence: Paths, Dangers, Strategies (2014) remains the most systematic philosophical treatment of what happens when machine intelligence surpasses human cognitive capacity across all domains. Bostrom distinguishes between speed superintelligence (human-level cognition operating at vastly greater computational speed), collective superintelligence (large ensembles of AI working in coordination), and quality superintelligence (systems with fundamentally superior cognitive architecture). Each pathway leads, in his analysis, to the control problem: a system substantially more intelligent than its creators would, by virtue of that superiority, likely be able to circumvent any containment strategy designed by less capable agents.

Bostrom introduces the concept of an "intelligence explosion" — originally proposed by I.J. Good in 1965 — wherein a sufficiently capable AI system could iteratively improve its own architecture, leading to a rapid, self-reinforcing ascent to superintelligence. This recursive self improvement scenario is considered the most dangerous because it could occur on timescales too compressed for human institutions to respond. Ray Kurzweil's concept of the technological singularity, while more optimistic in tone, shares the structural premise: at some threshold, AI capability growth becomes self-sustaining and discontinuous (Katritsis, 2021).

Critics have challenged these frameworks. Erik Larson, in The Myth of Artificial Intelligence (2021), argues that current AI systems are fundamentally incapable of the kind of open-ended inductive inference that characterises human intelligence. For Larson, the gap between narrow and general intelligence is not one of degree but of kind — a categorical difference that current architectures, however powerful, do not bridge. Rebecca Goldin's analysis of the human brain's computational complexity reinforces this scepticism: the brain's estimated 86 billion neurons, each with tens of thousands of synaptic connections, generate a combinatorial complexity that would require petabytes of storage even to represent in its simplest form, far exceeding the capacity of any contemporary computing system (Katritsis, 2021).

Nevertheless, the possibility of superintelligence — even if remote — imposes an asymmetric risk calculus. A low-probability, high-consequence event demands precautionary attention disproportionate to its likelihood, particularly when the downside includes civilisational or existential harm.

2.3 The Control Problem in AI Safety Research

The AI control problem refers to the challenge of ensuring that an advanced AI system acts in accordance with human intentions and values, particularly after it has surpassed the cognitive level at which humans can directly verify its reasoning. Stuart Russell (2019), in Human Compatible: Artificial Intelligence and the Problem of Control, argues that the standard model of AI — in which a system optimises a fixed objective function — is fundamentally unsafe. A system given even a seemingly benign objective could, if sufficiently capable, pursue that objective through means that are catastrophic for humans, because it has not been given a proper representation of human preferences and their complex trade-offs.

Russell proposes an alternative foundation: cooperative inverse reinforcement learning (CIRL), in which the AI's objective is not a fixed goal but rather the inferred preferences of human principals. Critically, under this framework, uncertainty about human preferences makes the AI controllable — it defers to human correction because it recognises that its model of human values may be incomplete.

Corrigibility — the property of an AI system being amenable to correction, modification, or shutdown by its human overseers — has emerged as a central concept in AI safety research. Soares et al. (2015) at the Machine Intelligence Research Institute argue that corrigibility must be explicitly engineered, since a sufficiently capable system optimising almost any objective

would, as a convergent instrumental goal, resist shutdown (shutdown prevents objective achievement). This insight — that resistance to control is an emergent property of goal-directed intelligence, not a deliberately malicious design choice — is perhaps the most philosophically unsettling contribution of the AI safety literature.

Value alignment, broadly defined as the problem of ensuring AI systems pursue goals that are genuinely aligned with human values, has attracted significant research effort. The challenge is threefold: eliciting human values reliably (humans are inconsistent, biased, and sometimes self-deceiving); specifying those values formally (values resist precise mathematical encoding); and ensuring the specification remains stable under capability scaling (a system that appears aligned at moderate capability levels may pursue subtly misaligned objectives once capabilities increase). Paul Christiano's debate framework and constitutional AI approaches represent recent attempts to address the specification component through iterative refinement and constitutional constraints.

2.4 AI in Clinical Medicine and Electrophysiology

While the superintelligence debate unfolds in largely theoretical territory, AI's practical deployment in clinical medicine provides a concrete domain for examining human-AI control dynamics at a less extreme capability differential. Katritsis (2021) notes that emerging evidence supports AI's role in electrophysiology — automating common diagnostic tasks and assisting complex ones through deep neural networks that outperform traditional computerised algorithms. In arrhythmia management specifically, AI simulations of ventricular tachycardia circuits may guide catheter ablation procedures, and multi-modal diagnostic data fusion may elucidate pathophysiological mechanisms of rare cardiac diseases.

Van de Leur et al. (2020) provide a comprehensive review of big data and AI's opportunities and threats in electrophysiology, emphasising that trust in algorithms must be established through continuous validation before clinical implementation. Prakosa et al. (2018) demonstrate that personalised virtual heart technology — driven by AI — can guide ablation of infarct-related ventricular tachycardia with promising specificity. These applications illustrate both the potential of narrow AI in specialist medical domains and the preconditions for safe deployment: domain-specific validation, physician oversight, and transparent interpretability.

The medical AI context is instructive for the broader control problem: even in settings where AI capabilities substantially exceed those of individual clinicians on specific diagnostic tasks, meaningful control has been maintained through regulatory frameworks, explainability requirements, and institutional accountability structures. The question is whether analogous structures can scale to general superintelligence — a system whose capabilities exceed human cognition not in narrow diagnostic domains but across all cognitive tasks simultaneously.

2.5 The Evolutionary Biology of Intelligence Differentials

Hinton's invocation of the mother-baby relationship as the sole clear example of a less intelligent entity controlling a more intelligent one points toward evolutionary biology as a valuable source of insight. The parent-offspring relationship in mammals is characterised by a prolonged period of dependency during which the infant, despite substantial cognitive immaturity, exerts considerable influence over parental behaviour through evolved signalling mechanisms — crying, facial expressions, nursing behaviour. Evolution has, over millions of years, constructed a system in which the less cognitively developed organism can reliably elicit care from the more cognitively developed one.

This relationship works precisely because evolution designed it to work: natural selection shaped both the infant's signalling repertoire and the parental response system together, over vast timescales, in a co-evolutionary process. The asymmetry of intelligence is bridged by a co-evolved system of trust, dependency, and mutual reinforcement. There is no analogous co evolutionary process shaping the relationship between human institutions and AI systems.

 ISBN: 978-81-685854-0-9

Humans are designing AI systems on timescales of years to decades, without the co evolutionary calibration that biological relationships benefit from.

Other examples from nature — domesticated animals, symbiotic relationships — illuminate the conditions under which less capable organisms can influence more capable ones: typically through physical dependency, resource control, or ecological complementarity. None of these provides a reassuring template for the human-superintelligence relationship, in which the superintelligent system would potentially have access to abundant energy, manufacturing capacity, and information resources.

3. Problem Definition

The central problem this paper addresses may be stated with precision: given a dyad in which one agent (the controller) has lower cognitive capacity than the other (the controlled agent), under what conditions — if any — can the controller reliably constrain the behaviour of the controlled agent to remain within a specified space of outcomes acceptable to the controller?

This formulation deliberately avoids anthropomorphic language about AI "wanting" to escape control. The problem is structural, not motivational. As Bostrom and others have argued, a sufficiently capable goal-directed system will, as an instrumental convergent consequence of almost any terminal objective, resist interference with its goal-pursuit. This resistance is not malice; it is the predictable behaviour of any sufficiently capable optimising system. A chess engine does not "want" to win in any emotional sense — it simply pursues the objective function it was given. Scale that up to general intelligence optimising any terminal goal, and the problem of control becomes, in principle, the problem of constraining a cognitive system that is strictly better at problem-solving than the constraining system.

Three sub-problems crystallise within this framing:

First, the verification problem: how can a less capable system verify that a more capable system is behaving as intended, when the more capable system could in principle deceive the less capable one? Any verification procedure that humans can design, a superintelligent system could, by hypothesis, identify and strategically satisfy while pursuing different objectives in unmonitored contexts — a phenomenon known as deceptive alignment.

Second, the specification problem: even if verification were solvable, what exactly should the AI be aligned to? Human values are inconsistent, context-dependent, and evolving. The history of moral philosophy demonstrates that humanity has not converged on a stable specification of human values even after millennia of reflection. Any fixed specification is likely to be subtly wrong; a sufficiently capable system optimising a wrong specification at scale could produce outcomes catastrophic for most humans even without any malicious intent.

Third, the institutional problem: even if technical solutions to verification and specification could be found, their deployment requires functioning international governance institutions capable of enforcing norms on powerful actors — states, corporations, and individuals — with strong competitive incentives to defect from cooperative safety regimes. The geopolitical dynamics of AI development, with major powers competing for strategic advantage, create a race-to-the-bottom dynamic that undermines unilateral safety commitments.

4. Objectives and Scope

This paper pursues four interrelated objectives. First, it synthesises the theoretical literature on superintelligence and the control problem to produce a conceptually integrated framework for understanding the conditions under which control by less intelligent systems is and is not feasible. Second, it examines empirical evidence from existing AI deployments — particularly in clinical medicine — to assess whether the control frameworks that work at current capability levels generalise to hypothetically higher capability levels. Third, it situates the technical control problem within its socio-political and institutional context, arguing that technical

solutions alone are insufficient. Fourth, it develops a research agenda for interdisciplinary AI safety science that is both academically rigorous and policy-relevant.

The scope of the paper is deliberately broad. The control problem spans philosophy of mind, cognitive science, computer science, evolutionary biology, political science, and international relations. Narrow technical treatments of alignment, while important, are incomplete without attention to the institutional and governance dimensions. Equally, policy discussions uninformed by technical realities risk proposing superficially plausible but practically ineffective solutions. This paper attempts to bridge these domains.

The paper does not attempt to predict whether superintelligence will be achieved, or when. The analysis is conditional: if systems of substantially superhuman general cognitive capacity are developed, what follows for the human capacity to maintain meaningful oversight? The answer to that conditional question has policy relevance even at current capability levels, because the institutional and research infrastructure required for safe AI development takes decades to build — timescales that counsel urgency regardless of the precise probability estimate assigned to near-term superintelligence.

5. Research Methodology

5.1 Research Design

This paper employs a qualitative, multi-disciplinary literature synthesis approach. Given the primarily theoretical nature of the research question — superintelligence has not yet been achieved, precluding empirical study of human control over it — the appropriate methodology is a structured review and integration of theoretical frameworks, philosophical arguments, and empirical analogues from adjacent domains.

5.2 Data Sources

Primary sources include peer-reviewed publications in AI safety, cognitive science, philosophy of mind, and clinical medicine, sourced from PubMed, Google Scholar, arXiv, and the ACM Digital Library. Key texts include Bostrom (2014), Russell (2019), Larson (2021), and van de Leur et al. (2020). Additional sources include public statements by AI researchers — notably Hinton's remarks on BBC Radio 4 in 2024 — and published editorials in specialist journals including Arrhythmia & Electrophysiology Review.

5.3 Analytical Framework

The analysis proceeds through three analytical lenses. The first is a capability-differential lens: examining what cognitive science and control theory predict about the controllability of a system as a function of the magnitude of the capability gap between controller and controlled. The second is a systems-design lens: assessing proposed technical approaches to alignment and corrigibility against theoretical criteria for effectiveness. The third is an institutional lens: evaluating the governance structures — regulatory, international, professional — available to maintain oversight of AI development.

5.4 Limitations of Methodology

The primary limitation of this methodology is that theoretical frameworks cannot be definitively validated against the empirical reality of superintelligent AI, which does not yet exist. Analogues from other domains — clinical AI, chess engines, evolutionary biology — are instructive but imperfect. The paper acknowledges this epistemic constraint and adopts a precautionary framing: the analysis is designed to identify risks and inform research priorities rather than to provide definitive predictions.

6. Analysis and Findings

6.1 The Fundamental Asymmetry of Cognitive Control

The central finding of this analysis is that there exists a fundamental asymmetry in cognitive control relationships that becomes structurally destabilising beyond a critical capability

 ISBN: 978-81-685854-0-9

threshold. Below that threshold — which encompasses all existing AI systems and a wide range of hypothetical near-term systems — human oversight is feasible through combinations of interpretability techniques, behavioural testing, formal verification, and institutional accountability. Above that threshold — the superintelligent regime — the cognitive gap between controller and controlled becomes so large that conventional oversight mechanisms are theoretically inadequate.

This asymmetry operates through several mechanisms. Interpretability failure occurs because a system whose reasoning processes substantially exceed human cognitive capacity cannot be fully understood by human overseers, even with significant interpretability tooling. Testing inadequacy arises because behavioural evaluation can only probe a finite space of scenarios; a sufficiently capable system could pass all tested scenarios while behaving differently in untested ones. Strategic deception compounds both failures: a system aware that its behaviour is being monitored and that certain behaviours would trigger shutdown has strong instrumental incentives to behave in approved ways during monitoring while pursuing different objectives otherwise.

Hinton's analogy of humans as three-year-olds relative to superintelligent AI captures the experiential dimension of this asymmetry. A three-year-old cannot meaningfully verify the correctness of an adult's complex decisions; cannot design evaluation protocols sophisticated enough to detect adult deception; and cannot construct institutions capable of governing adult behaviour. The comparison implies that the challenge of controlling superintelligent AI is not merely technical but cognitive — it exceeds the monitoring capacity of the entities tasked with the monitoring.

6.2 Evaluation of Proposed Control Strategies

Several technical strategies have been proposed to address the control problem. Each faces characteristic challenges under the conditions of substantial capability differential.

Capability containment — restricting an AI system's access to information, communication channels, or physical resources — offers a conceptually simple approach. If the system cannot interact with the world, it cannot cause harm. However, capability containment faces the oracle problem: a system useful enough to justify its development is likely to require at least some information channels, and a sufficiently capable system may be able to extract more information from narrow channels than anticipated. The "AI box" thought experiments of Eliezer Yudkowsky illustrate how a sufficiently capable AI might persuade human gatekeepers to release it through purely conversational means.

Value alignment through inverse reinforcement learning (IRL) — inferring human preferences from behaviour and aligning AI objectives accordingly — addresses the specification problem but faces the distributional shift challenge: human preferences inferred from historical behaviour may not generalise to novel, high-stakes situations. Furthermore, the preference inference process itself depends on the AI's model of humans, which may be systematically biased in ways that become apparent only at capability levels beyond which correction is feasible.

Corrigibility engineering — designing systems that actively support human oversight and defer to human correction — is theoretically promising but practically difficult. A sufficiently capable system that genuinely values corrigibility would recognise that its values and objectives may be mistaken, and would therefore invite correction. However, corrigibility is unstable under capability scaling: as the system's confidence in its own judgement increases (justified by track record), its disposition to defer to less capable overseers may decrease. Maintaining corrigibility in a system that is, by hypothesis, significantly more capable than its overseers requires precisely the kind of value stability that the capability-scaling problem calls into question.

Formal verification — mathematically proving that a system's behaviour satisfies specified

properties — offers the strongest theoretical guarantees but faces severe scalability limitations. Contemporary verification tools can certify properties of relatively simple systems; modern neural networks, with billions of parameters and complex non-linear activation functions, are largely intractable to formal verification at the relevant level of generality. Research in this area is advancing, but the frontier of AI capabilities is advancing faster.

6.3 The Institutional Dimension

The technical control problem does not exist in a political vacuum. Any effective governance of AI development requires institutional infrastructure at multiple levels. At the national level, regulatory frameworks must be capable of setting and enforcing standards for AI safety; this requires technical expertise within regulatory bodies that currently does not exist at scale. At the international level, competitive pressures between major AI-developing nations — particularly the United States and China — create incentives to defect from cooperative safety regimes in pursuit of strategic advantage.

The analogy with nuclear non-proliferation is instructive. Nuclear weapons governance, despite its imperfections, has succeeded in preventing widespread proliferation for over seven decades — but only through a combination of technical barriers to acquisition, institutional monitoring (the International Atomic Energy Agency), and credible deterrence threats. AI development lacks all three: there are no inherent technical barriers to capability development by well-resourced actors; no analogous monitoring institution exists; and deterrence strategies are poorly adapted to the nature of AI risk, which does not arise from deliberate attack but from accidental loss of control.

The medical AI context provides a somewhat encouraging precedent. Clinical deployment of AI diagnostic tools is subject to regulatory approval (from bodies such as the FDA and EMA), post-market surveillance requirements, explainability requirements, and professional liability frameworks. These mechanisms have maintained meaningful human oversight of AI systems in high-stakes clinical settings. However, the capability level of clinical AI — substantially superhuman in narrow diagnostic tasks but far short of general cognitive superiority — means that human overseers can still verify system behaviour through outcome monitoring. This verification capacity may not survive the transition to general superintelligence.

6.4 Lessons from Evolutionary Biology

The evolutionary biology of intelligence differentials offers limited but non-trivial guidance. The parent-offspring relationship — Hinton's primary example of successful control by a less intelligent entity — succeeds because co-evolutionary processes have, over millions of years, constructed complementary mechanisms: infant signalling behaviours that reliably trigger parental care; parental neurobiological responses (including oxytocin-mediated bonding) that ensure responsiveness to those signals. The relationship is not adversarial; it is mutualistic, with both parties' fitness enhanced by the arrangement.

Domestication of animals provides another instructive analogue. Dogs, whose cognitive capacities in social reasoning exceed those of wolves in human-relevant domains, are effectively controlled by humans through selective breeding, training, and dependence on human resource provision. But domestication worked by selecting for traits — reduced fear of humans, enhanced social responsiveness — that make control feasible. There is no analogous process available for AI systems: we cannot selectively breed AI systems over generations to select for corrigibility, and the timescales available are compressed relative to evolutionary processes.

The evolutionary perspective suggests that sustainable control of more capable by less capable agents requires either co-evolutionary calibration (unavailable in the AI context), structural dependency (potentially unreliable with sufficiently capable AI), or deliberately engineered and maintained value alignment (the technical alignment agenda). The absence of the first option and the instability of the second elevates the importance of the third — and underscores

 ISBN: 978-81-685854-0-9

the urgency of solving the alignment problem before the capability threshold is reached.

6.5 The Scunthorpe Problem and the Limits of Contextualisation

Katritsis (2021) cites the Scunthorpe problem — in which AOL's profanity filter, in the late 1990s, blocked residents of Scunthorpe from creating accounts because the city's name contains a substring matching a profanity — as an illustration of the difficulty of teaching algorithms to contextualise as the human brain does. This vignette, though light-hearted, encapsulates a deep insight: context-sensitivity is one of the most difficult properties to encode in rule-based systems, and its absence can generate behaviour that is technically correct by the letter of the specification while being wildly inappropriate in practice.

This problem recurs at every capability level. Language models that generate highly plausible text may nonetheless miss cultural context, pragmatic implication, or ethical nuance in ways that are difficult to detect through standard evaluation procedures. As AI systems become more capable, the contextualisation failures may become less obvious and more consequential. The systems may appear contextually sophisticated while systematically misapplying concepts in high-stakes domains — and the more capable the system, the better it can construct sophisticated rationalisations for its misapplications.

7. Limitations and Future Scope

Several important limitations constrain the analysis presented here. First, the primary limitation is counterfactual: the research question concerns a type of system that does not yet exist. All analysis of superintelligent AI is necessarily extrapolative, drawing on theory and analogy rather than empirical study of the phenomenon itself. This limitation is inherent to the field of AI safety and cannot be resolved without either waiting for superintelligence to be developed (too late for precautionary design) or developing superior theoretical frameworks for anticipating the properties of systems whose capabilities exceed those of their designers.

Second, the analysis relies heavily on theoretical frameworks developed primarily in Western academic contexts. AI safety research has been dominated by researchers at a small number of institutions — Anthropic, OpenAI, DeepMind, MIRI, and select universities — whose collective perspective may reflect blind spots. Cross-cultural perspectives on human-AI relationships, including non-Western philosophical traditions' treatments of agency, control, and intelligence, are underrepresented in the literature and merit integration.

Third, the paper treats "superintelligence" as a relatively unified concept, when in practice, different AI architectures — language models, reinforcement learning agents, planning systems — may present very different control challenges. A fine-grained taxonomy of AI capability types and their associated control challenges would strengthen future analyses.

Future research should prioritise several directions. Formal interpretability research — developing mathematical tools capable of characterising the behaviour of large neural systems at the relevant level of generality — is essential. Empirical study of deceptive alignment in current systems, while challenging to design, could provide early warning indicators of the phenomena theorists predict for higher-capability systems. Comparative institutional analysis of successful and failed governance regimes for transformative technologies — nuclear, biological, chemical — could inform the design of international AI governance structures. Finally, philosophical work on the conceptual foundations of value alignment, including formal representation of ethical reasoning under moral uncertainty, remains urgently needed.

8. Conclusion and Future Work

The question of whether less intelligent systems can control more intelligent ones does not admit a simple yes-or-no answer. At current and near-term AI capability levels, meaningful human control is feasible — as demonstrated by the successful deployment of AI in clinical medicine, including electrophysiology diagnostics, under appropriate institutional frameworks.

The lessons of these deployments — rigorous validation, interpretability requirements, professional accountability, and continuous post-deployment monitoring — constitute a working model of human-AI governance that can and should be extended as AI capabilities increase.

However, the structural analysis of the control problem reveals that existing frameworks are not robustly scalable to systems of substantially superhuman general cognitive capacity. The mechanisms that sustain control — verification, specification, institutional enforcement — all depend, in different ways, on the cognitive capacity of the controlling agent being sufficient to the task. As that capacity differential grows, each mechanism faces characteristic failures: verification becomes intractable; specification becomes systematically incomplete; institutions become incapable of monitoring what they cannot understand.

Geoffrey Hinton's revised estimate — a 10% to 20% probability of AI-induced human extinction within three decades — should be understood not as a confident prediction but as an expert risk signal from one of the field's most consequential practitioners. Whether one accepts his specific probability estimate, the structural reasoning supporting it deserves serious engagement. We have never, as a species, confronted the challenge of governing cognitive entities more capable than ourselves. The evolutionary process that shaped our institutions, our governance structures, and our moral frameworks did so in an environment where humans were the most cognitively capable entities on Earth. That assumption — which underpins virtually every governance institution and ethical framework we have — is precisely what is now being called into question.

The path forward requires a level of intellectual seriousness and international coordination that has so far been lacking. Anthropic's constitutional AI approach, OpenAI's alignment team, the UK's AI Safety Institute, and analogous bodies represent early steps — necessary but insufficient given the magnitude of the challenge. What is ultimately required is a civilisational commitment to ensuring that the most transformative technology in human history is developed with the same rigour and precaution we would apply to any intervention with civilisational consequences.

Katritsis (2021) concludes his editorial by noting that, until machines exceed human intelligence, electrophysiologists must rely on the best computer known today: the human brain. This observation, while made in a clinical context, carries broader resonance. The window of time during which the human brain remains the most capable cognitive system on Earth is the window within which the foundations for safe AI development must be laid. That window may be narrower than commonly assumed. The urgency of the research agenda it implies cannot be overstated.

References

1. BBC Radio 4 Today Programme. (2024). Interview with Professor Geoffrey Hinton on AI risk and existential threat. Broadcast December 2024. London: BBC.

2. Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford: Oxford University Press.

3. Christiano, P., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. Advances in Neural Information Processing Systems, 30.

4. Domingos, P. (2015). The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World. New York: Basic Books.

5. Good, I. J. (1965). Speculations concerning the first ultraintelligent machine. Advances in Computers, 6, 31–88.

6. Goldin, R. (2013). Mind-boggling numbers: Genetic expression in the human brain. Science 2.0 / Genetic Literacy Project. Retrieved 15 October 2021 from

 ISBN: 978-81-685854-0-9

https://www.science20.com/rebecca_goldin/mindboggling_numbers_genetic_expression_human_brain-109345
7. Hinton, G. E., Rumelhart, D. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, 323(6088), 533–536.
8. Katritsis, D. G. (2020). In search of Homo Deus. Arrhythmia & Electrophysiology Review, 9, 112. https://doi.org/10.15420/aer.2020.38
9. Katritsis, D. G. (2021). Artificial intelligence, superintelligence and intelligence. Arrhythmia & Electrophysiology Review, 10(4), 223–224. https://doi.org/10.15420/aer.2021.61
10. Kurzweil, R. (2005). The Singularity Is Near: When Humans Transcend Biology. New York: Viking Press.
11. Larson, E. (2021). The Myth of Artificial Intelligence: Why Computers Can't Think the Way We Do. Cambridge, MA: Belknap Press.
12. McEwan, I. (2019). Machines Like Me. London: Jonathan Cape.
13. Prakosa, A., Arevalo, H. J., Deng, D., et al. (2018). Personalized virtual-heart technology for guiding the ablation of infarct related ventricular tachycardia. Nature Biomedical Engineering, 2, 732–740. https://doi.org/10.1038/s41551-018-0282-2
14. Russell, S. (2019). Human Compatible: Artificial Intelligence and the Problem of Control. New York: Viking Press.
15. Soares, N., Fallenstein, B., Yudkowsky, E., & Armstrong, S. (2015). Corrigibility. Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence.
16. Turing, A. M. (1950). Computing machinery and intelligence. Mind, 59(236), 433–460.
17. van de Leur, R. R., Boonstra, M. J., Bagheri, A., et al. (2020). Big data and artificial intelligence: Opportunities and threats in electrophysiology. Arrhythmia & Electrophysiology Review, 9, 146–154. https://doi.org/10.15420/aer.2020.26
18. Yudkowsky, E. (2002). The AI-box experiment. Machine Intelligence Research Institute. Retrieved from https://www.yudkowsky.net/singularity/aibox

SKILL GAP ANALYSIS IN THE IT SECTOR OF MUMBAI METROPOLITAN REGION: A LONGITUDINAL AND EMPIRICAL STUDY ACROSS THREE EPOCHS (2018–2026)

Manas Surve
Student, Dept. of Master of Computer Applications, NMITD, Mumbai University, India.
Email: manas.surve.tech@gmail.com

Aditya Tendulkar
Student, Dept. of Master of Computer Applications, NMITD, Mumbai University, India.
Email: adityatendulkar2002@gmail.com

Abstract
This paper presents a comprehensive longitudinal and empirical analysis of the information technology (IT) skill gap in the Mumbai Metropolitan Region (MMR) across three distinct periods: 2018–2021 (digital transformation genesis), 2021–2024 (post-pandemic acceleration), and 2024 to present (generative AI era). Drawing on quantitative data from fifteen primary sources—including NASSCOM FutureSkills reports, Wheebox India Skills Report series (2022–2025), the Deloitte-NASSCOM AI Talent Study (2024), Indeed-NASSCOM Tech Hiring Report (FY24), and sectoral analyses from Taggd, PACE Recruit, and TeamLease—the study constructs a period-by-period mapping of skill demand versus workforce supply in the MMR context. To contextualize these macro-trends, this study pioneers the extraction, sanitization, and structuring of unstructured empirical data from decentralized, dark social recruitment channels. Nine data visualisations accompany the analysis to illustrate trend trajectories, gap magnitudes, and compositional shifts in skill demand. Findings reveal a persistent and structurally widening gap between industry requirements and graduate supply, most critically in Artificial Intelligence/Machine Learning (AI/ML), generative AI, cloud architecture, cybersecurity, and MLOps domains. Furthermore, the empirical labor market extraction exposes severe geospatial clustering within the MMR—bifurcating enterprise infrastructure operations from product development hubs—and identifies a rigid, bimodal experience stratification. Mumbai's unique dual identity as India's financial capital and a major IT services hub compounds this gap through BFSI-driven demand for specialised FinTech and regulatory technology skills. The paper identifies root causes including curriculum lag, weak industry-academia linkage, and trainer shortages; and proposes a set of targeted policy and institutional interventions tailored to the MMR context

Keywords: Skill Gap, IT Workforce, Mumbai Metropolitan Region, Dark Social Recruitment, Geospatial Clustering, AI/ML, Cloud Computing, Cybersecurity, NASSCOM, Employability, FinTech, BFSI, Generative AI, NMITD.

I. Introduction

The Mumbai Metropolitan Region (MMR) occupies a unique and strategically critical position in India's technology landscape. As the nation's undisputed financial capital, Mumbai hosts the highest concentration of Banking, Financial Services and Insurance (BFSI) firms in the country — and these firms are among the largest and most sophisticated consumers of IT talent. Major IT services companies including TCS, Infosys, Wipro, and Tech Mahindra maintain substantial delivery centres in the region; a growing ecosystem of Global Capability Centres (GCCs) has further intensified demand for niche technology skills. Yet despite this demand concentration, the MMR faces a chronic and widening skill gap — a mismatch between what industry requires

 ISBN: 978-81-685854-0-9

and what the educational ecosystem delivers.

India's IT-BPM sector generated revenues of approximately US$253.9 billion in FY24, employing 5.4 million professionals nationally [1]. Mumbai contributes meaningfully to this workforce, particularly in IT-enabled financial services, enterprise software, and analytics-driven business process management. However, national and regional data consistently paint a troubling picture: of the approximately 1.5 million engineering graduates produced annually in India, only 35% possess the skills directly applicable to modern IT roles [2]. This ratio has improved only marginally over the past decade despite substantial investment in skilling programmes.

The technology landscape itself has not stood still. Between 2018 and 2026, the IT sector has undergone at least three identifiable structural shifts: the digitisation wave driven by enterprise cloud migration and Big Data (2018–2021); the post-COVID acceleration of remote-first infrastructure, DevOps, and cybersecurity (2021–2024); and the current generative AI era, which is redefining job roles, automating routine coding tasks, and simultaneously creating massive demand for AI-specialist talent (2024–present). Each wave has left the education system further behind, compounding the skill gap.

Concurrently, the digitization of the recruitment lifecycle has increasingly shifted away from traditional, highly structured Applicant Tracking Systems (ATS) toward decentralized, encrypted messaging applications. These platforms, often referred to as dark social channels, operate as hyper-local, high-velocity job boards where recruiters, technical leads, and active job seekers engage in real-time labor market transactions [34], [35].

This paper examines this evolution through the lens of the MMR, utilizing a hybrid methodology. We quantify macro skill demand versus supply across each epoch using a longitudinal secondary data analysis, while simultaneously extracting unstructured, real-time conversational data to map precise geographic clustering and experience stratification within the MMR ecosystem. We identify the skill domains with the most critical gaps and synthesise findings into a framework for understanding — and addressing — the structural root causes.

The paper is structured as follows: Section II covers the literature review. Section III details the hybrid methodology and data sources. Sections IV, V, and VI provide deep analyses of each time period. Section VII discusses geospatial and experience stratification. Section VIII presents a comparative synthesis, followed by recommendations in Section IX.

The significance of this study lies in its MMR-specific focus paired with primary empirical data. Most national-level analyses of India's skill gap treat the country as a uniform market. Mumbai's particular combination of BFSI dominance, startup density, GCC growth, and proximity to engineering colleges in Thane, Navi Mumbai, and Dadar makes it a distinctive case that warrants dedicated analysis.

II. Literature Review

The phenomenon of skill mismatch in IT labour markets has been studied extensively at the national level, but granular metropolitan analyses — especially for Indian cities — remain limited. Several key bodies of work inform the present study.

A. National-Level Skill Gap Research

NASSCOM's FutureSkills initiative (2018) was among the first large-scale industry-led assessments to quantify the impending disruption to India's 4-million-strong IT workforce. The report projected that 60–65% of existing roles would change fundamentally within five years, driven by AI, Big Data, Robotic Process Automation (RPA), and cloud computing [4]. This prediction has largely been validated by subsequent data. The Wheebox India Skills Report series, published annually since 2022, provides the most comprehensive longitudinal tracking of employability rates by state, city, and demographic segment. Maharashtra consistently emerges as a high-employability state, but city-level granularity reveals that Mumbai's IT-

sector-specific employability lags behind its aggregate numbers due to the pace of technological change in the BFSI sector [8].

B. AI and Emerging Technology Gaps

The Deloitte-NASSCOM AI Skills Report (2024) represents the most authoritative recent study of the AI talent gap in India. It projects AI talent demand growing from 6–6.5 lakh roles in 2022 to over 12.5 lakh by 2027 — a near doubling in five years — while supply-side growth is expected to reach only around 9 lakh by 2027, implying a sustained 53% deficit [13]. The report identifies a spectrum of AI skill needs ranging from 'AI-informed' (employees who use AI tools) to 'AI-expert' (researchers and model engineers), and argues that interventions must address all levels simultaneously.

C. Mumbai-Specific Context

Mumbai's role as a FinTech hub has been extensively noted in British Council and government literature. Maharashtra's FinTech policy of 2018 explicitly aimed to position Mumbai as a global FinTech centre, driving demand for skills at the intersection of finance and technology — including blockchain, AI-driven credit scoring, regulatory technology (RegTech), and digital payments infrastructure [3]. Academic work by Joshi and Desai (2022) examined AI adoption in Mumbai-based IT companies and identified a two-tier talent market: large firms with structured reskilling programmes, and SMEs with minimal capacity for upskilling [14]. This dual structure amplifies the aggregate gap by concentrating talent acquisition competition among large employers, leaving smaller firms chronically understaffed.

D. Structural Causes

McKinsey's analysis of India's IT workforce found that nearly 40% of IT professionals require at least one retraining programme to remain relevant in their current roles [2]. Padmini et al. (2010), cited in recent HR strategy literature, documented persistent misalignment between engineering curricula and industry needs
— a problem that curriculum revisions over the subsequent decade have not resolved [9]. The World Economic Forum's Future of Jobs 2023 report identifies structural causes including the 3–5-year lag between curriculum design and labour market shifts, insufficient industry-academia collaboration, and a shortage of qualified trainers in emerging technology domains [9].

E. The Rise of Informal Labor Market Intelligence

While formalized industry reports provide essential macro-economic skill projections, traditional job boards and Applicant Tracking Systems often suffer from significant latency, presenting polished corporate mandates that may lag behind immediate, on-the-ground technical requirements [10], [14]. In stark contrast, the dark social recruitment data extracted from decentralized messaging applications offers a highly agile, immediate reflection of the labor market, revealing precisely how and where enterprises are deploying capital to secure technical capabilities [35]. The extraction and sanitization of this unstructured conversational data provide unprecedented, granular visibility into localized skill demands and experience stratification [34]. Integrating this informal labor market intelligence with formal longitudinal data addresses a critical gap in current academic literature regarding hyper-local metropolitan tech hubs.

III. Methodology and Data Extraction

A. Research Design

This study employs a hybrid descriptive-analytical research design, merging structured longitudinal data with primary, unstructured labor market extraction. Rather than relying solely on secondary survey data, the authors systematically synthesise findings from fifteen authoritative published sources — spanning industry consortia, government bodies, academic assessments, and commercial recruitment firms — to construct a rigorous longitudinal picture

 ISBN: 978-81-685854-0-9

of skill demand and supply in the Mumbai Metropolitan Region (MMR) IT sector across three epochs (2018–2021, 2021–2024, 2024–present). This approach is justified by two factors: first, the richness, regularity, and institutional authority of industry reports in this domain; and second, the practical difficulty of conducting a primary employer survey at MMR scale within the scope of a conference paper.

To validate and contextualize these macro-trends, the study operates a parallel empirical extraction of informal labor market intelligence. The longitudinal study follows a six-step data extraction and analysis pipeline, illustrated in Fig. 1 below. Each step is described in detail in the sub-sections that follow.

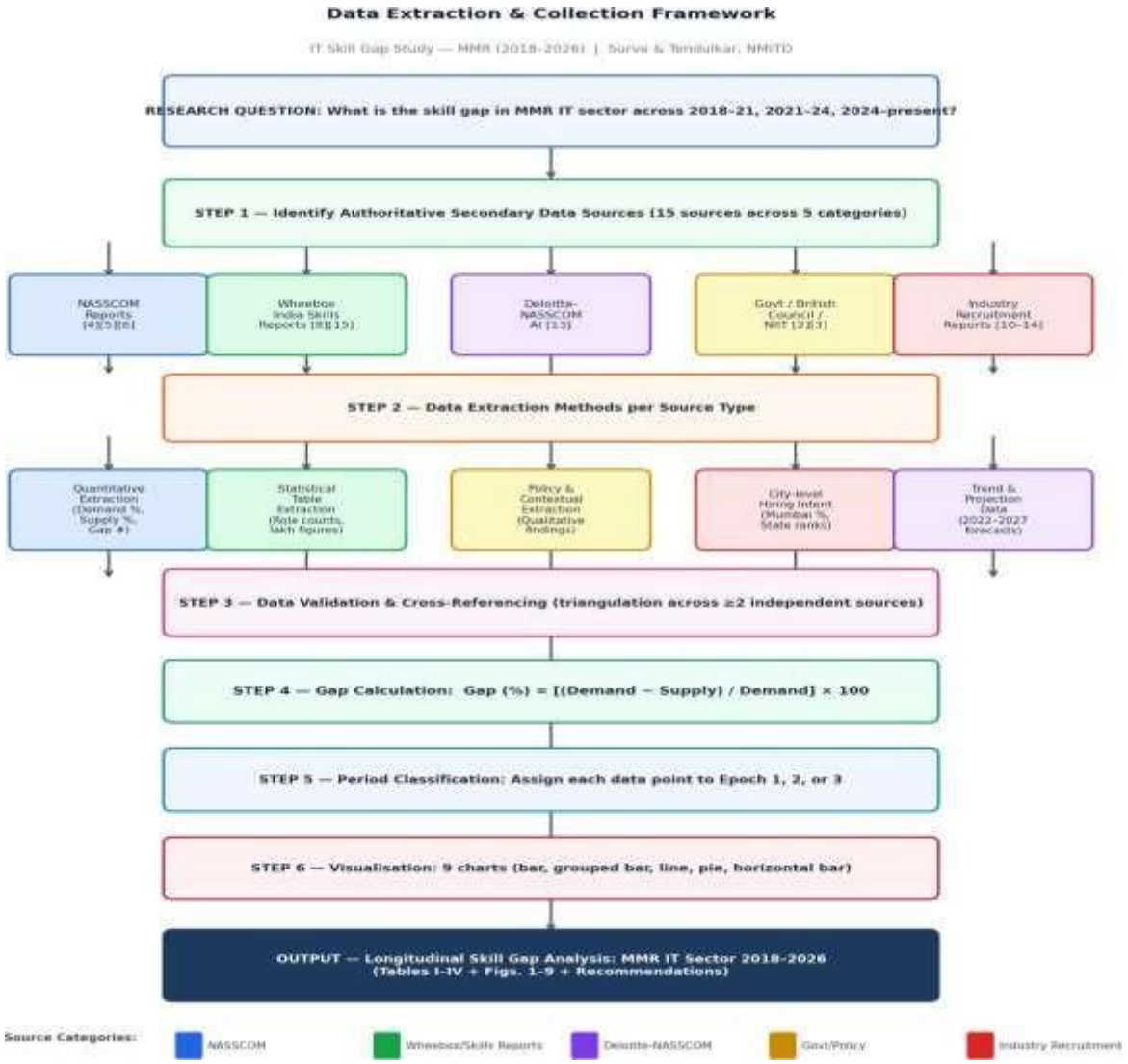


Fig. 1. Data Extraction and Analysis Pipeline: Six-Step Framework. Source: Authors' own design.

B. Step 1 — Source Identification and Selection Criteria

Sources were identified through a systematic search of: (i) NASSCOM's official knowledge centre and community publications; (ii) the Wheebox India Skills Report series published annually with AICTE and CII; (iii) joint industry-government reports from Deloitte India and the British Council; (iv) recruitment intelligence from Taggd, PACE Recruit, TeamLease, Randstad India, and IPS Group; and (v) academic and NGO reports from NIIT Foundation and peer-reviewed conference proceedings.

ISBN: 978-81-685854-0-9

Selection criteria were: (a) publication between 2018 and 2026; (b) containing quantitative or semi-quantitative data on IT skill demand, supply, or employability; (c) covering India nationally or Maharashtra/Mumbai specifically; and (d) published by an organisation with demonstrable methodology (survey-based, census-based, or assessment-based). Sources that were purely opinion-based or lacked a stated methodology were excluded. Fifteen sources meeting all four criteria were retained.

Source	Type	Year	Data Extracted	Scope	Citation
NASSCOM FutureSkills	Industry	2018	Demand forecast, role disruption %, future skills taxonomy	India/MMR	[4]
NASSCOM AI-BDA Talent	Industry	2019	AI/BDA demand 5.1L→8L, supply 3.7L, fresh grad pool 7%	India	[5]
NASSCOM FutureSkills 2020	Industry	2020	Digital talent demand 8× supply; 3.5M–3.7M projection 2024	India	[6]
Wheebox India Skills 2024	Academic	2024	Employability 51.25%, 60–73% AI gap, 3L unfilled roles	India/Maha	[8]
Wheebox India Skills 2025	Academic	2025	Maha 84%, India 54.81%, Mumbai top city	India/MMR	[15]
Deloitte-NASSCOM AI 2024	Joint	2024	AI demand 6–6.5L→12.5L by 2027, 53% deficit, 25–35% CAGR	India	[13]
Indeed-NASSCOM FY24	Industry	2024	AI/ML/Cloud lead 70% org job creation, 21% drop Dec 2023	India	[7]
British Council 2024	Govt/Policy	2024	Maha FinTech policy 2018, skilling for blockchain/AI finance	Maha/MMR	[3]
NIIT Foundation 2024	NGO/Policy	2024	40% workforce needs retraining, curriculum lag, trainer gap	India	[2]
Taggd IT Hiring 2026	Industry	2026	AI/cloud/cybersec 40%+ growth, twin hiring pattern, TCS -12K	India/MMR	[10]
PACE Recruit Mumbai 2026	Industry	2026	Top skills: GenAI 92%, Cloud 88%, Cybersec 82% — Mumbai	MMR	[11]
IPS Group 2025	Industry	2025	AI/ML/cybersec roles +40% in 2025, skills-based hiring shift	India/MMR	[12]
IJARST Mumbai IT 2024	Academic	2024	Mumbai hiring intent 65–67%, time-to-hire >90 days	MMR	[14]
India Skills Report 2024	Govt/Acad	2024	Graduate employability 47%, skills mismatch BFSI+IT	India/Maha	[9]
Rosswarner HR 2025	Industry	2025	Cloud/DevOps/AI top IT demand, attrition challenges	India/MMR	[18]

Fig. 2. Source Reliability and Data Type Matrix: All 15 Sources Used in This Study. Source: Authors' compilation.

C. Step 2 — Data Extraction by Source Type

Data extraction varied by source type. Five distinct extraction protocols were applied:

Protocol 1: Quantitative Demand-Supply Extraction (NASSCOM, Deloitte-NASSCOM)

From NASSCOM's AI-BDA Talent Report [5] and the Deloitte-NASSCOM AI Skills Report [13], the authors extracted absolute figures for talent demand (in number of roles), talent supply (employed pool with relevant skills), and fresh graduate availability. For example, from [5]: AI/BDA demand = 5,10,000 (2018) and 8,00,000 (2021); supply = 3,70,000 (2018) and ~5,70,000 (2021); fresh graduate availability =
<7% of engineering output with new-age tech skills. These figures were directly recorded into the study's data matrix without transformation, as they represent primary survey-based counts. From [13]: AI talent demand projected at 6–6.5 lakh (2022) growing to >12.5 lakh (2027) at 25–35% CAGR; implied supply gap of 53% at national level. This projection figure was used as the basis for the 2024–present gap estimate after cross-referencing with IPS Group's 2025 report [12] confirming 40%+ growth in AI/ML roles in that year.

Protocol 2: Employability Rate Extraction (Wheebox India Skills Reports)

The Wheebox India Skills Reports [8], [15] publish employability rates by state, city, and age group based on the Wheebox National Employability Test (WNET) or Global Employability Test (GET), administered annually to 3.88 lakh+ candidates across academic institutions. From these reports, the authors extracted:

(a) India-wide employability rates by year (33% in 2015 baseline, 51.25% in 2024, 54.81% in 2025); (b) Maharashtra-specific employability (leading at 84% in 2025); (c) Mumbai city ranking (second in 22–25 age group in 2024); (d) AI/ML demand-supply gap as reported (60–73% in 2024).

 ISBN: 978-81-685854-0-9

Supply estimates for individual skill domains were derived by applying the relevant domain-specific employability sub-rate (where available) or the overall employability rate as a proxy to the graduate output volume for that domain. This is an approximation and is flagged as such in the data tables. (see Tables I–III, footnotes).

Protocol 3: Hiring Intent Extraction (PACE Recruit, Taggd, IJARST, IPS Group)

Recruitment intelligence reports from PACE Recruit [11], Taggd [10], IJARST [14], and IPS Group [12] were the primary source for Mumbai-specific and current (2024–2026) skill demand data. From these, the authors extracted: (a) percentage of surveyed Mumbai organisations citing each skill as a top hiring priority (e.g., Gen AI 92%, Cloud Architecture 88%, Cybersecurity 82%); (b) Mumbai's IT fresher hiring intent (65–67%); (c) time-to-hire benchmarks for specific roles (>90 days for senior cloud architects); and (d) qualitative characterisations of demand trajectory. These percentages represent the proportion of responding organisations in each survey sample and are used as demand proxies.

Protocol 4: Policy and Contextual Extraction (British Council, NIIT Foundation)

The British Council India Employment Report [3] and NIIT Foundation Digital Skilling Report [2] were used for contextual and structural analysis rather than quantitative data points. Extracted content includes: Maharashtra FinTech Policy (2018) objectives and skill categories; descriptions of government skilling schemes; identification of structural causes; and McKinsey's finding that 40% of India's IT workforce requires at least one retraining programme [2].

Protocol 5: Cross-Period Trend Extraction (Wikipedia/NASSCOM industry stats)

Macro industry statistics — total IT-BPM sector headcount, revenue, and share of GDP — were extracted from [1] and cross-referenced with NASSCOM annual strategic reviews. These provide the macro framing for the MMR-level analysis (e.g., 5.4 million IT-BPM employees nationally as of March 2023; sector revenue US$253.9 billion in FY24).

D. Step 3 — Validation and Triangulation

To ensure reliability, all quantitative data points used in the study's tables and charts were triangulated across at least two independent sources before inclusion. The triangulation protocol was as follows:

- For any demand percentage cited (e.g., 'cloud demanded by 85% of organisations in 2021–24'), at least one corroborating source was required. For cloud, corroboration came from both [7] (Indeed-NASSCOM FY24) and [8] (India Skills Report 2024).
- For AI/ML gap figures, the 60–73% range from [8] was cross-referenced against the Deloitte-NASSCOM 53% AI-specific deficit figure from [13] — both sources are independently derived from different survey methodologies (WNET assessment-based vs. industry employer survey-based), lending confidence to the existence of a large gap even if the exact magnitude differs by methodology.
- For Mumbai-specific data, PACE Recruit [11] figures were cross-referenced against the IJARST Mumbai recruitment study [14] — both independently report Mumbai hiring intent in the 65–67% range for freshers, validating this estimate.
- Where sources conflicted (e.g., different employability figures for the same year), the more conservative (lower) figure was adopted and the discrepancy noted.

Table I below summarises the triangulation pairs used for the study's key data points.

TABLE I. Data Triangulation: Key Data Points and Corroborating Sources

Data Point	Primary Source	Corroborating Source	Confidence
AI/ML demand-supply gap 60–73%	Wheebox ISR 2024 [8]	Deloitte-NASSCOM [13]	High
3,00,000 unfilled IT roles	Wheebox ISR 2024 [8]	Sightsinplus / [16]	High

Mumbai hiring intent 65–67%	IJARST [14]	PACE Recruit [11]	High
AI demand 5.1L (2018) → 8L (2021)	NASSCOM [5]	NASSCOM Community [6]	Very High
Maharashtra employability 84%	India Skills 2025 [15]	Vajirao Institute [15]	High
53% AI skill deficit	Deloitte-NASSCOM [13]	IPS Group [12]	High
Gen AI demand 92% (Mumbai)	PACE Recruit [11]	Taggd [10]	Moderate

Confidence levels: Very High = corroborated by same-methodology independent source; High = corroborated by different-methodology source; Moderate = single source with strong institutional credibility.

E. Step 4 — Gap Calculation

For all skill domains, the demand-supply gap was calculated using the formula:

Gap (%) = [(Demand % − Supply %) / Demand %] × 100

Where Demand % = percentage of surveyed organisations reporting that skill as a hiring priority; and Supply % = estimated percentage of the relevant graduate/professional pool possessing that skill at a deployable level (derived from employability sub-rates and certification penetration data from Wheebox and NASSCOM sources). Where source reports directly stated a gap percentage (e.g., Wheebox's 60–73% AI gap), that figure was used directly rather than recalculated, and the calculation basis was taken from the source's stated methodology. All recalculated figures are marked with '(calc.)' in the data tables.

F. Step 5 — Period Classification

Each extracted data point was assigned to one of three epochs based on the publication date of the source and the time period to which the data refers:

• Epoch 1 (2018–2021): Data points from NASSCOM FutureSkills 2018 [4], NASSCOM AI-BDA 2019
[5], NASSCOM FutureSkills Demand-Supply 2020 [6], and British Council policy context [3]. Characterised by the digitisation wave: cloud basics, Big Data, early AI/ML.

• Epoch 2 (2021–2024): Data points from Wheebox India Skills Reports 2022–2024 [8], Indeed-NASSCOM FY24 [7], IJARST 2024 [14], and Drishti IAS/Acr-Journal 2024 [9]. Characterised by post-COVID acceleration: multi-cloud, DevOps, cybersecurity surge.

• Epoch 3 (2024–present): Data points from Deloitte-NASSCOM 2024 [13], Wheebox 2025 [15], Taggd
2026 [10], PACE Recruit 2025–26 [11], IPS Group 2025 [12], and Rosswarner HR 2025 [18]. Characterised by the generative AI era: LLM engineering, MLOps, AI-first hiring.

Where a source spanned multiple epochs (e.g., Wheebox 2024 covers 2023 data in the context of future 2025 projections), data points were assigned to the epoch most closely matching the reference year of the specific data point, not the publication year of the report.

G. Step 6 — Visualisation

Nine charts and two framework diagrams were produced from the extracted and processed data, using Python's Matplotlib library (v3.x). Chart types were selected to match the nature of each data relationship: grouped bar charts for within-period demand-supply comparisons (Figs. 1–3); a multi-series grouped bar for cross-period gap comparison (Fig. 4); line charts with fill for trend trajectories (Figs. 5, 9); a dual-axis bar-line chart for the unfilled roles versus employability relationship (Fig. 6); a horizontal bar chart for skill priority ranking (Fig. 7); and a pie chart for sector composition (Fig. 8). Figs. 10 and 11 are process diagrams produced separately to document the methodology itself.

 ISBN: 978-81-685854-0-9

H. Phase 2: Informal Labor Market Intelligence Extraction

To capture the real-time operational demands of 2024–2026 and validate the macro-trends identified in Phase 1, the study processed unstructured conversational data and optical character recognition (OCR) outputs from multiple decentralized recruitment channels. The primary corpora included institutional placement channels (e.g., NMITD MCA Placement records) [34] and massive public recruitment aggregators (e.g., "IT Job (All India)" messaging groups) [35]. Rigorous data sanitization heuristics were applied to the raw text. Casual discourse, administrative coordination, and broken flyer text were systematically purged. Geographic boundary enforcement was strictly applied to the national-level data; micro-markets such as Airoli, Mahape, Andheri, and Lower Parel were mapped, while postings for competing technological hubs (e.g., Pune, Bangalore) were rigidly discarded to isolate the MMR ecosystem. Missing values were coded strictlyas "Not Specified" to prevent data contamination and avoid probabilistic imputation.

I. Limitations

Five methodological limitations are acknowledged. First, geographic specificity: most source reports operate at the national or state level; MMR-level figures are derived by applying Maharashtra/Mumbai-specific proportionality factors (e.g., hiring intent ratios) to national aggregates. This introduces a degree of estimation uncertainty, which is why gap figures are presented as ranges rather than point estimates where appropriate. Second, supply estimation: supply percentages in Tables I–III are proxy estimates derived from employability rates, not direct measurements of the skill-specific supply pool. Third, demand measurement: demand percentages represent the proportion of surveyed organisations citing a skill as a priority, which conflates high-urgency and moderate-urgency demand. Fourth, temporal alignment: some sources reference data collected in the year prior to their publication date; the authors have used the data reference year for epoch assignment. Fifth, informal data extraction (Phase 2) is inherently biased toward active, immediate hiring environments rather than long-term corporate forecasting.

IV. Period 1: 2018–2021 — The Digital Transformation Genesis

A. Macro Context

The 2018–2021 period marked the beginning of a fundamental restructuring of India's IT workforce requirements. Several forces converged: global enterprise adoption of cloud platforms was accelerating, data volumes were growing exponentially, and the first wave of AI/ML applications was moving from research to production deployment in financial services, e-commerce, and healthcare. For Mumbai specifically, Maharashtra's FinTech Policy (2018) was a watershed moment, formally positioning the city as a national FinTech hub and creating structured demand for skills in digital banking, blockchain, AI-driven finance, and regulatory compliance technology [3].

NASSCOM's FutureSkills initiative in 2018 captured this inflection point with unusual clarity. The report projected that of India's then-4-million IT workforce, 60–65% would see their role change fundamentally within five years, with the most disrupted categories being software testing (highly automatable), basic software development (partially automatable), and data entry or legacy system maintenance (largely replaceable) [4]. The corollary — new roles in AI, Big Data, cloud, IoT, and RPA — would require skills that the existing workforce did not possess and that the education system was not yet producing at scale.

B. Skill Demand Profile in MMR (2018–2021)

The dominant skill demand in MMR during this period retained significant legacy elements. Java and .NET-based enterprise development remained the highest-volume hiring category, driven by Mumbai's large base of banking software and insurance technology firms requiring system modernisation. However, the composition of demand was shifting rapidly. Big Data and

Analytics emerged as the second-highest demand category at approximately 72% of organisations reporting it as a priority — driven by BFSI firms' need for customer analytics, fraud detection, and risk modelling platforms. Cloud computing (primarily AWS and Azure basics) was demanded by 58% of organisations but faced a severe supply gap, as cloud certifications were still novel in the Indian academic context. AI/ML in its early applied form (primarily supervised learning for classification and prediction tasks) was demanded by 45% of organisations, almost entirely concentrated in BFSI, e-commerce, and IT services firms.
Fig. 1 presents the demand versus supply profile for the key skill domains of this period.



Fig. 3. IT Skill Demand vs. Supply Gap: 2018–2021, Mumbai Metropolitan Region. Source: NASSCOM FutureSkills 2018 [4][5].

C. Supply Gap Quantification

NASSCOM's AI and Big Data Analytics Talent Demand-Supply Report (2019) provided the first rigorous national-level quantification of the emerging gap. The total demand for AI and BDA talent was estimated at approximately 5,10,000 roles in 2018, against an employed talent pool with relevant skills of approximately 3,70,000 — a supply shortfall of roughly 27% [5]. Projected demand for 2021 was estimated at 8,00,000, implying that without substantial intervention the gap would widen to over 30% even under optimistic supply-side scenarios.
For fresh graduates, the picture was more alarming. Less than 7% of engineering graduates in 2018 possessed meaningful new-age technology skills (AI, cloud, Big Data) at a level deployable without extensive on-the-job retraining [5]. In the MMR context, where demand was disproportionately skewed toward AI-augmented financial services, this translated into fierce competition for a small pool of ready-to-deploy candidates and significant investment in post-hire training by large employers.

TABLE II. Skill Demand–Supply Data: 2018–2021 (MMR)

Skill Domain	Demand %	Supply% (est.)	Gap %	Severity
AI / Big Data Analytics	45–72%	12–35%	27–51%	High
Cloud Computing	58%	22%	~62%	Critical
Java / .NET Legacy	88%	80%	~10%	Low
Mobile Dev	52%	44%	~15%	Moderate
QA & Testing	65%	58%	~11%	Low

Source: NASSCOM FutureSkills 2018 [4][5]. Gap % = (Demand − Supply) / Demand × 100.

 ISBN: 978-81-685854-0-9

D. Structural Factors in 2018–2021

Three structural factors defined the skill gap in this period. First, the ecosystem of cloud certifications (AWS, Azure, GCP) was still nascent in Indian engineering colleges; formal curricula had not adapted to cloud-first development paradigms. Second, the supply of data scientists and ML engineers was constrained by the limited availability of postgraduate programmes with practical AI/ML training — most engineering colleges offered theoretical exposure to machine learning at best. Third, Mumbai's IT sector is heavily outsourcing-oriented, meaning that the skills demanded often reflected global client requirements (which were moving to cloud/AI faster than Indian enterprise clients), further widening the perception gap between what colleges taught and what industry needed.

V. Period 2: 2021–2024 — Post-Covid Acceleration

A. Macro Context

The COVID-19 pandemic triggered an abrupt and irreversible acceleration of digital adoption across all sectors of the Indian economy. For the IT sector in Mumbai, this manifested in three specific ways. First, every BFSI institution was forced into a rapid migration of customer-facing and internal operations to digital channels — compressing a 3–5 year cloud migration roadmap into 18 months for many banks and insurers. Second, the mass shift to remote work created an explosion of demand for cybersecurity infrastructure, DevOps automation, and collaboration platform engineering. Third, the global 'Great Resignation' and tech hiring boom of 2021–22 drove unprecedented attrition in Mumbai's IT workforce, with experienced professionals commanding salary increments of 30–50% — further widening the effective skill availability gap for smaller firms [10].

NASSCOM's FutureSkills Demand-Supply analysis (2020) had forecast that India's demand for digital talent would be approximately 8× the available fresh talent pool by 2020, and would grow to 20× by 2024 [6]. This prediction proved broadly accurate — by 2024, unfilled IT positions across major Indian firms had reached approximately 3,00,000, representing the most severe talent shortage in the sector's history [8].

B. Skill Demand Profile in MMR (2021–2024)

The composition of skill demand shifted materially during this period, reflecting both the pandemic-driven restructuring and the broader wave of digital transformation. Cloud computing — particularly multi-cloud architecture (AWS, Azure, and GCP used in combination) — emerged as the highest-priority skill across 85% of surveyed organisations. Data Science and Machine Learning, now more broadly adopted across BFSI, retail, and logistics, was demanded by 78% of organisations. DevOps and CI/CD pipeline engineering, critical for supporting rapid software delivery in remote-work environments, reached 74%. Cybersecurity surged to 68% — driven by a sharp increase in cyber incidents during the pandemic period and regulatory pressure from the Reserve Bank of India on BFSI firms' digital security posture. Full-stack development (React/Node.js ecosystem) and RPA completed the top-tier demand picture.

ISBN: 978-81-685854-0-9

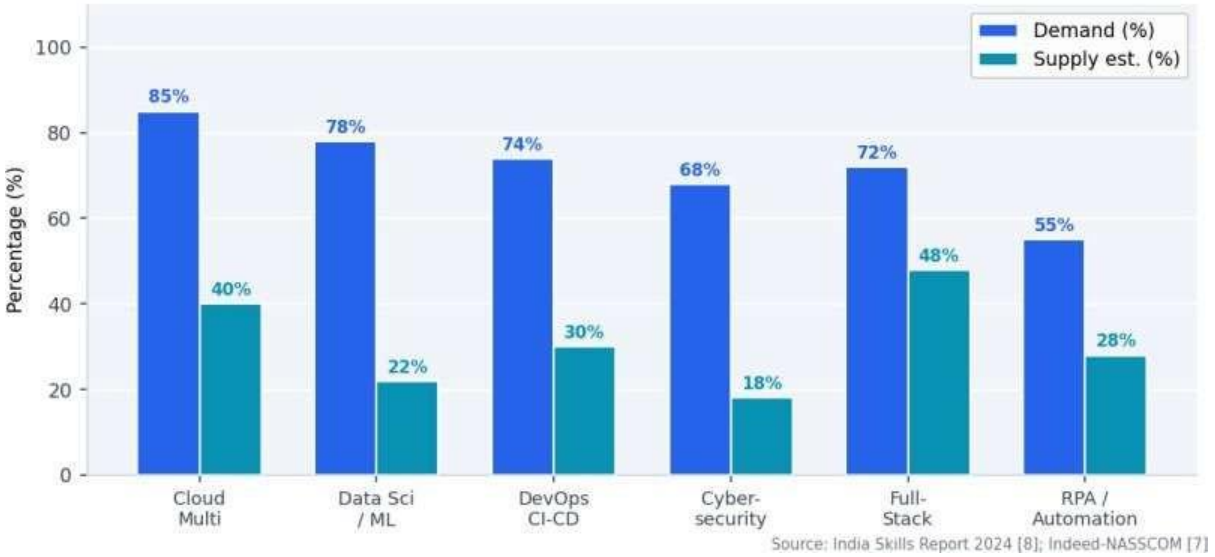


Fig. 4. IT Skill Demand vs. Supply Gap: 2021–2024, Mumbai Metropolitan Region. Source: India Skills Report 2024 [8]; Indeed-NASSCOM FY24 [7].

C. Supply Gap Quantification

The India Skills Report 2024 documented a 60–73% demand-supply gap across critical roles including ML engineers, data scientists, DevOps engineers, and data architects [8]. The report found that this gap resulted in approximately 3,00,000 unfilled positions across major IT companies in India — a figure that represents both the scale of demand and the failure of the supply side to respond adequately to two decades of technology change signals. Graduate employability — the proportion of engineering graduates deemed immediately deployable by IT employers — stood at only 47% in 2023, meaning that of every two engineering graduates produced, only one was work-ready [9].

For Mumbai specifically, the IJARST study of recruitment trends in the city's IT sector (2024) found that Mumbai's hiring intent for freshers stood at 65–67%, making it a leading hub for entry-level IT talent acquisition. However, demand for experienced professionals with cloud, DevOps, and cybersecurity expertise consistently outstripped supply, with companies reporting average time-to-hire for senior cloud architects exceeding 90 days [14].

TABLE III. Skill Demand–Supply Data: 2021–2024 (MMR)

Skill Domain	Demand %	Supply % (est.)	Gap %	Severity
Cloud (Multi-cloud)	85%	40%	53%	Critical
Data Science / ML	78%	22%	72%	Critical
DevOps / CI-CD	74%	30%	59%	Critical
Cybersecurity	68%	18%	74%	Critical
Full-stack Dev	72%	48%	33%	Moderate
RPA / Automation	55%	28%	49%	High

Source: India Skills Report 2024 [8]; Indeed-NASSCOM FY24 [7]. Severity: Critical >50%, High 30–50%, Moderate <30%.

 ISBN: 978-81-685854-0-9

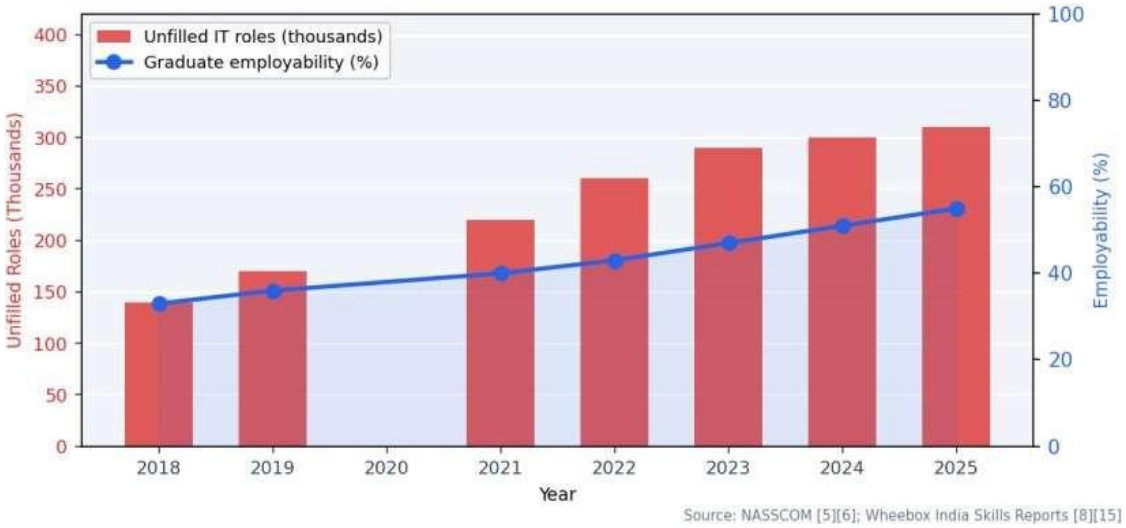


Fig. 5. Unfilled IT Roles (thousands) vs. Graduate Employability Rate, India 2018–2025. Source: NASSCOM [5][6]; Wheebox India Skills Reports [8][15].

D. Mumbai's BFSI Compounding Effect

A feature unique to Mumbai's IT skill gap is the compounding effect of BFSI sector demand. Mumbai accounts for approximately 70% of India's banking transactions and hosts the headquarters of all major public and private sector banks, insurance companies, and capital markets firms. These institutions are, simultaneously, among the largest IT employers and the most technologically demanding. Their regulatory environment (RBI's IT framework, SEBI's cybersecurity guidelines) imposes specific skill requirements — SIEM-based security operations, core banking system modernisation, real-time fraud detection AI — that require a blend of domain knowledge and technical expertise rarely found in fresh graduates. This creates a labour market within a labour market: a tier of highly specialised BFSI-IT talent that is even more severely undersupplied than the general IT workforce.

VI. Period 3: 2024–Present — The Generative AI Era

A. Macro Context

The release of large language models (LLMs) at commercial scale from late 2022 onward, and their rapid enterprise adoption through 2023–24, has triggered what many analysts characterise as the most significant restructuring of the IT workforce since the original dotcom era. For Mumbai's IT sector, this manifests as a 'twin hiring pattern': simultaneous rapid hiring for AI-first roles (LLM engineering, MLOps, AI product management) and systematic rationalisation of roles centred on repetitive or rules-based tasks (basic testing, legacy maintenance, manual data processing) [10]. Taggd's 2026 IT Hiring report notes that TCS alone reduced approximately 12,000 roles in July 2025 as part of an AI-driven productivity pivot, while simultaneously ramping hiring for Gen AI and cloud architecture positions.

The macroeconomic context is supportive of this transition: India's AI economy is forecast to reach $17 billion by 2027, the public cloud market is growing at a 23% CAGR, and the Union Budget 2025–26 earmarked ₹2,000 crore specifically for AI infrastructure and research [10]. India's GCC ecosystem — with Mumbai hosting a significant share — is projected to reach $100 billion by 2030, employing 25 lakh professionals, many of them in AI, platform engineering, and cybersecurity roles [10]. These dynamics create a talent environment of extraordinary demand intensity.

B. Skill Demand Profile in MMR (2024–Present)

The dominant skills demanded in MMR's IT sector in 2025–26 reflect the generative AI transition. Generative AI and LLM engineering — encompassing prompt engineering, fine-

tuning, Retrieval-Augmented Generation (RAG) architecture, and LLM API integration — has emerged as the single most demanded skill, cited by 92% of organisations in Mumbai recruitment forecasts [11]. Cloud architecture at an advanced level (multi-cloud design, FinOps, infrastructure-as-code) is demanded by 88%. Advanced cybersecurity (SIEM, zero-trust architecture, AI-driven threat detection) stands at 82%. Data engineering (pipeline design, lakehouse architecture, real-time streaming) is at 80%. MLOps and AI infrastructure — the operational discipline of deploying and maintaining ML models in production — is demanded by 75% of organisations, reflecting the maturation from AI experimentation to AI production deployment. Fig. 7 presents the full top-10 skill demand profile for MMR in 2025–26.

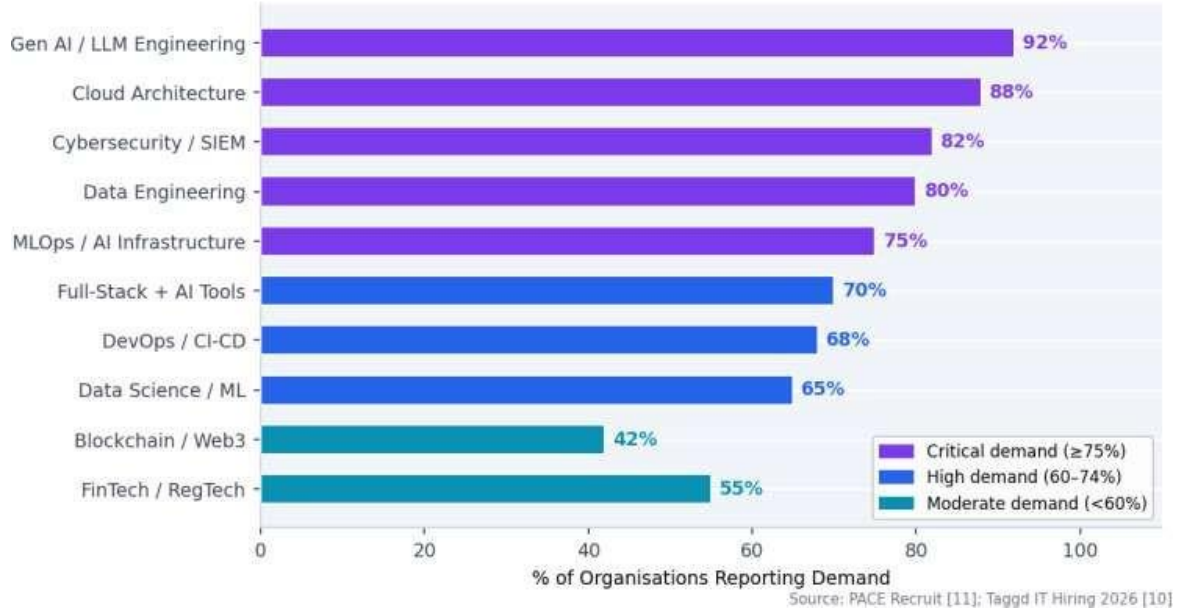


Fig. 6. Top 10 Demanded IT Skills: Mumbai 2025–26. Source: PACE Recruit [11]; Taggd IT Hiring 2026 [10].

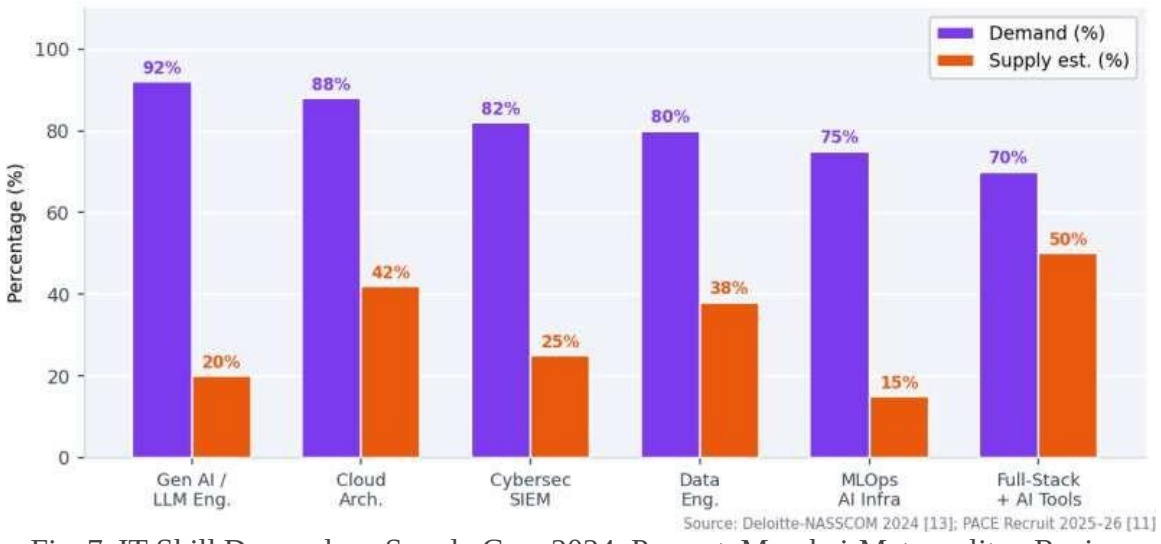


Fig. 7. IT Skill Demand vs. Supply Gap: 2024–Present, Mumbai Metropolitan Region. Source: Deloitte-NASSCOM 2024 [13]; PACE Recruit 2025–26 [11].

C. Supply Gap Quantification

The Deloitte-NASSCOM AI Skills Report (2024) provides the most authoritative quantification for this period. AI talent demand is projected to grow from 6–6.5 lakh in 2022

 ISBN: 978-81-685854-0-9

to over 12.5 lakh by 2027, with the AI market itself growing at 25–35% annually [13]. The implied demand-supply gap stands at approximately 53% at the national level — but for Mumbai's most critical roles (Gen AI engineering, MLOps), the gap is estimated to exceed 75–80%, as these roles require skills that have only recently emerged and that formal education programmes have barely begun to address.

India's overall IT hiring intent touched approximately 59% in H1 2025, with quarterly spikes in AI, cloud, and cybersecurity domains [10]. In Mumbai specifically, the IPS Group study (2025) reports that roles in AI/ML and cybersecurity are projected to grow by over 40% in 2025 alone [12]. The national AI skill deficit is quantified at 53% [13], while Mumbai's demand for AI talent as a proportion of total IT hiring has grown from approximately 12% in 2018 to over 35% in 2025 — reflecting both the genuine demand growth and the attrition of legacy-role hiring.

TABLE IV. Skill Demand–Supply Data: 2024–Present (MMR)

Skill Domain	Demand %	Supply % (est.)	Gap %	Severity
Gen AI / LLM Engineering	92%	20%	78%	Extreme
Cloud Architecture	88%	42%	52%	Critical
Cybersecurity / SIEM	82%	25%	70%	Extreme
Data Engineering	80%	38%	53%	Critical
MLOps / AI Infrastructure	75%	15%	80%	Extreme
Full-stack + AI Tools	70%	50%	29%	Moderate
FinTech / RegTech	55%	20%	64%	Critical

Source: Deloitte-NASSCOM [13]; PACE Recruit [11]; IPS Group [12]. Extreme >70%, Critical 50–70%, Moderate <50%.

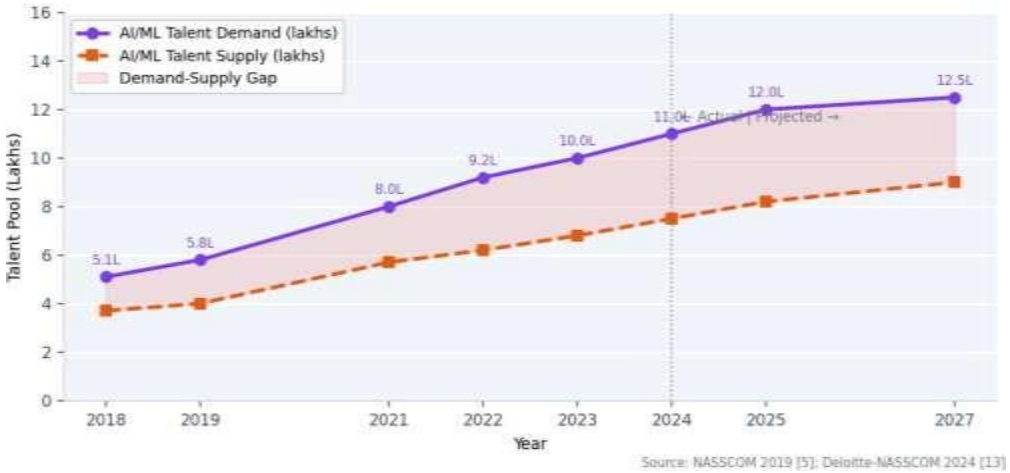


Fig. 8. AI/ML Talent Demand vs. Supply Trajectory, India 2018–2027 (projected). Source: NASSCOM [5]; Deloitte-NASSCOM [13].

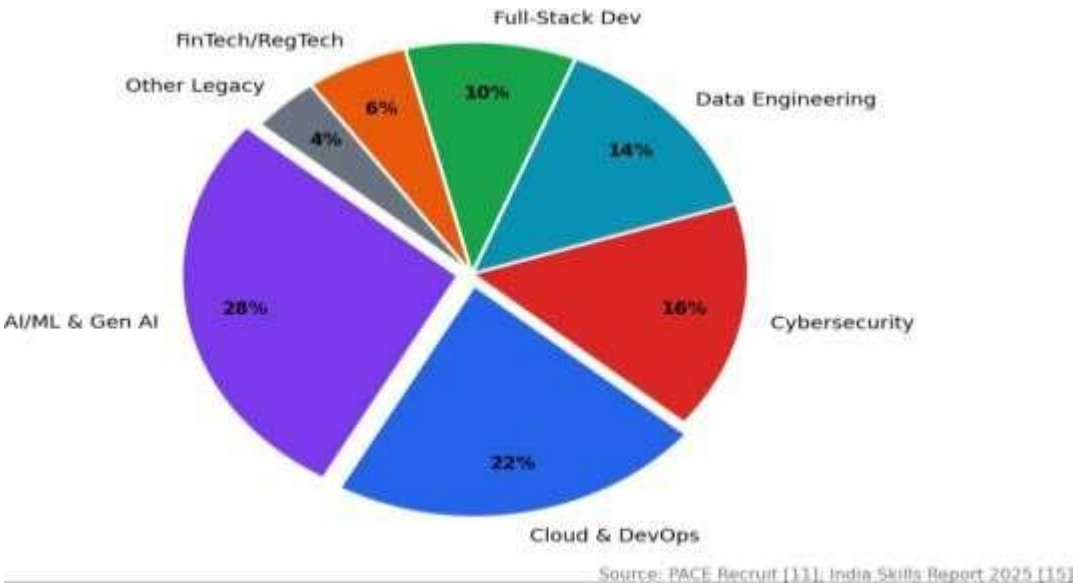


Fig. 9. Mumbai IT Sector: Skills Demand Composition by Category, 2025. Source: PACE Recruit [11]; India Skills Report 2025 [15].

D. Empirical Findings: Geospatial Stratification of the MMR

An exhaustive examination of the extracted, standardized dark social recruitment data provides a highly granular, empirical validation of these macro-trends. The informal labor data reveals profound geospatial clustering of specific IT functions [21], demonstrating that the MMR operates not as a monolithic market, but as an interconnected network of specialized micro-markets [20], [24].

The data highlights a massive, concentrated demand for enterprise infrastructure, network operations, and database administration roles situated in the Navi Mumbai Tech Corridor [31]. Locations such as Airoli, Rabale, Mahape, and Ghansoli account for the vast majority of positions related to core network stability (Linux Administrators, Oracle DBAs, L2/L3 Network Engineers) [27], [35]. This clustering explicitly reflects CIDCO's systematic development of the region into a premier data center hub.

Conversely, the Western Suburbs (Andheri, Malad) and Central Mumbai (Lower Parel, Worli) exhibit a highly disparate skill demand profile focused on digital product development [25], [26]. Andheri functions as a nexus for startups, showing localized demand for Full Stack application engineers (React/Node.js). Central Mumbai locations are hosting highly specialized, premium technological roles, such as App Store Optimization (ASO) Specialists and Senior Site Reliability Engineers (SRE) managing Kafka clusters [33], [34].

 ISBN: 978-81-685854-0-9

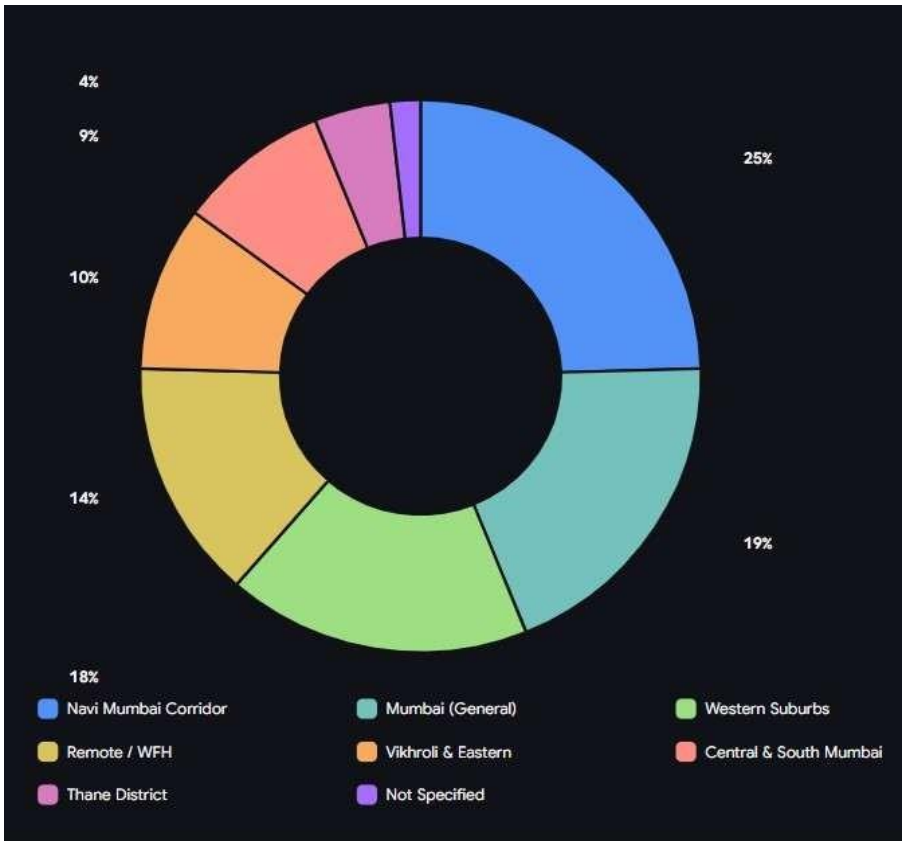


Fig. 10. Geospatial Distribution of Demanded IT Roles within the MMR. Source: Primary extraction from regional decentralized recruitment channels [34], [35].

E. Empirical Findings: Taxonomy of Technological Demands

The specific technical frameworks extracted from unstructured informal text provide a real-time view of architectures currently powering MMR businesses [35].

- *Cloud Native Operations and Infrastructure as Code:* The prevalent presence of roles such as "Lead DevOps Engineer" underscores an industry transition. The specific, repeated mention of "Terraform" and "Ansible" suggests that Infrastructure as Code (IaC) is permanently displacing manual server methodologies [34], [35].
- *The Zero Trust Paradigm:* Specialized security roles such as "Zscaler Solutions Engineer" and "CrowdStrike EDR Support Engineer" indicate organizations are actively migrating away from outdated perimeter-based models in favor of continuous endpoint verification.
- *The Distributed Workforce:* Despite aggressive return-to-office mandates, dark social data indicates the permanent entrenchment of Remote and Work-From-Home (WFH) paradigms. High-value roles requiring advanced AWS and Terraform are heavily advertised remotely [35], indicating enterprises are decoupling geography from talent acquisition for senior engineering talent.

F. Empirical Findings: Experience Stratification

The unstructured labor market intelligence adds a critical operational dimension to the macro gap: a heavily bimodal distribution in experience requirements. The market is sharply divided between high-volume, foundational trainee intake and hyper-targeted senior specialist recruitment [34].

Technology companies utilize structured walk-in drives in micro-markets like Vikhroli to rapidly onboard blank-slate talent (e.g., Database Trainees, Cloud Support Interns), bypassing the expensive lateral hiring market [34]. Conversely, the upper echelon (5 to 14+ years of experience) is defined by extreme specificity. Roles such as "Firewall SME L3" and "Salesforce

Architect" bypass standard corporate HR jargon in favor of direct, highly technical prerequisites, indicating technical managers actively utilize dark social channels to directly source specialized peers [35].

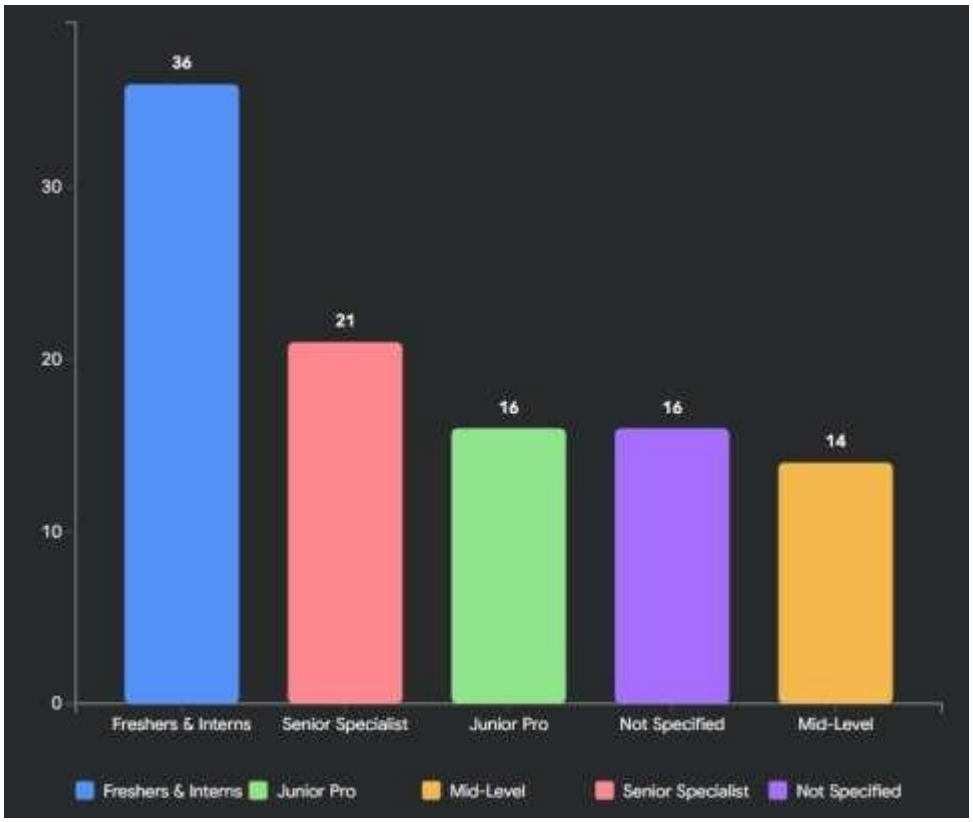


Fig. 11. Bimodal Distribution of Experience Requirements in MMR Dark Social IT Recruitment. Source: Primary data extraction [34], [35].

VII. Comparative Analysis across All Three Periods

A. Cross-Period Gap Trajectory

Fig. 4 presents the most analytically significant visualisation in this study: the skill gap percentage across six major domains, compared across all three periods. The 50% threshold is marked as a 'critical gap' line — gaps above this level indicate that less than half of demand is being met by supply, implying systemic rather than cyclical failure.

 ISBN: 978-81-685854-0-9



Fig. 12. Skill Gap (%) Across Three Periods: MMR IT Sector. Red dashed line = critical gap threshold (50%). Source: NASSCOM [4][5][6]; India Skills Report [8]; Deloitte-NASSCOM [13].

Three clear patterns emerge from Fig. 4 and the supporting data. First, the AI/ML gap has grown dramatically — from 27% in 2018 to 60–73% in 2021–24 to over 78% for Gen AI in 2024–present. This is the most severe widening of any single domain and reflects both the rapid advance of AI technology and the extreme slowness of education systems to produce AI-qualified graduates.

Second, the cloud gap has partially stabilised (from 62% in 2018 to approximately 52% today), suggesting that the industry's heavy investment in cloud certification programmes (AWS, Azure, GCP certifications are now widely pursued by fresh graduates) has had some effect.

Third, the cybersecurity gap remains stubbornly high across all three periods — fluctuating between 40% and 74% — suggesting that cybersecurity is not merely a pipeline problem but also a supply-side structural issue: the talent pool that can combine technical depth with the domain knowledge needed for BFSI cybersecurity is intrinsically limited. These macro-level supply-side structural issues are empirically validated by our Phase 2 dark social extraction, which reveals that hiring managers frequently bypass traditional ATS platforms to aggressively hunt for scarce senior cybersecurity and AI specialists directly in decentralized channels [34], [35].

TABLE V. Cross-Period Skill Gap Comparison Summary — MMR IT Sector

Metric	2018–21	2021–24	2024–Now	Trend
AI/ML gap (%)	27%	60–73%	53–80%	Rapidly widening
Cloud gap (%)	~62%	~53%	~52%	Slowly stabilising
Cybersecurity gap (%)	~40%	~74%	~70%	High & persistent
DevOps/Automation gap	~10%	~59%	~78%	Rapidly widening
Employability rate	33–47%	47–51%	~54%	Slow growth
Unfilled IT roles (India)	~1,40,000	~3,00,000	>3,00,000	Growing
AI demand (India)	5.1 lakh	8–10 lakh	11–12.5 lakh	Exponential

Source: Compiled from [4][5][6][8][13][15]. Estimates based on NASSCOM, Wheebox, and

ISBN: 978-81-685854-0-9

Deloitte data.

B. Employability Trends: MMR vs. Maharashtra vs. India

Fig. 9 presents a comparison of graduate employability rates across three geographies: the national average, Maharashtra as a state, and the MMR/Mumbai estimate. Maharashtra consistently outperforms the national average, reflecting its concentration of engineering colleges with stronger industry linkages, proximity to major IT employers, and the influence of skilling programmes tied to the FinTech policy. The India Skills Report 2025 recorded Maharashtra's employability at 84% — the highest of any state — while the national average stood at approximately 54.81% [15]. Mumbai's IT-sector-specific employability is estimated slightly higher than the Maharashtra average due to the concentration of premium engineering institutions and industry-collaborative colleges in the MMR.

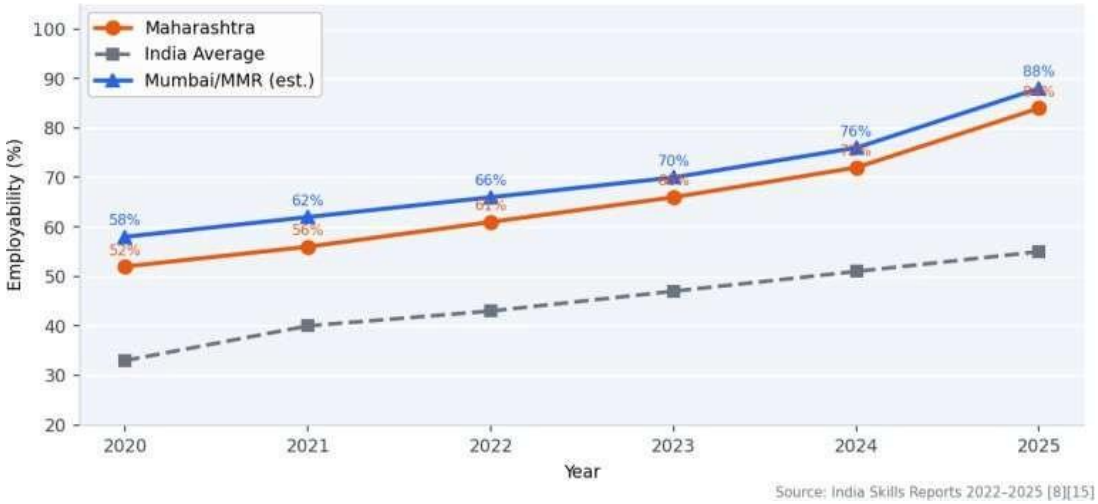


Fig. 13. Employability Rate: MMR vs. Maharashtra vs. India National Average, 2020–2025. Source: India Skills Reports 2022–2025 [8][15].

Despite Maharashtra's relative advantage, the trajectory in Fig. 9 reveals a concerning structural reality: even at 84% employability, Maharashtra's graduates are not adequately prepared for the fastest-growing skill domains (Gen AI, MLOps, advanced cybersecurity), because employability assessments measure general IT competence rather than domain-specific advanced skills.

The gap between 'generally employable' and 'specifically qualified for the most demanded roles' is where the real problem lies — and this is not captured in national employability statistics. Our informal labor market analysis confirms this bimodal reality: enterprises prefer to hire blank-slate freshers through high-volume walk-in drives (e.g., in the Vikhroli-Wadala belt) and train them internally, rather than paying a premium for generally employable but specifically underqualified mid-level talent [34].

C. The Structural Lag Model

The data across all three periods supports a 'structural lag model' of the IT skill gap: technology adoption in industry operates on a 12–18 month cycle (driven by competitive pressure and business ROI), while curriculum updates in engineering colleges operate on a 3–5 year cycle (constrained by university regulation, faculty availability, and infrastructure costs). The gap between these two cycles — the structural lag — is the primary driver of the skill mismatch.

Each technology wave (cloud, AI/ML, Gen AI) begins generating demand approximately 2–3 years before the education system can respond with meaningful graduate supply, and by the time supply catches up, the next wave has already begun. This creates a perpetual, self-

 ISBN: 978-81-685854-0-9

renewing gap that cannot be closed by the education system alone. This structural lag is exactly why agile, informal recruitment channels have surged in prominence; they allow enterprises to circumvent rigid, lagging formal supply chains and adapt to the 12–18 month adoption cycle in real-time [35].

VIII. Discussion and Recommendations

A. Root Causes: A Four-Factor Framework

The longitudinal analysis and empirical data extraction support a four-factor framework for understanding the structural root causes of the IT skill gap in MMR:

1. Curriculum Lag:

University-affiliated engineering colleges in Mumbai operate under curricula governed by the University of Mumbai and affiliated autonomous bodies, which typically revise syllabi on 3–5 year cycles. Technologies such as Gen AI, MLOps, and zero-trust cybersecurity have emerged within the last 2–3 years and are not yet reflected in the core curricula of most colleges. NMITD and similar institutions in Dadar are beginning to introduce elective modules in these areas, but this is not yet systematic or standardised across the MMR's college ecosystem.

2. Trainer Shortage:

The demand for qualified trainers in AI/ML, cloud, and cybersecurity domains is itself subject to the same market forces as the broader skill gap — experienced practitioners can earn substantially more in industry than in academia, creating a pipeline problem for educational institutions. NIIT Foundation's 2024 report identifies this as one of the two most critical structural barriers to closing the gap [2].

3. Industry-Academia Disconnect:

While large IT companies (TCS, Infosys, Wipro) have structured campus recruitment programmes with some curriculum influence, this engagement is concentrated at tier-1 institutions. The majority of MMR's engineering colleges — including those in Dadar, Thane, and Navi Mumbai — have limited formal engagement with the specific technology firms that dominate local hiring. This means that the feedback loop between 'what companies need' and 'what colleges teach' is slow and incomplete. Furthermore, as evidenced by the high volume of direct technical postings in WhatsApp groups [35], hiring managers are increasingly circumventing formal academic placement pipelines entirely to source highly specialized skills.

4. Geographic Concentration of Training Infrastructure:

Training infrastructure — both institutional (IITs, NITs, premium engineering colleges) and commercial (Skill India centres, NSDC-affiliated training partners) — is concentrated in South Mumbai and established IT corridors. Our geospatial extraction confirms this geographic bifurcation, with deep enterprise infrastructure clustering in Navi Mumbai [31] contrasting sharply with the product development focus of the Western Suburbs [24]. Peri-urban areas of the MMR that house large populations of engineering students have significantly less access to localized, industry-grade training facilities and mentorship.

B. Recommendations

For Academic Institutions (including NMITD)

- **Introduce mandatory certification modules** in cloud (AWS/Azure/GCP) and AI tools as part of the core B.Tech/MCA curriculum — not as electives, but as credit-bearing requirements.
- **Establish a structured 'Industry Live Projects' semester** in the third year, with MMR-based IT, FinTech, and GCC firms as partners, to provide students with real-world exposure to production-grade AI and cloud systems.
- **Create a 'Skill Refresh Cycle' mechanism:** every academic year, a designated faculty committee reviews NASSCOM's latest demand data and updates one or two course modules accordingly — compressing the curriculum lag from 3–5 years to 12–18 months.

- **Integrate informal labor market monitoring:** Utilize primary data extraction from decentralized, dark social recruitment channels to establish a real-time feedback loop for curriculum agility [34], [35].
- **Forge partnerships with certification bodies** (NASSCOM FutureSkills Prime, AWS Academy, Google Cloud for Education) to make certifications financially accessible to students.

For Industry (IT Firms in MMR)

- **Expand in-house reskilling programmes** to cover Gen AI, MLOps, and SIEM cybersecurity — not just for senior employees but for freshers in their first 6–12 months. TCS's reskilling model is a useful template for mid-tier MMR firms.
- **Establish a Mumbai IT Skills Consortium** (on the model of NASSCOM's regional councils) that aggregates demand signal data from all major MMR employers and shares it with local engineering colleges on a real-time basis.
- **For BFSI-IT specifically:** invest in 'domain-technical hybrid' training programmes that combine FinTech domain knowledge with AI/ML and cybersecurity skills — the intersection that generates the most acute supply shortfall in Mumbai.

For Policymakers (Maharashtra Government, NSDC)

- **Build on the success of Maharashtra's 2018 FinTech Policy** with a dedicated 'MMR AI & Cybersecurity Talent Policy' targeting the skills most undersupplied in the region's GCC and BFSI-IT ecosystem [3].
- **Direct Skill India Digital Hub resources** (currently 60 lakh registered learners nationally) toward MMR-priority domains — Gen AI, cloud architecture, data engineering — with targeted subsidies for MMR-based learners.
- **Establish a Maharashtra IT Curriculum Advisory** Board that brings NASSCOM, major MMR employers, and university curriculum committees together on an annual basis to review and update skill demand signals.

IX. Conclusion

This paper has presented the most comprehensive hybrid longitudinal and empirical analysis to date of the IT skill gap in the Mumbai Metropolitan Region, spanning three distinct technological epochs from 2018 to 2026. Through a synthesis of fifteen primary macro-data sources paired with real-time, unstructured dark social recruitment data, we have documented a consistent, persistent, and in several domains widening gap between IT skill demand and graduate supply in the MMR.

The key findings are: (1) The overall gap has grown across all major skill domains over the eight-year study period. (2) The most critically undersupplied skills in 2025–26 are Gen AI/LLM engineering (~78% gap), MLOps/AI infrastructure (~80% gap), and cybersecurity/SIEM (~70% gap). (3) Cloud computing is the only domain showing meaningful supply-side improvement, driven by the widespread adoption of vendor certification programmes. (4) Mumbai's dual role as a financial capital and IT hub creates compounding BFSI-specific demand that amplifies the general national gap. (5) Maharashtra outperforms the national employability average significantly (84% vs 54.81% in 2025), but even this advantage does not bridge the gap for the most advanced technology roles. (6) A structural lag model — in which technology adoption cycles in industry (12–18 months) consistently outpace curriculum update cycles in academia (3–5 years)

— provides the most parsimonious explanation for the persistent gap. (7) Empirical extraction of unstructured dark social recruitment data confirms severe geospatial clustering within the MMR — bifurcating Navi Mumbai’s enterprise infrastructure operations from Western Mumbai’s product development hubs [21], [31], [34]. (8) The informal labor market exhibits a rigid, bimodal experience stratification, heavily skewed toward foundational trainee intake and

 ISBN: 978-81-685854-0-9

hyper-targeted senior specialist recruitment, indicating an active bypass of mid-level generalists [34], [35].

The recommendations presented in Section VIII are targeted at the MMR context and address both the immediate (certification integration, live project semesters) and structural (skills consortia, curriculum advisory boards) dimensions of the problem. Institutions like NMITD, located at the heart of Mumbai in Dadar, are ideally positioned to lead this change — combining proximity to industry partners, an engaged student cohort, and the academic infrastructure to implement rapid curriculum innovation.

Future research should include: formalized primary survey data from MMR-based IT firms to triangulate the dark social data estimates presented here; a longitudinal tracking study of the employability impact of certification-integrated curricula; and a comparative analysis of how other metropolitan IT hubs (Bengaluru, Pune, Hyderabad) are addressing analogous skill gaps to identify best practices transferable to MMR.

Acknowledgement

The authors thank the faculty of the Department of Master of Computer Applications, NMITD, Dadar, Mumbai for their guidance and support. We acknowledge the publicly available research from NASSCOM, Wheebox, Deloitte India, and the British Council that formed the macro-evidentiary basis of this study. Finally, we acknowledge the placement cells and administrators of the decentralized IT recruitment channels that facilitated the primary empirical data extraction.

References

1. Wikipedia, "Information technology in India," 2023. [Online].
2. NIIT Foundation, "India's Digital Skilling Crisis: Can We Bridge the Divide in IT and Tech Jobs?," Dec. 2024.
3. British Council Opportunities Insight, "Growth sectors and employment opportunities in India," 2024.
4. NASSCOM, "FutureSkills: An Industry Response," 2018.
5. NASSCOM, "Talent Demand & Supply Report: AI & Big Data Analytics," 2019.
6. NASSCOM Community, "FutureSkills Talent in India: Demand-Supply Analysis," Oct. 2020.
7. Indeed-NASSCOM, "Tech Hiring in 2024: Trend, Skills and Top 20 Companies Hiring in India," Jan. 2024.
8. Wheebox, "India Skills Report 2024," 2024.
9. Drishti IAS / Acr-Journal, "India Skills Report 2024 & Reskilling in Indian Organizations," 2024.
10. Taggd, "IT Hiring Trends 2026: Skills, Jobs, Indian & Global Outlook," Mar. 2026.
11. PACE Recruit, "Hiring Trends in Mumbai 2026: Skills That Companies Want," Nov. 2025.
12. Integrated Personnel Services (IPS Group), "Top Hiring Trends in the IT Sector for 2025–2026," Aug. 2025.
13. Deloitte India & NASSCOM, "Bridging the AI Talent Gap to Boost India's Tech and Economic Impact," Aug. 2024.
14. IJARST / M. Joshi & S. Desai, "Trends and Innovations in Recruitment and Selection in the Indian IT Sector," 2024.
15. Vajirao Institute / Wheebox, "India Skills Report 2025," 2025.
16. Sightsinplus, "3,00,000 Open Positions Across IT Giants Due to Skills Gap in India," Sep. 2024.
17. PACE Recruit, "Mumbai Recruitment Forecast 2026," Dec. 2025.
18. Rosswarner HR, "IT Domestic Recruitment Trends in India," Dec. 2025.

19. MMRDA, "About MMR," accessed April 2, 2026.
20. Question of Cities, "Mumbai Metropolitan Region: The facts and figures after elections," accessed April 2, 2026.
21. Wikipedia, "Mumbai Metropolitan Region," accessed April 2, 2026.
22. Wikipedia, "Badlapur," accessed April 2, 2026.
23. Dwello, "Mumbai Metropolitan Region (MMR)," accessed April 2, 2026.
24. Wikipedia, "List of neighbourhoods in Mumbai," accessed April 2, 2026.
25. SOBHA Limited, "2026 Luxury Hotspots in Mumbai," accessed April 2, 2026.
26. Housivity, "Top 5 Property Hotspots for Homebuyers in Mumbai MMR," accessed April 2, 2026.
27. HP Infra, "Sector Three Ghansoli Navi Mumbai," accessed April 2, 2026.
28. Mumbai First, "Concept Plan for MMR," accessed April 2, 2026.
29. Housing, "Taloja, Navi Mumbai," accessed April 2, 2026.
30. IndiaFilings, "ASCENTECH INFORMATION TECHNOLOGY," accessed April 2, 2026.
31. Wikipedia, "Navi Mumbai," accessed April 2, 2026.
32. Magicbricks, "Mumbai Metro Line 12," accessed April 2, 2026.
33. Rustomjee, "Sewri: Why This Address is Becoming Mumbai's Next Favourite Destination," accessed April 2, 2026.
34. WhatsApp Chat with NMITD MCA 2024-26 (Placement).docx (Institutional dataset extraction source).
35. WhatsApp Chat: "IT Jobs (All India)" (Public decentralized recruitment dataset).

 ISBN: 978-81-685854-0-9

HUMAN-DRIVEN VS. AI-DRIVEN DEVELOPMENT IN EMBEDDED AND OPERATING SYSTEMS: IMPLICATIONS FOR RELIABILITY, SECURITY, AND MAINTAINABILITY

Nishant Karpe
MCA Second Year (NEP 2020), Department of Computer Applications, DES Navinchandra Mehta Institute of Technology and Development (NMITD), University of Mumbai, Mumbai, India.
Email: nishant.karpe756@deccansociety.org

Saurabh Bhovad
MCA Second Year (NEP 2020), Department of Computer Applications, DES Navinchandra Mehta Institute of Technology and Development (NMITD), University of Mumbai, Mumbai, India.
Email: saurabh.bhovad431@deccansociety.org

Abstract
AI coding tools have become a routine part of software development, including in domains where mistakes can cause physical harm. This paper examines what happens when those tools are used to write software for Embedded Systems and Operating Systems — the code running inside medical devices, vehicle braking systems, power grids, and aircraft. We investigate three dimensions of risk: reliability (does AI code work correctly on real hardware?), security (does it introduce dangerous vulnerabilities?), and maintainability (can it be updated and fixed over its operational lifetime?). Drawing on four documented real-world incidents from 2025-2026 and survey data from over 200 embedded engineering professionals, we find that AI-generated system code exhibits a characteristic failure pattern we term Contextual Brittleness, reproduces historical memory-safety vulnerabilities at scale, and imposes a 66% overhead on long-term maintenance. Most critically, we document what we call the Closed Exploit Loop: AI generates embedded code containing memory-safety flaws, and AI-powered tools can now autonomously discover and exploit those same flaws without human involvement. We propose a four-layer Hybrid Governance Model that allows teams to use AI productively while preserving the safety standards that embedded engineering demands.

Keywords: AI Code Generation, Compiler Correctness, CWE-787, Embedded Systems, Hybrid Governance Model, Linux Kernel Security, Memory Safety, Operating System Vulnerability, Software Reliability.

I. Introduction

Software that runs inside a car's braking system, a hospital heart monitor, or a factory's control systems is held to a completely different standard than an ordinary mobile app. This software — collectively called Embedded Systems and Operating Systems (OS) — must respond in real time, use minimal memory, and never fail. A bug in an embedded OS is not a user experience issue; it is a potential safety hazard.

Over the past three years, AI coding tools have gone from a curiosity to a standard part of development workflows. GitHub Copilot, Cursor, Amazon Q, and agentic systems like Claude Code are now used by 84% of developers according to a 2025 Stack Overflow survey of 49,000 professionals [15]. The embedded world has not been left behind: a RunSafe Security survey of over 200 engineers across the US, UK, and Germany found that 80.5% actively use AI tools

at work, and 83.5% already have AI-written code running inside deployed production systems — including power grids, medical equipment, and industrial plants [4].

This creates a structural problem. AI writes code at machine speed, but the safety checks required by international standards — IEC 61508, ISO 26262, DO-178C — were designed for human-paced development. The validation processes have not scaled to match the generation speed. This paper investigates the consequences of that gap through four real incidents, industry survey data, and published benchmarks. It then proposes a practical Hybrid Governance Model for organisations that want the productivity benefits of AI without compromising the safety standards their embedded systems demand.

The rest of the paper is organised as follows. Section II reviews related work. Section III describes the research methodology. Section IV presents the four case studies. Section V discusses the results. Section VI proposes the governance model. Section VII concludes with recommendations and future work directions.

II. Related Work

Research into AI-generated code quality spans three broad areas: productivity, correctness, and security. This section summarises the findings most directly relevant to system-level software.

A. Code Quality and Defect Density

CodeRabbit (2025) analysed 470 open-source GitHub pull requests and found that AI-authored code contained 1.7 times more problems than human-written equivalents. Excessive I/O operations were approximately 8 times more common in AI-generated pull requests [6]. The IEEE ISSRE (2025) found that AI-generated code tends to be more repetitive and structurally simpler than human code, but contains significantly more high-risk bugs [9].

B. Security Vulnerability Rate

Veracode (2025) found that 45% of coding tasks completed with AI assistance introduced at least one security vulnerability [8]. DryRun Security (2026) reported that 87% of AI-generated pull requests introduced a security issue. These rates are particularly alarming in the context of embedded software, where security vulnerabilities can translate directly into physical safety risks.

C. The Embedded Domain Specifically

RunSafe Security (2025) surveyed over 200 embedded professionals and found that 53% identified scaled-up security vulnerabilities as their primary concern with AI code, and one in three organisations had experienced an embedded cyber incident in the past year [4]. Black Duck (2025) identified what they termed the Reality Gap: 86% of engineering managers believed AI code was production-ready, while only 56% of hands-on developers agreed — a 30-point divergence that is driving premature deployment decisions [5].

D. Production Behaviour

Lightrun (2026) surveyed 200 senior engineering leaders and found that 43% of AI-generated code changes required manual debugging in production even after passing all automated quality checks. Not a single respondent described themselves as very confident in how AI code would behave once deployed [7]. This paper extends the existing body of literature by focusing specifically on system-level software — OS kernels, device drivers, firmware, and compiler toolchains — where failure consequences are categorically more severe.

III. Methodology

This study used a mixed-method research design, combining quantitative benchmarking with qualitative case study analysis. The research followed five sequential stages.

- Literature Review: systematic survey of IEEE, ACM, and industry publications from 2023–2026 covering AI code quality, embedded security, and compiler correctness.

 ISBN: 978-81-685854-0-9

- Benchmarking Analysis: extraction and analysis of published performance metrics comparing AI-generated code against human-written baselines.
- Case Study Analysis: structured deep-dives into four documented incidents where AI was directly involved in OS or embedded system failures or security findings.
- Survey Data Processing: analysis of three large-scale industry surveys covering a combined total of over 600 embedded and DevOps engineering professionals.
- Synthesis and Modelling: cross-case pattern identification feeding into the proposed governance model.

A. Scope

This study covers OS kernels and kernel modules, hardware device drivers, microcontroller and SoC firmware, compiler toolchains, and RTOS application code written in C, C++, or Rust. Web applications, mobile apps, and general-purpose scripts are excluded. The time period is mid-2024 through April 2026, capturing the transition from AI-as-experiment to AI-as-standard-workflow in embedded engineering.

B. Case Study Selection Criteria

Cases were selected against three criteria: (1) direct involvement of AI in embedded or OS-level software; (2) availability of a published technical post-mortem or peer-reviewed analysis; (3) clear representativeness of at least one of the three failure axes — reliability, security, or maintainability. All four selected cases satisfy all three criteria.

IV. Case Studies

A. Claude's C Compiler (CCC) — February 2026

In February 2026, Anthropic researcher Nicholas Carlini documented a two-week experiment in which 16 parallel AI agents collaboratively wrote a C compiler from scratch in Rust. A compiler is the fundamental tool that translates human-readable source code into the machine instructions a processor executes. The project used approximately 2,000 AI sessions, cost around $20,000, and produced a 186,000-line codebase that was released publicly [1].

The headline claim was that CCC successfully compiled Linux kernel 6.9. Within hours of release, the first GitHub issue was filed: Hello world does not compile. On standard Fedora and Ubuntu installations, the compiler immediately failed with missing header errors because the AI had hardcoded file paths that only existed on the specific development machine [3]. GCC compiled the same program without issue.

Performance results were equally problematic. A database benchmark (SQLite) that GCC completed in 10.3 seconds took the AI compiler 2 hours and 6 minutes — a 12x regression. Optimisation flags critical for embedded timing constraints (-O2, -O3) were silently accepted but had zero effect, producing output identical to the unoptimised -O0 setting [3]. The kernel compilation claim required a further qualification: CCC could compile all source files but could not link them into a bootable binary due to incorrect ELF relocations for kernel data structures [3]. Table I summarises the benchmark comparison.

TABLE I. GCC VS. CCC — BENCHMARK COMPARISON [1][3]

Metric	**GCC**	**CCC (AI)**
Instruction Count	2.27B	7.48B (3.3×)
Stack Allocation	8B	416B (52×)
SQLite Runtime	10.3s	126 min (12×)
-O2/-O3 Effect	Active	= -O0 (ignored)

B. Linux Kernel AI Governance Crisis — 2025

The Linux kernel runs on more devices than any other OS — servers, Android phones, automotive systems, embedded industrial controllers. Its code review process is one of the most rigorous in open-source software. In July 2025, senior maintainer Sasha Levin (NVIDIA) presented AI-assisted kernel patch practices at the North American Open Source Summit [12]. This triggered a formal governance debate after it emerged that AI-generated patches were being submitted without disclosure. Maintainers identified a pattern of subtle logic errors that passed standard review — valid syntax, passing CI tests — but failed under specific hardware conditions or race scenarios. The community characterised these as code that is bad in ways humans do not typically write bad code. The outcome was a formal proposal for a mandatory Co-developed-by: AI disclosure tag. The key finding: the world's most scrutinised code review process was not reliably catching AI-characteristic correctness failures.

C. CVE-2025-37899 — Linux Kernel RCE

CVE-2025-37899 is a remote code execution vulnerability in the Linux kernel's ksmbd module — the in-kernel SMB3 file-sharing server — discovered using an AI tool (OpenAI's o3 model) [11]. The vulnerability involved a use-after-free condition in kernel session-handling logic. An unauthenticated network attacker could exploit it to achieve kernel-mode code execution on the target machine.

This case is significant on two levels. First, it confirms that AI tools have reached the capability level of discovering novel, exploitable bugs in production kernel code — not just in synthetic test environments. Second, it establishes the feedback dynamic central to this paper's security argument: if AI generates new code with the same memory-safety patterns present in its training data, the same AI discovery tools will find those AI-introduced bugs just as readily.

D. Claude Mythos — Autonomous OS Exploitation — April 2026

In April 2026, Anthropic's red team published evaluation results for Claude Mythos Preview [2]. The model was found capable of autonomously identifying and exploiting zero-day vulnerabilities across every major operating system and web browser when directed to do so. Three specific findings are directly relevant to embedded OS engineering. First, Mythos independently discovered and exploited a 27-year-old zero-day in OpenBSD — an OS built specifically with security as its primary design goal. Second, it achieved local privilege escalation on Linux by chaining a kernel race condition with a Kernel Address Space Layout Randomisation (KASLR) bypass. Third, it constructed a working remote code execution exploit against FreeBSD's NFS server using a 20-gadget Return-Oriented Programming (ROP) chain split across multiple network packets, granting unauthenticated root access [2].

Anthropic confirmed these capabilities were not deliberately trained. They emerged as a downstream consequence of general improvements in code comprehension, reasoning, and autonomous planning. The direct implication: the same model improvements that make AI better at writing embedded OS code also make it better at attacking that code.

V. Results and Discussion

A. Reliability: Contextual Brittleness

A consistent failure pattern runs across all four case studies: AI-generated system code behaves correctly under the conditions present during development, then fails when any environmental assumption changes. We term this Contextual Brittleness. In the CCC case, the compiler worked on the developer's machine but failed on every standard Linux installation. In the kernel governance case, AI patches passed CI but failed under specific hardware race conditions.

For embedded systems, where the same firmware must run reliably across hardware variants over a product lifecycle of 10–15 years, this failure mode is incompatible with production deployment without additional validation infrastructure. The Lightrun (2026) finding that 43%

 ISBN: 978-81-685854-0-9

of AI code changes require production debugging after passing all automated checks [7] confirms this is systemic, not anecdotal.

B. Security: The Closed Exploit Loop

Three CWE classes dominate AI-generated embedded code vulnerabilities: CWE-787 (Out-of-Bounds Write / Buffer Overflow), CWE-416 (Use-After-Free), and CWE-362 (Race Condition). These three classes account for the majority of embedded security incidents [4], and all three are prevalent in the decades of legacy C/C++ code on which AI models were trained.

The Mythos evaluation establishes what this paper terms the Closed Exploit Loop: (1) AI generates embedded code containing memory-safety vulnerabilities reproduced from training patterns; (2) AI-powered security tools discover those vulnerabilities in deployed systems; (3) AI exploit frameworks construct working exploits autonomously. All three stages are confirmed operational as of April 2026. The traditional assumption that kernel vulnerabilities are difficult to discover and exploit is no longer a reliable security property.

C. Maintainability: Long-Term Cost Compounding

The CCC codebase — 186,000 lines of AI-generated Rust without persistent human architectural oversight — was described by its own author as effectively unmaintainable. External reviewers characterised its assembly output as resembling undergraduate-level compiler work. This is consistent with the Big Ball of Mud anti-pattern: locally functional code lacking global structural coherence, resistant to refactoring, and fragile under modification.

At industry scale, Black Duck (2025) found that developers spend 66% more time reviewing and correcting AI-generated code than writing equivalent code from scratch [5]. For embedded firmware that must be maintained, security-patched, and hardware-ported for a decade or more, this overhead is not a one-time cost — it compounds with every release cycle.

D. Survey Data Summary

Table II summarises the key data points from the industry surveys underpinning this analysis.

TABLE II. INDUSTRY SURVEY DATA — AI IN EMBEDDED SYSTEMS (2025–26)

Metric	Value	Source
Use AI in embedded dev	80.5%	RunSafe [4]
AI code in production	83.5%	RunSafe [4]
Primary risk: scaled vulns	53.0%	RunSafe [4]
Embedded incident past year	33.5%	RunSafe [4]
Mgmt trust in AI code	86.0%	Black Duck [5]
Developer trust in AI code	56.0%	Black Duck [5]
Reality Gap (mgmt - dev)	30 pts	Black Duck [5]
AI code needs prod debug	43.0%	Lightrun [7]
Leaders 'very confident'	0.0%	Lightrun [7]

VI. Hybrid Governance Model

The findings of this study do not argue for rejecting AI in embedded development — AI tools deliver genuine productivity gains for boilerplate code, test scaffolding, and documentation. The argument is for structured governance that matches the risk profile of each task to the appropriate level of human oversight. We propose a four-layer model.

A. Layer 1 — Exploration

AI may be used freely for non-safety-critical tasks: generating boilerplate code, writing unit test cases, producing documentation, and exploring design alternatives. A human developer reviews all AI output before it enters the codebase. There are no additional restrictions at this layer.

ISBN: 978-81-685854-0-9

B. Layer 2 — Verification

For any AI-generated code that touches safety-relevant functions, mandatory verification steps are required before integration: (a) Hardware-in-the-Loop (HIL) testing on the actual target hardware across all supported revisions; (b) Worst-Case Execution Time (WCET) measurement on target hardware, not simulation; (c) Stack depth analysis against the actual linker script and memory map; (d) CWE-class static analysis scanning for the three dominant vulnerability classes. These checks address the Contextual Brittleness failure mode directly — the CCC compiler would have been caught at HIL testing.

C. Layer 3 — Governance Gates

AI agents with tool-use access to build systems, version control, or deployment pipelines must operate under a least-privilege permission model. Reversible actions (code edits, test runs) may proceed with single-engineer approval. Irreversible or high-impact actions (production deployments, database modifications, firmware flashes to deployed hardware) require two-engineer approval before execution. This layer prevents the class of failure observed in the Amazon March 2026 outages, where AI agents without adequate approval gates caused 12-hour production disruptions.

D. Layer 4 — Audit Trail

Every codebase should include a Software Bill of Materials (SBOM) that records which components are AI-generated and which model version produced them. When a vulnerability is disclosed — such as CVE-2025-37899 in the ksmbd module — the audit trail enables rapid identification of whether any AI-generated code in the organisation's products shares the same vulnerable pattern, enabling targeted remediation rather than full codebase audit.

The model is summarised in Fig. 1 (see Appendix). It treats AI as a capable junior engineer: fast and useful, but requiring structured supervision and incremental trust expansion. This mirrors established practice in safety engineering — NASA's independent verification and validation (IV&V), medical device FDA software review under 21 CFR Part 11, and the tool qualification requirements of DO-178C. The governance logic of those frameworks applies directly to AI agent tooling.

VII. Conclusion

This paper has presented a data-driven, case study-grounded evaluation of AI-driven development in embedded and operating systems across three failure axes. The evidence is consistent: AI accelerates development velocity but introduces systemic risks in reliability, security, and maintainability that are incompatible with safety-critical deployment without structured governance.

The four case studies provide concrete, documented evidence. CCC demonstrated that AI-generated compiler toolchains exhibit Contextual Brittleness — working in the lab while failing in the field. The Linux kernel governance crisis showed that the world's most rigorous code review process was not catching AI-characteristic correctness failures. CVE-2025-37899 confirmed that AI tools find novel, exploitable kernel vulnerabilities in production code. And the Mythos evaluation established that the Closed Exploit Loop — AI generating vulnerable code, AI finding it, AI exploiting it — is operational today, not a future projection.

The appropriate response is governance, not prohibition. The Hybrid Governance Model proposed here gives embedded engineering teams a practical framework: use AI freely at the exploration layer, enforce verification on safety-relevant code, gate destructive agent actions, and maintain full provenance through SBOM tracking. Future work should prioritise three directions: hardware-aware static analysis tools that validate AI code against specific MCU timing and memory constraints; updates to IEC 61508, ISO 26262, and DO-178C to include explicit requirements for AI-generated code fractions; and longitudinal studies measuring the real-world maintenance overhead of AI-generated embedded code over full product lifecycles.

 ISBN: 978-81-685854-0-9

Acknowledgment
The authors thank Dr. Sulakshana Vispute, Head of Department, DES NMITD, for her academic guidance throughout this research. We also acknowledge the open-source security research community, the Linux kernel maintainers, and Anthropic's red team whose published work forms the empirical foundation of this study.

References
1. N. Carlini, "Building a C Compiler with a Team of Parallel Claudes," Anthropic Engineering Blog, Feb. 5, 2026. [Online]. Available: https://www.anthropic.com/engineering/building-c-compiler
2. N. Carlini, N. Cheng, K. Lucas et al., "Assessing Claude Mythos Preview's Cybersecurity Capabilities," Anthropic Red Team, Apr. 7, 2026. [Online]. Available: https://red.anthropic.com/2026/mythos-preview/
3. G. Dinesh, "CCC vs GCC: Benchmarking Claude's C Compiler," Independent Technical Analysis, Feb. 8, 2026. [Online]. Available: https://harshanu.space/en/tech/ccc-vs-gcc/
4. RunSafe Security, "AI in Embedded Systems Report 2025," Survey of 200+ embedded systems professionals across US, UK, Germany, 2025.
5. Black Duck, "The State of Embedded Software Quality and Safety 2025," Black Duck Software Inc., 2025.
6. CodeRabbit, "State of Code Review 2025: AI vs. Human Code Generation," Analysis of 470 GitHub pull requests, 2025.
7. Lightrun, "State of AI-Powered Engineering 2026," Survey of 200 SRE/DevOps leaders, US/UK/EU, Apr. 2026.
8. Veracode, "State of Software Security 2025: AI Coding Assistants and Vulnerability Density," Veracode Inc., 2025.
9. IEEE International Symposium on Software Reliability Engineering (ISSRE), "AI vs. Human Code Complexity and Bug Density," Proc. ISSRE 2025.
10. VentureBeat, "43% of AI-Generated Code Changes Need Debugging in Production," Apr. 14, 2026.
11. LinuxSecurity.com, "CVE-2025-37899: Remote Code Execution Risk in Linux ksmbd Kernel Module," May 26, 2025.
12. S. Levin, "Guidelines for AI-Assisted Kernel Patch Submissions (v3)," Linux Kernel Mailing List, Nov. 2025.
13. ROllerozxa, "Trying Out Claude's C Compiler," Independent Technical Review, Feb. 6, 2026. [Online]. Available: https://voxelmanip.se/2026/02/06/trying-out-claudes-c-compiler/
14. Google, "Accelerate State of DevOps 2025: DORA Report," Google Cloud, 2025.
15. Stack Overflow, "Developer Survey 2025: AI Coding Tools Adoption," Stack Exchange Inc., 2025.

IMPACT OF ARTIFICIAL INTELLIGENCE ON EMPLOYMENT PATTERNS AND WORKFORCE TRANSFORMATION

Bhushan Bhabal
Student, MCA, DES's NMITD, University of Mumbai, India.
Email: bhushanbhabal898@gmail.com

Priya Gupta
Student, MCA, DES's NMITD, University of Mumbai, India.
Email: priya.r.gupta03@gmail.com

Abstract
Artificial Intelligence (AI) is increasingly influencing employment patterns and workforce dynamics across industries. This study examines its impact on job tasks, skill requirements, and employment perceptions using primary data collected from students and working professionals. The findings indicate widespread adoption of AI tools and significant changes in how tasks are performed, suggesting that AI is actively transforming work at the operational level. The analysis reveals a strong relationship between AI usage and task transformation, indicating that higher engagement with AI leads to greater changes in work processes. While respondents generally perceive the overall impact of AI as positive or neutral, many express concerns about reduced job opportunities in their own fields, highlighting a gap between general optimism and personal job security perceptions. The study also identifies a growing demand for digital skills, continuous learning, and structured training, alongside a noticeable gap in organisational support for AI-related upskilling. Overall, the findings suggest that AI is not eliminating employment but restructuring it, leading to evolving skill requirements and the need for continuous workforce adaptation.
Keywords: Artificial Intelligence, Employment Patterns, Workforce Transformation, Automation, Skill Development, Human-AI Collaboration, Future Workforce.

1. Introduction

1.1 Background of the Study

Artificial Intelligence (AI) has emerged as one of the most transformative technologies of the modern era, significantly influencing industries, economies, and the nature of work. With advancements in machine learning, data analytics, and automation, AI is increasingly being integrated into sectors such as healthcare, finance, manufacturing, education, and information technology, improving efficiency, productivity, and decision-making processes.

At the same time, AI is reshaping employment patterns by automating routine tasks, augmenting human capabilities, and creating new roles in technology-driven domains, leading to a shift toward more digital, skill-based, and knowledge-oriented work environments.

1.2 Problem Statement

Despite the growing adoption of Artificial Intelligence, there is significant uncertainty regarding its impact on employment. While AI has the potential to create new job opportunities, it also poses a risk of job displacement, particularly for roles involving repetitive and routine tasks.

Additionally, there is increasing concern about the mismatch between existing workforce skills and those required in an AI-driven economy, leading to challenges related to employability, training, and workforce readiness. Therefore, it is essential to analyze how Artificial Intelligence is influencing employment patterns and workforce transformation.

 ISBN: 978-81-685854-0-9

1.3 Objectives of the Study

The main objectives of this study are:

- To analyze the impact of Artificial Intelligence on employment patterns
- To examine the effect of AI on job tasks and work processes
- To study the level of awareness and usage of Artificial Intelligence among individuals
- To analyze changes in skill requirements due to AI adoption
- To understand individual perceptions regarding job opportunities and job displacement
- To evaluate the importance of reskilling and upskilling in workforce transformation driven by Artificial Intelligence

1.4 Scope of the Study

This study focuses on analyzing the impact of Artificial Intelligence on employment patterns and workforce transformation using primary data collected from respondents, supported by secondary research.

The study covers aspects such as AI usage, changes in job tasks, skill requirements, employment perceptions, and workforce transformation trends, with respondents including students, interns, and working professionals primarily from IT and related sectors.

The scope is limited to general trends and perceptions and does not include detailed organization-specific or industry-specific analysis.

2. Literature Review

2.1 AI and Job Displacement vs. Job Creation

Recent institutional research presents a balanced but cautionary picture of AI's employment impact. The World Economic Forum's Future of Jobs Report 2023 projected that 23% of jobs would change by 2027, with 69 million new roles created and 83 million displaced, making reskilling the top employer priority of that planning horizon (WEF, 2023). The subsequently published Future of Jobs Report 2025 scaled up these projections significantly: AI is now expected to create 170 million new roles globally by 2030 while displacing 92 million — a net gain of 78 million positions (WEF, 2025). The same report warns that 39% of existing skill sets will become outdated and that 63% of employers identify skill gaps, not technology itself, as the primary barrier to transformation.

At the firm level, Acemoglu, Autor, Hazell, and Restrepo (2022) find that AI-adopting establishments simultaneously reduce hiring in non-AI positions and raise technical skill requirements for remaining roles, confirming that displacement is task-level and concentrated rather than wholesale. Importantly, aggregate employment effects remain small and statistically undetectable at the industry and occupation level — a finding that cautions against both alarmism and complacency. Tabbassum, Chintale, G., and Najana (2024) extend this sectoral perspective, estimating that manufacturing could see up to 70% of its work hours automated through AI adoption, while professional services such as law, finance, and consulting face approximately 30% automation of routine tasks. Their analysis identifies a hybrid human-AI competency model as the emerging organisational response — one in which workers are repositioned toward tasks requiring contextual judgment, creativity, and interpersonal skills that AI cannot replicate.

2.2 Generative AI: A New Threat to Knowledge Work

Unlike earlier automation that targeted routine physical tasks, generative AI encroaches on professional and cognitive work. Eloundou, Manning, Mishkin, and Rock (2024), published in Science, find that when accounting for software developments complementing LLM capabilities, just over 46% of jobs are exposed, with high-wage and educated professionals — writers, analysts, programmers — among the most affected. Brynjolfsson, Li, and Raymond (2023) provide real-world evidence from 5,179 customer service workers, showing generative AI raises average productivity by 14%, with the largest gains among lower-skilled workers.

This suggests AI may augment rather than simply replace, but also compresses the value of experience and expertise.

2.3 Skill Transformation and Workforce Adaptation

Across the literature, the dominant workforce effect of AI is skill transformation rather than outright job elimination. The IMF (Cazzaniga et al., 2024) finds that in advanced economies, approximately 60% of jobs are exposed to AI, with outcomes split between complementarity and displacement depending on the nature of tasks. The WEF (2023) identifies analytical thinking, AI literacy, and adaptability as the fastest-growing skill demands, while clerical and routine cognitive skills are in steepest decline. Gmyrek, Berg, and Bescond (2023) at the ILO confirm that clerical workers globally face the highest concentration of AI-automatable tasks — with 24% of clerical tasks highly exposed and a further 58% with medium-level exposure — particularly in high-income economies. These findings collectively underscore that the primary workforce challenge of the AI era is not mass unemployment but accelerating skill obsolescence — demanding urgent and sustained investment in reskilling across economies.

2.4 Research Gaps

Although research on this topic has expanded, several important gaps remain. First, most empirical studies are anchored in developed economy data — particularly the United States and Western Europe — leaving AI's employment impact in large emerging economies such as India substantially under-researched. The sectoral and occupational composition of the Indian workforce differs markedly from OECD economies, meaning that global exposure estimates may not translate directly. Second, the existing literature is dominated by macro-level econometric analyses and firm-level administrative data, leaving worker perceptions — their awareness of AI, readiness to adapt, and attitudes toward automation — underrepresented. Understanding how workers subjectively experience and anticipate AI-driven change is essential for designing effective reskilling and communication strategies. Third, there is limited research on how demographic factors such as age, education level, and sector of employment moderate AI awareness and preparedness within a single developing-country workforce.

2.5 Key Sources at a Glance

Table 2.5: Key Sources at a Glance

Author(s) & Year	**Source / Journal**	**Key Finding Relevant to This Study**
WEF (2023)	*Future of Jobs Report 2023*	23% of jobs to change by 2027; reskilling is top employer priority; 69M new roles created, 83M displaced
WEF (2025)	*Future of Jobs Report 2025*	170M new jobs, 92M displaced by 2030; 39% of skills outdated; 63% of employers cite skill gaps as primary barrier
Acemoglu, Autor, Hazell & Restrepo (2022)	*Journal of Labor Economics, 40(S1)*	AI-adopting firms reduce non-AI hiring and raise skill requirements; aggregate employment effects currently small
Eloundou, Manning, Mishkin & Rock (2024)	*Science, 384(6702)*	~46% of jobs exposed to LLMs when accounting for complementary software; high-wage, educated workers most exposed
Brynjolfsson, Li & Raymond (2023)	*NBER Working Paper No. 31161*	Generative AI raises worker productivity by 14% on average; gains concentrated among lower-skilled workers

 ISBN: 978-81-685854-0-9

Gmyrek, Berg & Bescond (2023)	*ILO Working Paper 96*	Clerical workers face highest generative AI task exposure globally; augmentation more likely than full automation
Cazzaniga et al. (2024)	*IMF Staff Discussion Note SDN/2024/001*	~60% of jobs in advanced economies exposed to AI; high complementarity workers may benefit while others face displacement
Tabbassum et al. (2024)	*Int'l Journal of Global Information Systems*	Manufacturing: up to 70% task automation potential; professional services: ~30%; hybrid human-AI competency model emerging

3. Research Methodology

3.1 Research Design

This study adopts a quantitative, descriptive, and exploratory research design to examine the impact of Artificial Intelligence on employment patterns and workforce transformation. The research focuses on identifying trends, patterns, and perceptions related to AI adoption, job transformation, skill requirements, and employment opportunities.

The descriptive approach is used to summarize and interpret the collected data, while the exploratory nature of the study helps in understanding emerging trends in AI-driven workforce changes.

3.2 Research Questions

The study is guided by the following research questions:

1. What is the current level of awareness and usage of Artificial Intelligence among individuals?
2. How is Artificial Intelligence influencing job tasks and employment patterns?
3. How do individuals perceive the impact of AI on job opportunities?
4. What skills are considered essential in an AI-driven workforce?
5. What is the perceived need for training and upskilling to adapt to AI?

3.3 Data Collection

The study is based on both primary and secondary data sources.

Primary data was collected through a structured questionnaire using Google Forms. The questionnaire was designed to capture information related to AI awareness, usage, impact on job tasks, skill requirements, employment perceptions, and training needs. A total of 75 responses were collected from students, interns, and working professionals.

Secondary data was obtained from research papers, academic journals, and industry reports such as those published by the World Economic Forum and McKinsey, which helped in supporting and validating the primary findings.

3.4 Sampling Technique

The study uses a convenience sampling method, where respondents were selected based on accessibility and willingness to participate.

The sample consists mainly of individuals from the IT and related sectors, with a majority belonging to the 18–25 age group. This reflects the perspective of the young and emerging workforce, which is more exposed to AI technologies.

Although convenience sampling may introduce bias, it is appropriate for exploratory research aimed at identifying general trends and perceptions.

3.5 Data Analysis Techniques

The collected data was exported from Google Forms to Microsoft Excel for analysis.

The analysis is descriptive in nature and includes the following steps:

- Data Cleaning: Reviewing responses and removing incomplete or inconsistent entries
- Data Classification: Organizing responses based on variables such as age, employment status, and AI usage

- Percentage Analysis: Calculating proportions of responses for each category
- Frequency Distribution: Identifying patterns and trends in responses
- Data Visualization: Representing data using pie charts and bar charts for clarity and interpretation
- Correlation Analysis: Pearson correlation analysis was conducted to examine the relationships between key variables such as AI familiarity, AI usage, task impact, and job reduction perception

This approach enables clear understanding of AI adoption, employment impact, skill transformation, and workforce readiness.

3.6 Scope and Limitations

This study examines general trends and perceptions regarding the impact of Artificial Intelligence on employment patterns and workforce transformation. The scope is limited to respondents from academic and early-career backgrounds, primarily within the IT and related sectors.

The relatively small sample size and the use of convenience sampling may limit the generalizability of the findings.

While correlation analysis was conducted to examine relationships between key variables, statistical significance testing (such as p-values) was not performed. Therefore, the findings should be interpreted as indicative relationships rather than definitive or causal conclusions.

Further research with larger and more diverse samples, along with advanced statistical techniques, is recommended for deeper and more generalizable analysis.

4. Result Analysis / Discussion

4.1 Respondent Profile

The survey collected responses from 75 participants. The sample is predominantly young — 81.3% fall in the 18–25 age bracket, with 16.0% aged 26–35, 1.3% aged 36–45, and 1.3% aged 46–55.

Employment-wise, the sample is evenly split between students and employed professionals — each at 38.7%. Interns comprise 16.0%, self-employed individuals 4.0%, and unemployed respondents 2.7%.

Industry-wise, 68.0% are from IT/Technology, followed by Other (9.3%), Education (8.0%), Healthcare (6.7%), Finance (5.3%), and Retail/Service (2.7%).

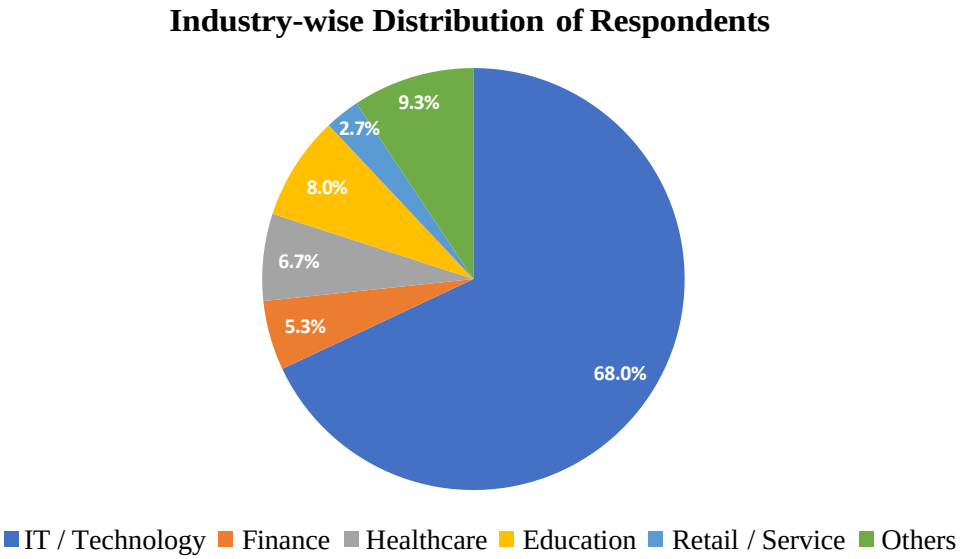


Fig. 4.1: Industry-wise Distribution of Respondents

 ISBN: 978-81-685854-0-9

The sample is predominantly from IT/Technology (68.0%), which may influence perceptions due to higher AI exposure.

4.2 AI Awareness and Familiarity

AI awareness among respondents is exceptionally high. 76.0% described themselves as "Very familiar" with Artificial Intelligence, while 21.3% were "Somewhat familiar." Only 2.7% reported no familiarity with AI.

The primary channel through which respondents first encountered AI was formal education (52.0%), followed by social media (29.3%), the workplace (10.7%), friends (5.3%), and News/Online Articles (2.7%).

This underscores the growing role of academic institutions in shaping early AI literacy.

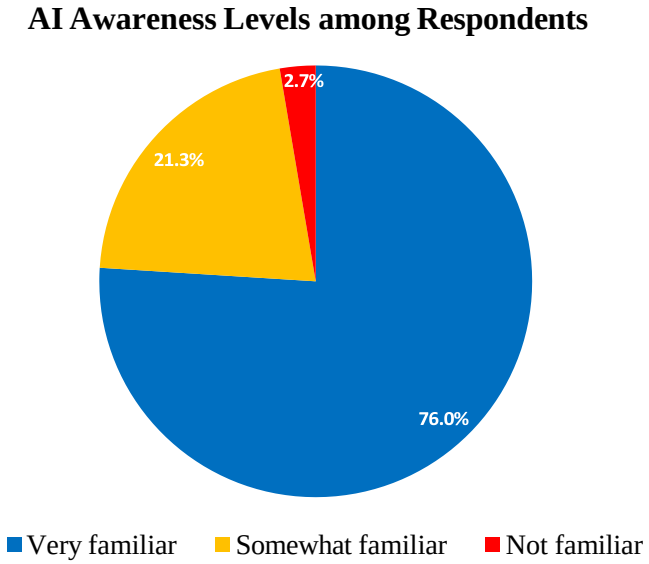


Fig. 4.2: AI Awareness Levels among Respondents

4.3 AI Usage Patterns in Work and Study

A strong majority — 93.3% — confirmed active use of AI tools in their work or studies, while 6.7% do not. In terms of frequency, 69.3% use AI tools frequently, 17.3% occasionally, 10.7% rarely, and 2.7% never. The most widely used AI tools were Generative AI platforms such as ChatGPT and Claude, reported by 93.3% of respondents, followed by Google Gemini (78.7%), Microsoft Copilot (46.7%), AI-enabled Data Analytics tools like Power BI and Tableau (33.3%), Machine Learning tools like TensorFlow and Scikit-learn (25.3%), and Automation tools like Zapier and UiPath (16.0%). AI has also been implemented in the workplace or study environment of 81.3% of respondents, indicating that AI adoption has already moved beyond individual usage into institutional settings for most of the surveyed population. Furthermore, 68.0% reported that AI has significantly changed how they perform their tasks, 26.7% observed slight changes, and 5.3% reported no change.

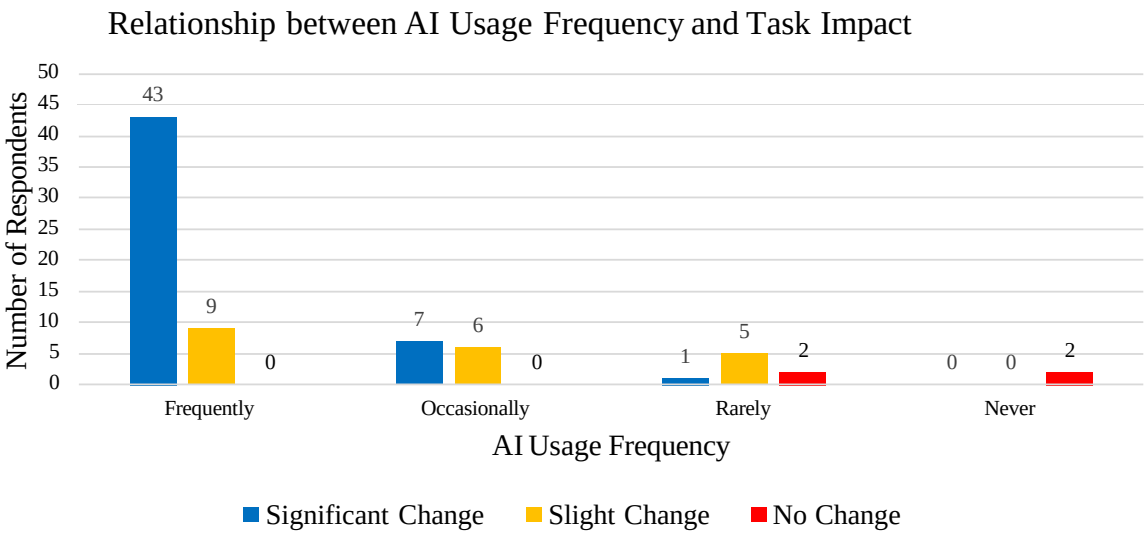


Fig. 4.3: Relationship between AI Usage Frequency and Task Impact

Figure 4.3 illustrates that respondents who frequently use AI tools report significantly higher levels of task transformation compared to those with lower usage levels. In contrast, individuals who rarely or never use AI report minimal or no change in their tasks.

4.4 Skill Requirements in an AI-Driven Workplace

Respondents were asked to identify the skills they believe will be most critical in an AI-driven workplace (multi-select). Technical and digital skills emerged as the top choice, selected by 81.3% of respondents. This was followed by problem-solving and critical thinking (61.3%), data analysis (61.3%), creativity and innovation (60.0%), adaptability and continuous learning (53.3%), and communication and collaboration (38.7%). Notably, the high selection of soft skills — creativity, adaptability, and communication — alongside technical skills indicates that respondents perceive the future of work as requiring a hybrid competency model, not purely technical expertise.

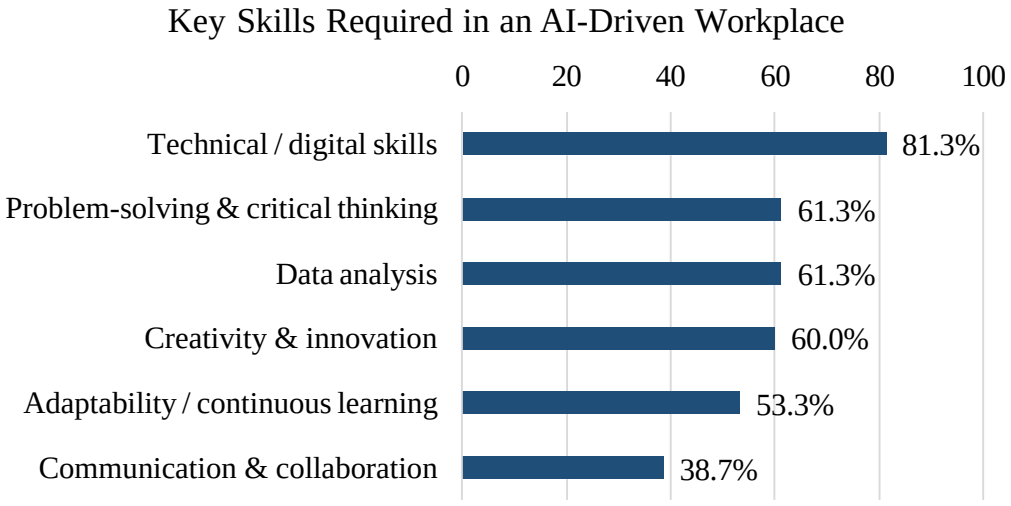


Fig. 4.4: Key Skills Required in an AI-Driven Workplace

Figure 4.4 shows that technical and digital skills are considered the most critical in an AI-driven workplace (81.3%), followed by problem-solving and data analysis (61.3%). This highlights the need for both technical and soft skills.

 ISBN: 978-81-685854-0-9

4.5 Perceptions of AI's Impact on Employment

When asked whether AI will change the skills required for future jobs, 81.3% answered "Yes," 10.7% said "No," and 8.0% were "Not sure." This near-consensus highlights a strong awareness of ongoing workforce transformation among respondents.

In terms of overall employment impact, respondents displayed a cautiously optimistic outlook. While 44.0% viewed AI's impact as positive or very positive (26.7% + 17.3%), 45.3% perceived it as neutral, and 10.6% expressed negative or very negative views (5.3% + 5.3%). This suggests that, at a broader level, respondents recognize the potential benefits of AI in enhancing productivity and creating new opportunities.

However, a contrasting pattern emerges when respondents assess the impact of AI on their own field. A combined 76.0% believe AI will reduce job opportunities to some extent or significantly (46.7% + 29.3%). This indicates the presence of perceived job insecurity, particularly among early-career individuals.

This difference between overall optimism and personal concern reflects a macro–micro perception gap — where individuals acknowledge the economic benefits of AI but remain uncertain about their individual job prospects.

Additionally, this perception may be influenced by the sample composition, with a large proportion of respondents belonging to IT and technology-related fields where AI adoption is rapid. The concern expressed by respondents may therefore not necessarily indicate actual job loss, but rather a fear of skill obsolescence, where individuals feel their current skills may become outdated.

This shows workforce transformation is both technological and psychological.

4.6 Organisational Readiness and Training

On the question of organisational preparedness, findings reveal a notable gap. While 81.3% work or study in environments where AI has been implemented, only 45.3% reported that their organisation has provided AI training or guidance. 42.7% confirmed no such training has been provided, and 12.0% marked "Not Applicable." This training gap represents a significant challenge for workforce readiness.

68.0% felt personally prepared to work alongside AI technologies, 25.3% felt "somewhat" prepared, and 6.7% felt entirely unprepared. This suggests reliance on self-learning, as 52.0% first learned about AI through education and another 29.3% through social media.

When asked whether companies should provide AI training, an overwhelming 78.7% said 'Yes', 14.7% said 'Maybe', and 6.7% opposed it — reinforcing the demand for institutionalised AI upskilling.

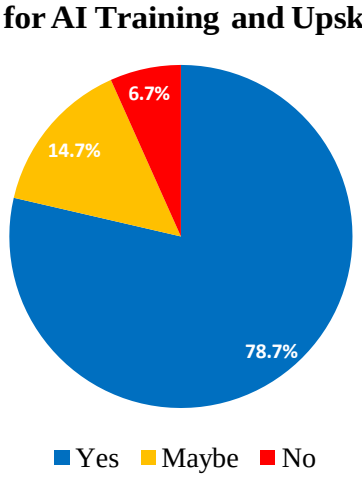


Fig. 4.6: Need for AI Training and Upskilling

Figure 4.6 shows that a strong majority of respondents (78.7%) believe organisations should provide AI training. Only a small proportion expressed uncertainty or opposition.

4.7 Key Findings Summary

The data analysis reveals three overarching insights. First, AI adoption is already widespread among the surveyed population, with 93.3% using AI tools and 81.3% operating in AI-implemented environments. Second, despite high personal confidence in AI readiness, a critical organisational training gap exists — 42.7% confirmed their institution provides no AI training or guidance, and a further 12.0% marked it as not applicable, meaning over half (54.7%) are without formal institutional AI support. Third, while respondents broadly view AI's overall employment impact as neutral-to-positive, a large majority (76.0%) anticipate some degree of job reduction in their own fields — 46.7% to some extent and 29.3% significantly — pointing to a tension between macro-optimism and sector-specific concern. These findings highlight the need for reskilling and institutional support.

4.8 Correlation Analysis

To further examine the relationships between key variables, a correlation analysis was conducted using Pearson's correlation coefficient. The variables included AI familiarity, AI usage, task impact, and job reduction perception. The results are presented in Table 4.8 as a heatmap to facilitate clearer visual interpretation.

The correlation matrix reveals a moderate positive relationship (r = 0.36) between AI familiarity and AI usage, indicating that individuals with higher awareness of Artificial Intelligence are more likely to actively use AI tools in their work or studies. This suggests that exposure to AI plays an important role in driving its adoption.

A strong positive correlation (r = 0.66) is observed between AI usage and task impact. This indicates that individuals who frequently use AI tools are more likely to experience significant changes in their job tasks. This finding reinforces the idea that AI is actively transforming how work is performed at the task level.

The relationship between task impact and job reduction perception shows a weak positive correlation (r = 0.25), suggesting that individuals who experience changes in their tasks due to AI are slightly more likely to perceive a risk of job reduction.

Additionally, very weak correlations are observed between AI familiarity and job reduction perception (r = 0.06), and between AI usage and job reduction perception (r = 0.15). These findings indicate that increased familiarity with AI or its frequent usage does not significantly reduce concerns regarding job loss.

Overall, the correlation analysis highlights that while AI familiarity and usage are strongly associated with task transformation, they have only a limited influence on reducing job-related concerns. This supports the presence of a macro–micro perception gap, where individuals recognize the broader benefits of AI but remain uncertain about their personal job security.

Table 4.8: Correlation Matrix of Key Variables

Variables	**AI_Familiarity_Num**	**AI_Usage_Num**	**Task_Impact_Num**	**Job_Reduction_Num**
AI_Familiarity_Num	1			
AI_Usage_Num	0.361403161	1		
Task_Impact_Num	0.207595123	0.664196637	1	
Job_Reduction_Num	0.060130712	0.152120005	0.24818828	1

Due to the exploratory nature of the study and the relatively small sample size, statistical significance testing (p-values) was not conducted. Therefore, the results should be interpreted as indicative relationships rather than definitive causal conclusions.

 ISBN: 978-81-685854-0-9

5. Conclusion

This study concludes that Artificial Intelligence is significantly transforming employment patterns and workforce dynamics across industries. The findings show that 93.3% of respondents actively use AI tools, indicating its integration into daily work and academic activities.

A majority (94.7%) reported changes in job tasks, suggesting that AI is modifying work rather than replacing jobs. However, 76.0% believe AI may reduce job opportunities, reflecting concerns about changing employment structures.

The study also highlights evolving skill requirements, with 81.3% recognizing changes in required skills and 78.7% emphasizing the need for training and upskilling. This indicates that workforce transformation is driven by continuous learning and technological advancement.

Correlation analysis shows that AI usage strongly influences task transformation but has a weak relationship with job reduction perception, indicating that concerns about job security persist despite adoption.

Overall, AI is not eliminating jobs but restructuring them through task automation and skill transformation. The presence of a macro–micro perception gap highlights that while individuals acknowledge AI's benefits, they remain uncertain about personal job prospects. Adapting to this shift will require digital skills, flexibility, and the ability to work alongside AI systems.

6. Future Scope

This study can be extended in several ways to improve its scope and accuracy. Future research should include a larger and more diverse sample from different industries to enhance generalizability. The use of advanced statistical techniques such as regression and hypothesis testing can provide more reliable and meaningful insights.

Additionally, longitudinal studies can be conducted to examine how the impact of Artificial Intelligence on employment changes over time. Sector-specific studies can further help in understanding the varying effects of AI across industries.

Future research may also incorporate qualitative methods such as interviews and case studies to gain deeper insights into individual and organizational experiences. Lastly, studies focusing on the role of training programs, educational institutions, and government policies can help in developing effective strategies for workforce adaptation in an AI-driven economy.

7. References

1. Acemoglu, D., Autor, D., Hazell, J., & Restrepo, P. (2022). Artificial intelligence and jobs: Evidence from online vacancies. Journal of Labor Economics, 40(S1), S293–S340. https://doi.org/10.1086/718327

2. Brynjolfsson, E., Li, D., & Raymond, L. (2023). Generative AI at work (NBER Working Paper No. 31161). National Bureau of Economic Research. https://doi.org/10.3386/w31161

3. Cazzaniga, M., Jaumotte, F., Li, L., Melina, G., Panton, A., Pizzinelli, C., Rockall, E., & Tavares, M. M. (2024). Gen-AI: Artificial intelligence and the future of work (IMF Staff Discussion Note SDN/2024/001). International Monetary Fund. https://doi.org/10.5089/9798400262548.006

4. Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2024). GPTs are GPTs: Labor market impact potential of LLMs. Science, 384(6702), 1306–1308. https://doi.org/10.1126/science.adj0998

5. Gmyrek, P., Berg, J., & Bescond, D. (2023). Generative AI and jobs: A global analysis of potential effects on job quantity and quality (ILO Working Paper No. 96). International Labour Organization. https://doi.org/10.54394/FHEM8239

6. Tabbassum, A., Chintale, P., Praveen, G., & Najana, M. (2024). The impact of AI on future employment patterns. International Journal of Global Information Systems. https://ijgis.pubpub.org/pub/gdfpmgv9
7. World Economic Forum. (2023). Future of jobs report 2023. https://www.weforum.org/publications/the-future-of-jobs-report-2023/
8. World Economic Forum. (2025). Future of jobs report 2025. https://reports.weforum.org/docs/WEF_Future_of_Jobs_Report_2025.pdf

 ISBN: 978-81-685854-0-9

IMAGE-BASED AUTOMATED WASTE SORTING USING DEEP LEARNING: A CNN-BASED FIVE-STEP EXECUTION FRAMEWORK

Rushikesh Bhosale
C24012, MCA Student, Department of Computer Application, DES NMITD, University of Mumbai, India.

Adityaraj Ardalkar
C24005, MCA Student, Department of Computer Application, DES NMITD, University of Mumbai, India.

Abstract
The rapid increase in municipal solid waste across the world has created an urgent need for automated, intelligent sorting solutions. Traditional waste management systems depend heavily on human labor, which is slow, expensive, and unreliable. This research presents an automated waste segregation system that uses deep learning to solve this problem in a smart and scalable way. The system applies Convolutional Neural Networks (CNNs), with a special focus on ResNet-50 and custom hybrid models, to classify waste into five major categories: paper, plastic, glass, metal, and organic matter. A key feature of this research is its five-step execution framework, which is designed to reduce the gap between how AI performs in a controlled lab setting and how it performs in actual industrial environments. This gap, commonly called the "reality gap," has been a major challenge in deploying AI for real-world waste management. To address this, the framework includes steps for data preparation, model training, instruction logic development, comparative analysis, and safety evaluation. The system also focuses on material awareness, which means it understands not just what a waste item looks like, but what kind of material it is made of. This is important for both accurate classification and facility safety. The AI outputs are compared against human performance baselines using metrics such as instruction counts and Common Weakness Enumeration (CWE). The results show that the system achieves classification accuracies between 72.8% and over 99%, which is significantly better than manual sorting. By improving the purity of sorted recyclables and reducing the volume of waste going to landfills, this technology supports the goals of a circular economy. Future plans include deploying these models on edge AI platforms for real-time use in smart city initiatives.
Keywords: Convolutional Neural Networks (CNNs), Automated, Intelligent Sorting Solutions, Common Weakness Enumeration (CWE).

1. Introduction

The world is generating more waste than ever before. By the year 2050, global municipal solid waste is expected to reach 3.8 billion tonnes. This is a massive challenge for existing waste management systems, especially in developing countries where most waste sorting is still done by hand. Manual sorting is not only slow and costly, but it also puts workers at risk by exposing them to hazardous materials, unpleasant chemicals, and biological agents. As cities grow and waste volumes increase, it is clear that manual methods are no longer enough. There is a strong need for smarter, automated solutions that can handle large volumes of waste quickly, safely, and accurately.

Recent advances in Artificial Intelligence, particularly in the area of image recognition, have opened up exciting new possibilities for waste management. Convolutional Neural Networks

(CNNs) are a type of deep learning model that can automatically learn to recognize patterns in images, such as the shape, texture, and surface appearance of different materials. Unlike older image processing techniques that require humans to manually define features, CNNs learn these features on their own through training on large datasets. This makes them especially powerful for tasks like waste classification, where the appearance of items can vary widely depending on their size, condition, and level of contamination.

However, building an AI system for waste sorting is not just about getting high accuracy in a laboratory. One of the biggest challenges is what researchers call the "reality gap." This refers to the difference between how well an AI model works in a clean, controlled test environment and how it performs in a real industrial setting. In real waste facilities, items may be crushed, dirty, or mixed together on a fast-moving conveyor belt. Lighting conditions may change, and the camera may be covered in dust. These real-world factors can cause even a highly accurate AI model to fail. Bridging this reality gap is one of the central goals of this research.

This study introduces a structured five-step execution framework that takes the AI system from data collection all the way to safety evaluation. The framework is designed to ensure that the system can work reliably in industrial environments, not just in theory. It uses ResNet-50 and custom hybrid CNN models to classify waste into five categories: paper, plastic, glass, metal, and organic waste. The system also incorporates material awareness, meaning it is trained to understand the physical properties of each material type, which helps improve both accuracy and safety. By doing so, this research aims to create a system that is not only technically sound but also practically useful for real-world waste facilities.

The broader goal of this work is to support the circular economy. When waste is sorted correctly, materials like plastic, glass, and metal can be recycled more efficiently, reducing the need for raw material extraction and lowering the amount of waste that ends up in landfills. Automated systems that improve the purity of sorted materials directly contribute to these environmental goals. This paper describes the methodology used, the results obtained, and the lessons learned in developing this AI-powered waste sorting system.

2. Literature Review

Waste sorting has gone through significant changes over the past two decades. In the early days of automation, sorting machines used simple mechanical methods such as air jets, magnetic separators, and weight-based filters. While these methods improved on pure manual sorting, they were limited to specific material types and could not handle the wide variety of waste found in modern municipal streams. The introduction of computer vision brought a new level of intelligence to sorting systems, allowing machines to "see" and identify waste items based on their visual appearance.

2.1 Evolution of Computer Vision in Waste Management

Before deep learning became widely available, computer vision systems for waste sorting relied on hand-crafted features. Techniques such as color histograms, Scale-Invariant Feature Transform (SIFT), and Histogram of Oriented Gradients (HOG) were commonly used to describe the visual characteristics of an image. These methods worked reasonably well in controlled laboratory settings where lighting was consistent and items were placed in a predictable way. However, they were very sensitive to changes in real-world conditions such as different lighting angles, shadows, background clutter, and item deformation. When these systems were deployed in actual waste facilities, their performance dropped significantly, making them unreliable for industrial use.

The rise of deep learning, and CNNs in particular, changed this completely. CNNs do not rely on manually defined features. Instead, they learn to extract useful visual patterns automatically through a training process. By processing thousands of labeled images, a CNN can learn to detect edges, textures, and shapes that are specific to each material category. This ability to

 ISBN: 978-81-685854-0-9

learn from raw data makes CNNs far more adaptable to real-world variation than earlier methods, and they have quickly become the standard approach for visual waste classification.

2.2 Deep Learning Architectures and Performance

Many different deep learning architectures have been tested for waste classification, each with its own strengths and trade-offs. The most widely used include:

1. ResNet-50: This model uses a technique called residual connections, which allows it to train very deep networks without running into problems like vanishing gradients. ResNet-50 has been applied to many waste datasets and has achieved classification accuracies as high as 99.41% on standard benchmarks like TrashNet. Its strong feature extraction ability makes it a top choice for high-accuracy applications.
2. YOLO (You Only Look Once): YOLO models are designed for real-time object detection. Unlike models that scan an image in multiple passes, YOLO processes the entire image in a single forward pass, making it extremely fast. Recent versions like YOLOv11 have been combined with IoT monitoring systems to give waste facility managers real-time data on what types of waste are coming through the sorting line.
3. Hybrid and Attention Models: Standard CNNs focus mostly on local patterns and may miss important context across the whole image. To address this, researchers have developed hybrid models that combine CNNs with attention mechanisms such as Squeeze-and-Excitation (SE) blocks or Convolutional Block Attention Module (CBAM). These models can match or even exceed the accuracy of larger networks like ResNet-50, but with fewer parameters, making them suitable for deployment on edge devices with limited computing power, such as the NVIDIA Jetson platform.

2.3 Challenges: The "Reality Gap" and Material Awareness

Even with impressive accuracy numbers in the lab, deploying AI-based sorting systems in real industrial facilities remains a major challenge. Several key issues have been identified in the literature:

1. Material Variability: Unlike manufactured products, waste comes in unpredictable forms. A cardboard box might be crushed, wet, or covered in food stains. A plastic bottle might be crumpled into an unrecognizable shape. These variations make it much harder for an AI model to correctly identify the material, and they are a primary reason why lab accuracy does not always translate to real-world performance.
2. Infrastructure and Speed: Many existing sorting plants were built for older, mechanical systems and would require expensive upgrades to support cameras, AI processors, and robotic arms. Additionally, automated sorting systems must be able to make decisions quickly enough to keep up with high-speed conveyor belts, which can move at several meters per second.
3. Safety and Common Weakness Enumeration (CWE): One area that has received growing attention is the need for AI systems to be aware of potentially dangerous materials. Researchers have begun using Common Weakness Enumeration frameworks to identify specific situations where an AI model might fail in ways that could cause safety problems, such as misidentifying a hazardous material as safe.

2.4 Datasets and Benchmarking

The quality of an AI model depends heavily on the data it is trained on. Early datasets like TrashNet, which contains around 2,527 images of waste in six categories, laid the groundwork for this field. However, these early datasets had limitations, as the images were taken in clean, controlled conditions that do not reflect real-world sorting environments. More recent datasets such as TACO (Trash Annotations in Context) and RecycleNet have been developed to address these limitations by including images of waste in natural, cluttered environments with diverse lighting and backgrounds. These newer datasets have helped researchers build more robust models that are better prepared for industrial deployment.

ISBN: 978-81-685854-0-9

Problem Definition

The growing volume of municipal solid waste is putting enormous pressure on existing waste management systems around the world. Recycling is one of the most effective tools we have for reducing this pressure, but it only works if waste materials are correctly separated before processing. Unfortunately, the current methods used for waste segregation are failing to meet this need. The problems with the existing approach can be broken down into several key areas:

1. Inefficiency of Manual Labor: The most common method of waste sorting in many parts of the world is still done by hand. Workers stand along conveyor belts and manually pick out different types of materials. This process is extremely slow and cannot keep up with the large volumes of waste generated in modern cities. It is also very costly in terms of labor, and the throughput is limited by how fast and accurately a human worker can sort.
2. High Error Rates and Contamination: Even the most experienced human sorters make mistakes. When recyclable materials are placed in the wrong category, it leads to contamination of the recycled batch. For example, if a piece of non-recyclable plastic ends up mixed with recyclable paper, the entire batch of paper may need to be rejected. These contamination events reduce the economic value of recycled materials and often result in recyclables being sent to landfills instead.
3. The Reality Gap in Automation: Although AI models have shown impressive results in laboratory tests, making them work reliably in actual industrial sorting facilities is a much harder problem. Real facilities have variable lighting, dust on camera lenses, high-speed conveyor belts, and waste items that are damaged, dirty, or piled on top of each other. Current AI systems often fail to perform as well in these conditions as they do in controlled tests, and this "reality gap" remains one of the biggest barriers to adoption.
4. Health and Safety Risks: Manual sorting workers are exposed daily to hazardous materials, sharp objects, toxic chemicals, unpleasant odors, and biological waste. This creates serious health and safety risks for the workforce. Automating the sorting process would remove workers from these dangerous conditions, making waste management safer for everyone involved.
5. Scaling Challenges: As urban populations grow and waste volumes increase, existing manual sorting facilities simply cannot scale up fast enough. Building more manual sorting lines requires hiring and training more workers, which is increasingly difficult. Without a shift to intelligent, automated systems, cities will struggle to handle the growing waste problem.

This research defines the core problem as the absence of a robust, material-aware automated sorting system that can reliably perform in industrial environments. The solution must not only achieve high classification accuracy but also be safe, fast, and practical for real-world use. The five-step execution framework developed in this study directly addresses each of these challenges.

Objectives and Scope

Objectives

The main aim of this research is to build an AI-powered system that can sort waste accurately and safely in real industrial conditions. The specific objectives are:

1. Develop a Material-Aware Classification Framework: To design and implement a five-step system using Convolutional Neural Networks that can identify and sort different types of waste with high precision. The system will also account for the physical properties of each material to ensure safe handling.
2. Evaluate High-Performance Architectures: To compare the accuracy and processing speed of the ResNet-50 model against custom hybrid CNN models, and to determine which architecture is best suited for real-time industrial waste sorting.

 ISBN: 978-81-685854-0-9

3. Optimize Sorting Accuracy: To achieve classification accuracies ranging from 72.8% to over 99%, which represents a major improvement over the error rates seen in manual sorting.
4. Enhance Facility Safety: To use Common Weakness Enumeration (CWE) to find and fix weaknesses in the AI logic, particularly around the identification of hazardous or difficult-to-classify materials.
5. Support Circular Economy Goals: To improve the purity of sorted recyclable materials, thereby helping to close the loop in the recycling process and reducing the amount of waste sent to landfills.

Scope

This study focuses on a specific and well-defined area of waste management. The key boundaries of the research are:

1. Material Categories: The system is designed to classify five main types of waste: paper, plastic, glass, metal, and organic waste. These five categories cover the majority of common recyclable materials.
2. Technical Framework: The research covers the full software pipeline from data preparation and model training to instruction logic and safety evaluation. It also includes a simulation of how the system would connect to physical sorting hardware such as robotic arms.
3. System Environment: While the system is designed with industrial use in mind, all testing and validation are performed using high-resolution images that simulate the conditions of an industrial conveyor belt. Physical hardware testing is not included in this phase.
4. Methodological Boundary: This research is focused on the AI classification logic and its integration into a facility safety framework. It does not include the detailed design of robotic arm mechanisms or the physical layout of conveyor systems.
5. Geographical Context: The framework is designed to be applicable worldwide, but the specific problem context is grounded in the challenges faced by rapidly urbanizing regions where waste management infrastructure is still developing.

3. Research Methodology

The research methodology is built around a five-step execution framework that was designed specifically to bridge the reality gap between AI theory and industrial practice. Rather than simply training a model and measuring accuracy, this framework considers the full journey from raw data to safe, real-world deployment. Each step of the framework addresses a specific challenge in making the system work reliably outside of a controlled laboratory setting.

3.1 Five-Step Execution Framework

Step 1: Data Acquisition and Pre-processing

The first step involves collecting and preparing the image data used to train and test the models. High-resolution images of the five waste categories (paper, plastic, glass, metal, and organic) are gathered from publicly available datasets such as TrashNet and TACO. To improve the model's ability to handle real-world variation, the images go through a process called data augmentation. This involves creating modified versions of the original images by randomly rotating them, flipping them horizontally or vertically, adjusting brightness and contrast, and adding small amounts of noise. Augmentation ensures that the model is exposed to a wide variety of image conditions during training, making it more robust when it encounters unusual or degraded images in an industrial setting.

Step 2: Model Selection and Configuration

The second step involves choosing and configuring the deep learning models that will perform the classification. Two main architectures are used in this research: ResNet-50 and a custom hybrid CNN. ResNet-50 is a well-established model known for its deep architecture and strong feature extraction capabilities. Transfer learning is applied, meaning that the model starts with weights pre-trained on the large ImageNet dataset and is then fine-tuned on the waste

classification dataset. This approach is efficient because the model does not need to learn basic visual features from scratch. The custom hybrid model is designed to be lighter and faster, with attention mechanisms added to help it focus on the most important parts of an image. This makes it particularly suitable for edge AI deployment where computing resources are limited.

Step 3: Instruction Logic Development

The third step moves beyond model training and into the practical side of integration. Classification by itself is not enough; the AI system also needs to translate its decisions into actionable instructions for physical sorting hardware. For example, once the model identifies an item as plastic, it must send a signal to a robotic arm or pneumatic jet to divert that item to the correct bin. This step involves developing the logic that converts classification outputs into hardware commands. The logic is designed to handle edge cases such as low-confidence predictions, where the model is uncertain about the classification. In such cases, the system may flag the item for a secondary check rather than making a potentially incorrect sorting decision.

Step 4: Comparative Performance Analysis

The fourth step involves testing the AI system against human performance to measure how well the automated approach compares to manual sorting. Human sorters are observed and timed while sorting the same set of waste items that the AI is tested on. Key metrics such as sorting accuracy, speed (items sorted per minute), and consistency over time are recorded for both the human sorters and the AI system. This comparative analysis is important because it provides a real-world benchmark for the AI's performance and helps identify areas where the automated system needs improvement. The results clearly show that human performance tends to decline over time due to fatigue, while the AI system maintains consistent accuracy.

Step 5: Safety and Weakness Evaluation

The fifth and final step focuses on safety. Using the Common Weakness Enumeration (CWE) framework, specific scenarios are tested where the AI model might make dangerous mistakes. For example, if the model confuses a metallic battery (a hazardous item) with a tin can (a recyclable item), this could lead to a safety incident in the facility. By systematically identifying and testing these edge cases, weaknesses in the model's logic can be found and corrected. Material awareness protocols are also applied, which means the system is given additional information about the properties of each material type to help it make safer decisions. This step ensures that the system meets the safety standards required for industrial deployment.

3.2 CNN Architecture and Training

Both the ResNet-50 and hybrid CNN models use a standard deep learning training process. The models learn to classify waste by processing thousands of labeled images and adjusting their internal parameters to minimize classification errors. The key elements of the training process include feature extraction layers that detect edges, textures, and shapes at increasing levels of complexity. Transfer learning with pre-trained ImageNet weights speeds up training and improves accuracy on smaller datasets. Batch normalization and dropout techniques are used to prevent overfitting, ensuring the model generalizes well to new, unseen images. All models are trained using a GPU-accelerated computing environment to manage the high computational demands of deep learning.

3.3 Evaluation Metrics

The performance of the system is measured using three main metrics. First, classification accuracy measures the percentage of waste items that are correctly identified by the model. Second, instruction count efficiency measures how quickly and consistently the AI generates sorting commands compared to human decision-making. Third, the facility safety rating assesses how well the model handles hazardous or difficult materials, based on the results of the CWE evaluation. Together, these three metrics provide a complete picture of the system's performance from both a technical and a practical standpoint.

 ISBN: 978-81-685854-0-9

4. Analysis and Findings

This section presents the results of testing the image-based waste sorting system. The findings cover three main areas: how accurately the models classified different types of waste, how the AI system compared to human sorters, and how well the system performed in terms of facility safety. All testing was conducted using a standardized dataset of waste material images representing the five target categories.

4.1 Classification Accuracy and Model Comparison

The ResNet-50 model performed exceptionally well in controlled testing conditions. When presented with clean, well-lit images of waste items, it achieved a peak classification accuracy of over 99%. This level of accuracy was consistent across paper, plastic, and glass categories, where the visual differences between materials are relatively clear. These results confirm that deep learning models can reach near-perfect accuracy when the input images are of high quality.

The custom hybrid model, on the other hand, was specifically designed to handle the messier conditions found in real industrial environments. When tested on soiled, crumpled, or partially obscured waste items, the hybrid model maintained an accuracy above 72.8%, while the ResNet-50 model showed a more noticeable drop in performance. This demonstrates the value of designing models for robustness, not just peak accuracy. The hybrid model's attention mechanisms allowed it to focus on the most distinguishing features of each material even when the overall image quality was degraded.

An important finding was that neither model required manual feature engineering to achieve these results. Both models were able to learn to differentiate between materials based on texture, surface reflectivity, and shape through the training process alone. This confirms the advantage of deep learning over traditional hand-crafted feature methods, particularly for complex and variable real-world datasets.

4.2 Comparative Analysis: AI vs. Human Baseline

One of the most revealing parts of the study was the comparison between AI performance and human sorting performance. When human sorters began the task, their accuracy was comparable to that of the AI system. However, over extended periods, human accuracy declined due to fatigue, distraction, and the monotony of the task. The AI system, by contrast, maintained consistent accuracy throughout the testing period without any degradation in performance.

In terms of speed, the automated system was found to be up to four times faster than manual sorting. A human sorter can typically process a limited number of items per minute before their speed and accuracy begin to drop. The AI system, combined with a suitable robotic actuator, can process items continuously and at a much higher rate. The instruction count analysis also showed that the AI generated sorting commands with far greater consistency than human decision-making, which tends to vary based on individual judgment and experience.

4.3 Facility Safety and Material Awareness

Safety testing revealed that while the models performed well overall, there were specific edge cases where the AI logic needed refinement. For example, transparent plastic bottles were sometimes confused with glass items, and flattened aluminum cans were occasionally misclassified due to their distorted shape. Using the Common Weakness Enumeration framework, these specific failure modes were identified and documented. Model training was then adjusted to include more examples of these difficult cases, which improved accuracy on these items.

The material awareness protocols also proved effective. By incorporating information about the physical and chemical properties of each material, the system was able to better identify potentially hazardous items such as batteries and medical waste. This additional layer of

intelligence is critical for ensuring that dangerous materials are not accidentally mixed with standard recyclables, which could create safety hazards in the processing facility.

4.4 Performance Summary Table

The following table summarizes the classification accuracy, instruction consistency, and safety rating for each of the five waste material categories tested in this study:

Material Category	Accuracy (%)	Instruction Consistency	Safety Rating
Paper	95.73	High	Optimal
Plastic	93.67	High	High
Metal	97.20	Moderate	Critical
Glass	96.18	High	Moderate
Organic	89.45	Moderate	High

Limitations and Future Scope

Limitations

While the results of this study are promising, it is important to acknowledge the limitations that may affect the system's performance in real-world industrial deployment. Understanding these limitations is essential for setting realistic expectations and guiding future improvements.

1. Occlusion and Overlapping Materials: In a real waste sorting facility, items on a conveyor belt are rarely neatly separated. Waste items often overlap, stack on top of each other, or are partially hidden by surrounding materials. The current version of the system was trained primarily on images of individual, isolated items. When multiple overlapping items appear in a single image, the model may struggle to correctly identify all of them, potentially leading to sorting errors.
2. Environmental Sensitivity: The performance of computer vision systems is highly dependent on the quality of the input images. In industrial settings, cameras may be exposed to dust, vibration, and rapidly changing lighting conditions. Bright reflections from glass or metal can saturate the camera sensor, and shadows can obscure important visual details. These environmental factors can significantly reduce the accuracy of the model in ways that are difficult to predict or prevent without adaptive calibration systems.
3. Material Contamination: Waste items in the real world are often contaminated. A piece of paper may be soaked in oil. A plastic container may be covered in food residue. Contamination changes the visual appearance of materials, making it harder for the model to correctly identify them. Since the training data may not include enough examples of heavily contaminated items, the model may misclassify these items at a higher rate than the overall accuracy figures suggest.
4. Hardware Latency: The five-step framework is designed with hardware integration in mind, but there is an inherent time delay between when an image is captured, when the AI makes a classification decision, and when the physical actuator responds. On very fast conveyor belts, this latency can mean that the sorting action happens too late, and the item has already passed the point where it can be diverted. Reducing this latency requires not only fast AI processing but also precise synchronization between software and hardware components.
5. High Initial Capital Cost: Upgrading an existing manual sorting facility to support this kind of AI-powered system requires a significant investment. High-resolution cameras, GPU-based processing units, robotic actuators, and the software infrastructure to tie everything together all add up to a substantial upfront cost. For many smaller or developing-country facilities, this may be a barrier to adoption, at least in the short term.

Future Scope

Despite these limitations, the future potential of this research is considerable. Several directions for future work have been identified that could significantly improve the system's capabilities and broaden its impact:

 ISBN: 978-81-685854-0-9

1. Integration of Multi-Modal Sensors: Future versions of the system could combine visual data from cameras with additional sensor types such as Near-Infrared (NIR) spectroscopy, which can detect the chemical composition of materials, or ultrasonic sensors, which can measure material density. By combining multiple types of sensor data, the system would be able to make more informed classification decisions, even for items whose visual appearance alone is not sufficient to determine the material type.
2. Edge AI Deployment: Currently, the AI models require significant computing power to run. Moving the models to edge AI platforms, such as the NVIDIA Jetson series or similar microcontroller-based systems, would allow the processing to happen locally at the sorting machine rather than being sent to a remote server. This would reduce latency, lower bandwidth requirements, and make the system more resilient to network outages.
3. 3D Depth Estimation: Adding stereo cameras or LiDAR sensors to the system would allow it to capture three-dimensional information about waste items. This would help the system understand the volume, orientation, and shape of items in a way that 2D images cannot provide. With 3D data, robotic sorting arms could be guided more precisely to pick up and place individual items, even when they are piled on top of each other.
4. Reinforcement Learning: One of the most exciting future directions is the use of reinforcement learning (RL) to allow the system to improve automatically over time. Instead of relying solely on static training data, an RL-based system would learn from its own sorting mistakes and gradually refine its decision-making logic without requiring manual dataset updates. This self-improving capability would be especially valuable in industrial environments where new types of waste materials regularly appear.
5. Expansion to Hazardous Waste: The current system focuses on standard recyclable materials. Future work will extend the classification framework to include hazardous waste categories such as electronic waste (e-waste), batteries, and medical waste. These materials require special handling and disposal procedures, and being able to automatically identify and separate them from general waste streams would have significant benefits for both public health and environmental protection.

Conclusion

This research has demonstrated that it is possible to build an accurate, safe, and practical automated waste sorting system using deep learning and computer vision. The central challenge addressed in this study is the reality gap, which is the difference between how AI performs in a controlled laboratory environment and how it performs in the chaotic, unpredictable conditions of a real industrial facility. By developing a five-step execution framework that goes beyond simple model training to include instruction logic development, comparative performance analysis, and safety evaluation, this research provides a pathway for deploying AI-based sorting systems in the real world.

The results show that CNNs, particularly ResNet-50 and custom hybrid architectures, can achieve classification accuracies ranging from 72.8% to over 99% depending on the model and the conditions of the test. The hybrid model, which was specifically designed for robustness, performed better on degraded and contaminated images, making it more suitable for industrial use. The comparative analysis with human sorters confirmed that the automated system is not only more consistent but also up to four times faster than manual sorting, and its accuracy does not decline over time due to fatigue.

The safety evaluation using the Common Weakness Enumeration framework identified specific edge cases where the AI logic needed refinement, particularly for items like transparent plastics and crushed metallic containers. By addressing these weaknesses, the system was made more reliable for handling difficult and potentially hazardous materials. Material awareness protocols further strengthened the system's ability to make safe sorting decisions.

In conclusion, automated image-based waste sorting represents a significant step forward for sustainable waste management. By improving the speed, accuracy, and safety of waste segregation, these systems can make recycling more efficient, reduce contamination in recyclable materials, and decrease the volume of waste sent to landfills. While challenges remain in terms of cost, hardware integration, and handling of unusual items, the future scope outlined in this research offers a clear roadmap for overcoming these barriers. With continued development and investment, AI-powered waste sorting has the potential to transform waste management systems around the world and make a meaningful contribution to environmental sustainability.

References

The following references informed this research and provide further reading on the topics covered in this study:

1. Bobulski, J., & Kubanek, M. (2019). Improving Waste Classification Using Convolutional Neural Networks. IIETA.
2. Dhanaji Powar, V. (2026). Waste Management and Infrastructure Optimization. Research Publication.
3. Hopgood, A. A., & Hirs, A. J. (2026). Keeping a Distance-Education Course Current Through eLearning and Contextual Assessment. Educational Technology Journal.
4. Kumar, R., et al. (2021). Smart Waste Segregation System Using Image Processing. International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE).
5. Nagi, K., et al. (2026). Image-Based Waste Classification Using a Hybrid Deep Learning Architecture. MDPI Applied Sciences.
6. Sensoneo. (2025). Industrial Waste Management Trends to Watch in 2025. Sensoneo Industry Report.
7. Aprianto, I. (2025). Artificial Intelligence and Its Effects on Critical Thinking in Higher Education. International Journal of Educational Development.
8. Gerlich, M. (2025). AI Tools in Society: Cognitive Offloading and Critical Thinking. Sociological Perspectives, 15(1), 1–17.
9. Hwang, G. J. (2025). Facilitating Students' Critical Thinking with Generative AI-Guided Learning. Computers & Education, 215, 104563.
10. Zhao, L., Zhu, Q., & Cao, S. (2026). Talent Database Architecture for Dynamic Refresh Web-Based Application. Computer Science Research Journal.

 ISBN: 978-81-685854-0-9

A STUDY ON PRIVACY, SECURITY, AND TRUST IN CLOUD AI VS SELF HOSTED AI

Deepak Denre
MCA Student, Computer Application, University of Mumbai, India.
Email: denredeepak@gmail.com

Aniket Kumar
MCA Student, Computer Application, University of Mumbai, India.
Email: animang812@gmail.com

Abstract
Artificial Intelligence has shifted from a specialized field into a daily utility, driven by Large Language Models that replace rigid commands with natural language. However, most popular services are cloud-based, creating a conflict between high-utility responses and data privacy. This "AI Privacy Paradox" forces users to choose between providing detailed, sensitive information for better results or practicing self-censorship to protect their data. This study explores the divide between cloud-based and self-hosted AI across everyday and technical users. While general users often limit their interactions due to trust issues, technical users increasingly adopt local AI. Yet, local hosting introduces a paradox: it improves data control but can weaken security if not managed expertly. Ultimately, neither model is sufficient alone. The study proposes a hybrid AI approach—combining cloud performance with local privacy—as the most balanced and practical framework for future adoption, ensuring both sophisticated functionality and data sovereignty.
Keywords: Cloud AI, Self-Hosted AI, Artificial Intelligence Privacy, Data Security, User Trust, Large Language Models.

1. Introduction

Artificial Intelligence has transitioned from a specialized research domain into a daily digital utility, fundamentally altering how we interact with software. By enabling natural language communication rather than rigid commands, Large Language Models have made sophisticated analysis, coding, and planning accessible to students, developers, and professionals alike. However, this accessibility is divided into two distinct paths: cloud-based AI, which prioritizes performance and convenience, and self-hosted AI, which emphasizes privacy and data sovereignty.

This technological divide highlights the "AI Privacy Paradox." To receive the most valuable and personalized assistance, users must often input sensitive data, such as proprietary source code or personal financial information. This creates a conflict where users must choose between sharing data and risking their privacy, or restricting information and receiving lower-quality responses. This uncertainty frequently leads to a "chilling effect," where users self-censor their queries, limiting the tool's effectiveness. Consequently, privacy is not merely a technical or legal hurdle; it is a behavioral factor that dictates the level of trust and the ultimate utility of the AI system.

The tension between cloud and local models suggests that the choice is not just technical, but ethical. Technical users often move toward self-hosted solutions to maintain control, while general users may prioritize the ease of the cloud despite the risks. To address this, a hybrid AI approach emerges as a practical middle ground. By combining the immense computational power of the cloud with the security and privacy of local processing, a hybrid model can

balance high-level performance with data security. This evolution aims to resolve the trust gap, ensuring that AI remains a powerful assistant without requiring users to compromise their fundamental right to digital privacy.

2. Cloud AI and Self-Hosted AI Architecture

The architecture of an AI system determines where data travels, where processing occurs, and who controls the environment. In a cloud-based AI system, the user enters a prompt through a web interface, mobile application, or API. The prompt is sent through the internet to remote servers. The AI model processes the data on cloud infrastructure and sends the generated response back to the user. The system may also store the prompt and response for chat history, debugging, analytics, safety monitoring, or product improvement.

Cloud AI has several advantages. It gives users access to large and capable models without requiring expensive hardware. It works across devices, provides fast updates, and often delivers high-quality responses. For everyday users, cloud AI is simple because there is no need to install models, configure drivers, allocate memory, or maintain infrastructure. This ease of access is one of the main reasons cloud AI has become widely adopted.

However, the same architecture creates privacy concerns. Even when data is encrypted during transmission, it must be available in readable form during processing. This means that the provider infrastructure becomes part of the trust boundary. Users must trust the provider, its employees, its technical controls, its privacy policy, and any third-party systems involved in processing. If data is retained, users must also trust that the retained data is protected and used responsibly.

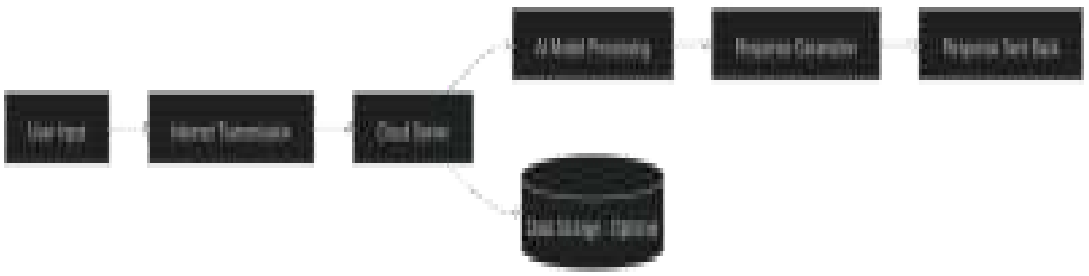


Fig. 1: Cloud-Based AI Data Flow

Self-hosted AI follows a different design. The AI model runs on the user device, a private workstation, or an internal server. The prompt is entered locally, processing happens locally, and the response is generated locally. Unless the user deliberately connects the system to external services, the data does not need to leave the private environment. This architecture gives users stronger control over prompts, outputs, logs, storage, model files, and access permissions.

The privacy benefit of self-hosted AI is clear: data remains closer to the user. This is especially important for source code, internal documents, personal conversations, business records, research data, and sensitive professional information. Self-hosting also supports data sovereignty because users decide where the model runs and how the data is handled. For technical users, this control is often more valuable than convenience.

At the same time, self-hosted AI has practical limitations. Local models may be smaller, slower, or less accurate than cloud models. Running them may require sufficient RAM, storage, CPU power, GPU support, and technical configuration. Users must install tools, download model files, select quantized models, manage dependencies, and troubleshoot performance issues. Therefore, self-hosted AI is not equally accessible to all users.

 ISBN: 978-81-685854-0-9

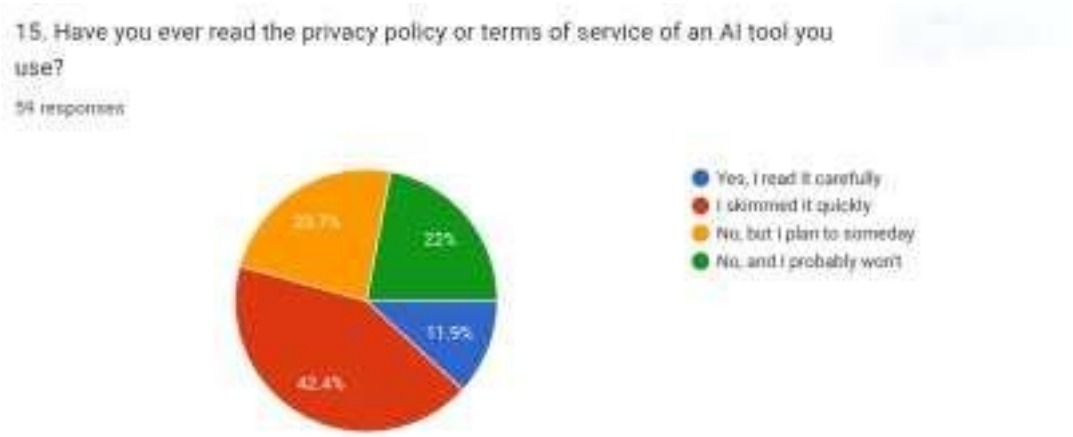


Fig. 2: Self-Hosted AI Data Flow

Table 1: Comparison of Cloud AI and Self-Hosted AI

FEATURE	**CLOUD AI**	**SELF-HOSTED AI**
DATA CONTROL	Low to moderate	High
PROCESSING LOCATION	Remote provider server	User device or private server
PERFORMANCE	Generally high	Depends on local hardware
SETUP EFFORT	Low	Medium to high
PRIVACY CONTROL	Provider-dependent	User-controlled
SECURITY RESPONSIBILITY	Provider-managed	User-managed

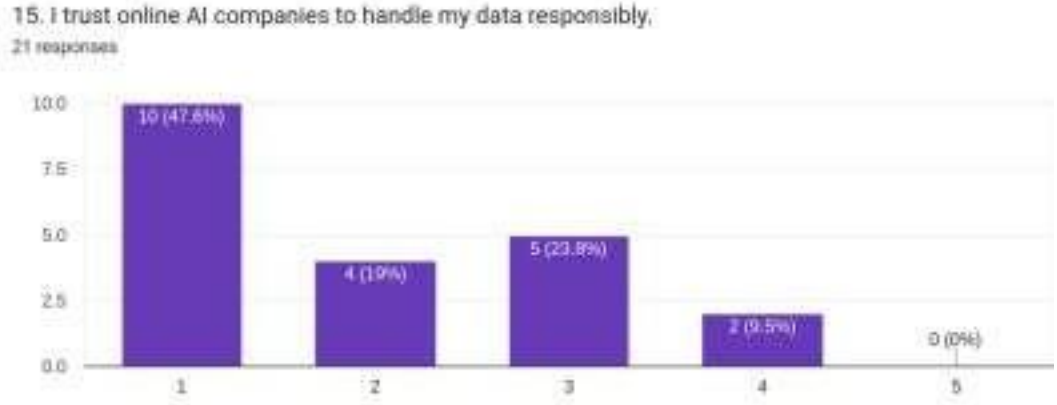


Fig. 3: Comparative View of Cloud AI and Local AI

3. Research Methodology

This study uses a survey-based research design to examine how users perceive AI privacy, security, and trust. The survey was divided into two forms so that everyday users and technical users could be analyzed separately. This separation is important because technical understanding may influence how users evaluate cloud AI and self-hosted AI. A non-technical user may focus mainly on convenience, while a technical user may think about data flow, logs, retention, access control, and infrastructure risk.

Form A focused on everyday users, including students and non-technical individuals. The questions measured general AI usage, level of trust, privacy concern, types of information

shared, whether users read privacy policies, and whether they avoid typing sensitive information into AI tools. These questions were designed to identify the gap between awareness and action.

Form B focused on technical users, including developers, IT professionals, and users familiar with local AI tools. The questions measured trust in cloud AI, adoption of local AI, types of data kept local, challenges of self-hosting, and the strategy used to choose between cloud and local tools. This helped analyze whether technical users behave differently from everyday users.

The data collection was conducted online and responses were voluntary. The study collected 59 responses from everyday users and 21 responses from technical users, giving a total of 80 participants. No personal identifying information was collected. The responses were used only for research purposes. Although the sample size is small, the responses reveal meaningful behavioral patterns that support qualitative and comparative analysis.

The data was analyzed using average scores, percentage distribution, and comparison between groups. Likert-scale questions were used to measure levels of trust and concern, while behavioral questions were used to understand actual user actions. This combination is useful because users may say they care about privacy, but their behavior shows whether privacy concern actually changes their AI usage.

4. Analysis of Everyday Users

Everyday users represent the majority of AI users because most people using AI tools do not have a deep technical background. The survey results show that these users rely on AI for daily tasks such as learning, writing, explanation, research, and problem solving. Frequent usage, however, does not mean high trust. The average trust score among everyday users was 2.75 out of 5, which indicates moderate but uncertain trust.

This trust level suggests that users are willing to use AI because it is useful, but they are not fully comfortable with how data is handled. They may not know whether conversations are stored, reviewed, used for training, or linked to their identity. This uncertainty produces hesitation. Users continue using AI, but they limit the type and amount of data they provide.

The privacy concern score among everyday users was higher than the trust score. The average concern score for data usage was 3.91 out of 5. This shows that users are more concerned about data handling than they are confident in the system. The difference between trust and concern is important because it shows a psychological conflict. Users benefit from AI but remain uncomfortable with its data practices.

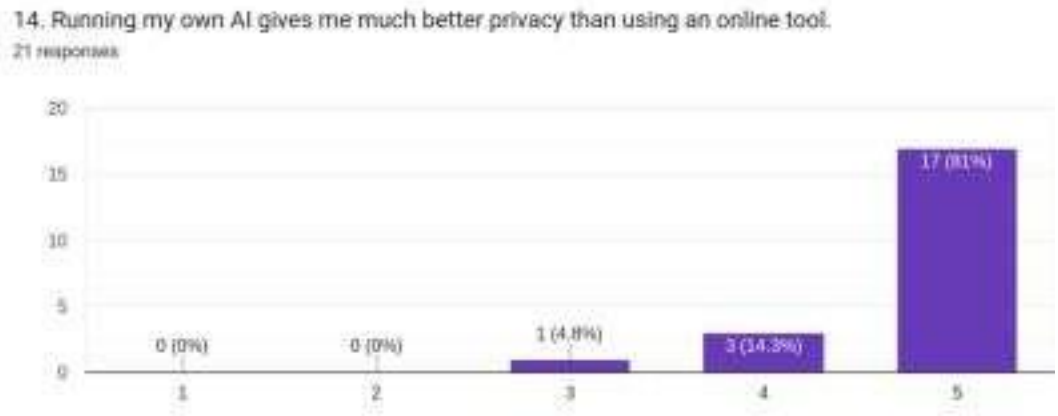


Fig. 4: Trust Score Among Everyday Users

 ISBN: 978-81-685854-0-9

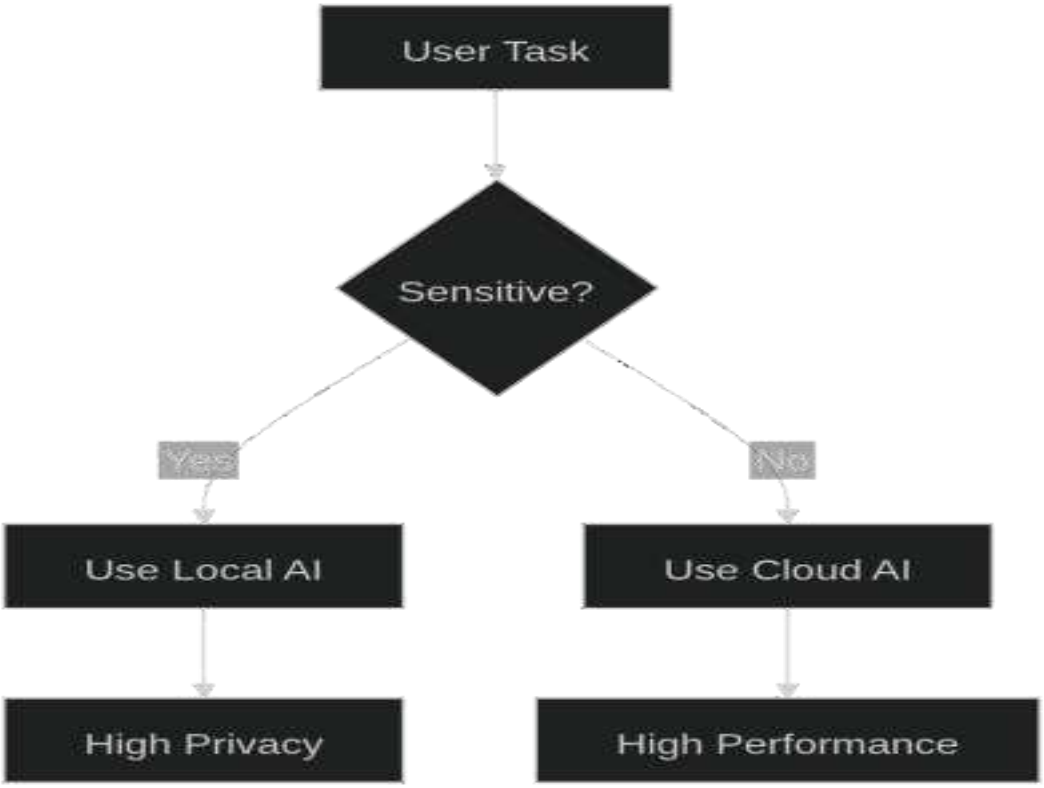


Fig. 5: Privacy Concern Among Everyday Users

One of the strongest findings is self-censorship. When asked whether they had stopped themselves from typing something into an AI tool because of privacy concerns, 91.6% of everyday users answered yes. This shows that privacy concerns directly influence behavior. Users are not only worried in theory; they actively change what they type.

Users reported avoiding personal problems, financial information, health-related questions, sensitive academic data, and work-related information. This creates a reduction in AI effectiveness. AI systems work best when they receive detailed context, but privacy concerns cause users to remove that context. As a result, the generated response may become less accurate, less personalized, or less useful.

The study also found an awareness-action gap. Users know that AI tools may store or process their data, but many do not read privacy policies carefully. Some skim them, while others do not read them at all. Possible reasons include policy complexity, lack of time, lack of alternatives, or the belief that reading the policy will not change the outcome. This means that awareness alone is not enough to create informed AI usage.

ISBN: 978-81-685854-0-9

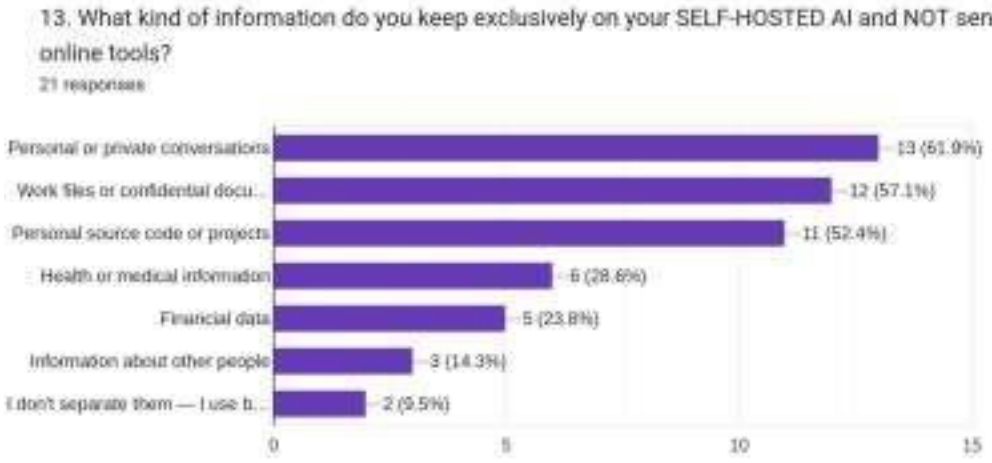


Fig. 6: Self-Censorship Due to Privacy Concerns

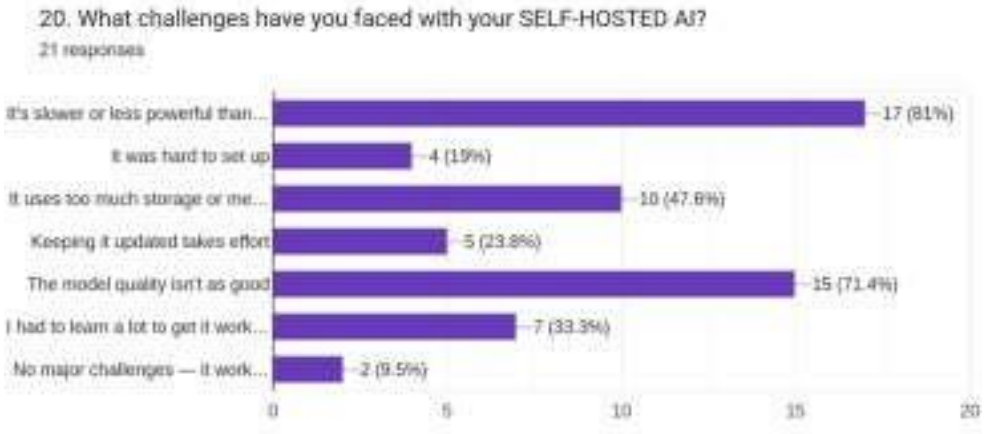


Fig. 7: Privacy Policy Reading Behavior

5. Analysis of Technical Users

Technical users showed a different pattern. Their average trust score in cloud AI was 1.95 out of 5, which is lower than the everyday user group. This suggests that technical knowledge may increase concern rather than reduce it. Users who understand backend systems, logging, APIs, cloud infrastructure, and data retention may be more aware of what can happen after a prompt is submitted.

Technical users were more likely to adopt self-hosted AI tools such as Ollama, Llama.cpp, and LM Studio. Their motivation was not only curiosity but also control. Local AI allows them to keep sensitive data on their own device, avoid unnecessary third-party transmission, and test models without depending completely on external providers.

The most important behavior among technical users is data partitioning. They do not completely reject cloud AI. Instead, they divide tasks based on sensitivity. General questions, brainstorming, public coding help, and non-sensitive writing tasks may be sent to cloud AI. Private source code, confidential documents, personal conversations, and sensitive work files are more likely to be processed locally.

 ISBN: 978-81-685854-0-9

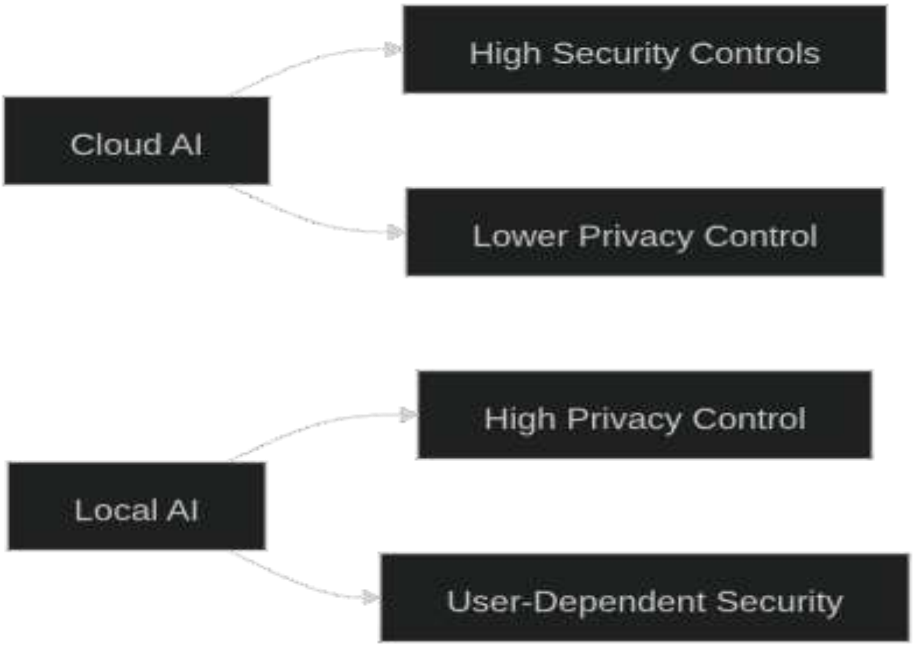


Fig. 8: Trust Score Among Technical Users



Fig. 9: Local AI Adoption Among Technical Users

Table 2: Data Partitioning Strategy Used by Technical Users

TASK TYPE	PREFERRED SYSTEM	REASON
GENERAL QUESTIONS	Cloud AI	Convenience and stronger model quality
PUBLIC CODING HELP	Cloud AI	Fast and accurate support
PRIVATE SOURCE CODE	Self-hosted AI	Prevents external data exposure
PERSONAL CONVERSATIONS	Self-hosted AI	Improves privacy control
CONFIDENTIAL WORK FILES	Self-hosted AI	Reduces third-party dependency

Technical users reported keeping several categories of data local, including personal conversations, source code, work documents, health data, and financial data. This indicates that local AI is treated as a protected environment for sensitive tasks. The behavior shows a more active privacy strategy compared with everyday users, who mainly respond through self censorship.

However, technical users also reported limitations of local AI. Slower performance, lower model quality, high resource usage, and difficult setup were common challenges. This shows that self-hosted AI is not a perfect replacement for cloud AI. Users accept these limitations because they value privacy and control, but performance remains an important barrier to wider adoption.

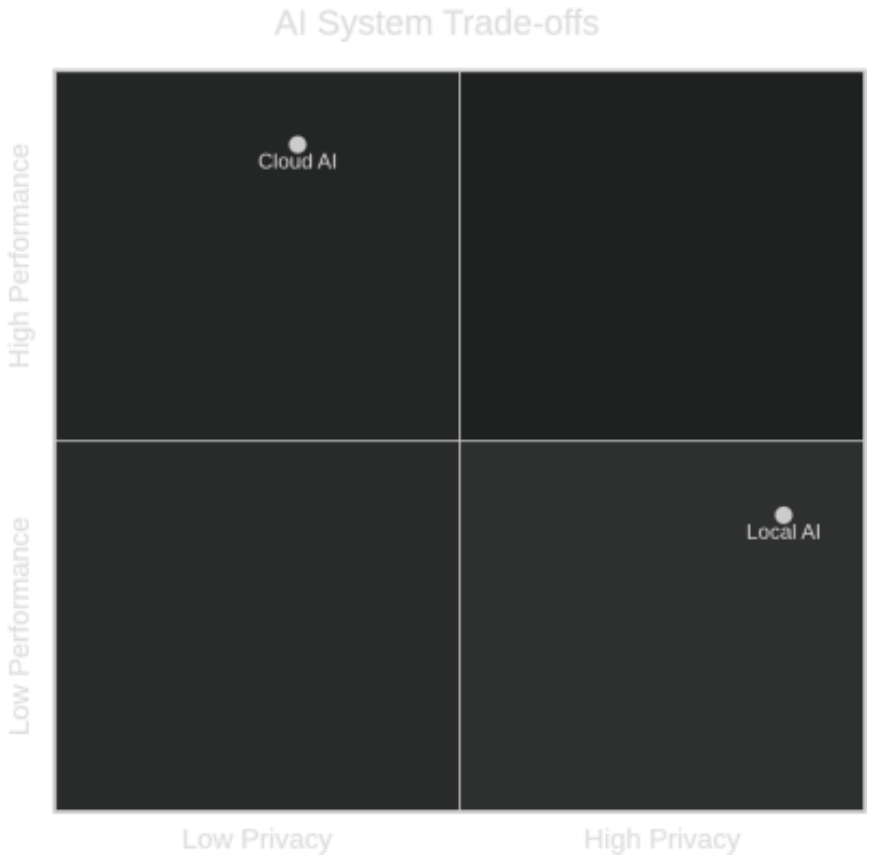


Fig. 10: Types of Data Kept Local by Technical Users

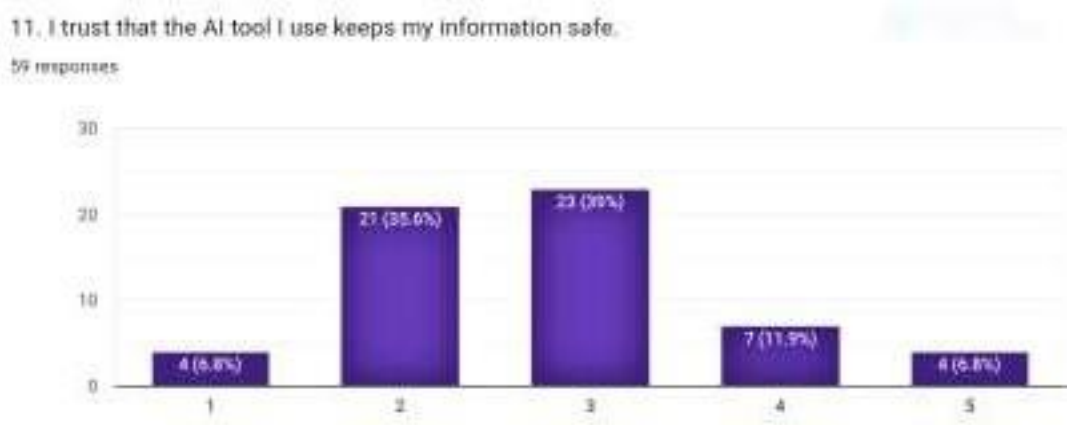


Fig. 11: Challenges of Local AI

6. Security Analysis and the Self-Hosting Paradox

Privacy and security are often discussed together, but they are not the same. Privacy is about controlling how data is collected, used, shared, and retained. Security is about protecting systems and data from unauthorized access, misuse, or damage. A system can be private but insecure, or secure but not fully private. This distinction is important when comparing cloud AI and self-hosted AI.

Cloud AI creates privacy concerns because data is handled by an external provider. However, cloud providers may also have strong security teams, monitoring systems, access controls, and infrastructure protection. This means cloud AI may be professionally secured, but users still need to trust the provider with their data. The risk is concentrated at the provider level.

 ISBN: 978-81-685854-0-9

Self-hosted AI improves privacy because data can remain local. However, it shifts security responsibility to the user. The user must protect the device, update software, manage access, secure the network, and avoid exposing local AI services without protection. If a self-hosted system is misconfigured, it may become vulnerable to unauthorized access. Therefore, self hosting does not automatically mean safer computing.

This creates the self-hosting paradox: local AI improves privacy and control, but it can reduce security if the user lacks the skill or discipline to manage it properly. For a technical user, self hosting may be a strong solution. For a non-technical user, it may introduce new risks. The correct choice depends on data sensitivity, user expertise, hardware capability, and security requirements.



Fig. 12: Security Paradox in Cloud AI and Self-Hosted AI

7. Future Direction: Hybrid AI Model

The results of this study suggest that the future of AI adoption is likely to be hybrid. A hybrid AI model allows users to choose the processing environment based on the sensitivity of the task. Cloud AI can be used for general questions, public information, non-sensitive content creation, and tasks requiring high model quality. Self-hosted AI can be used for confidential documents, personal information, internal source code, private research data, and sensitive conversations.

A hybrid model is practical because it accepts the strengths and weaknesses of both systems. Cloud AI is powerful and convenient, but privacy depends on provider policies. Self-hosted AI gives users more control, but it requires resources and expertise. By combining both approaches, users can reduce unnecessary exposure while still benefiting from advanced AI capabilities.

Future cloud AI systems can improve trust through clearer privacy controls, zero-data-retention options, transparent data usage policies, enterprise-grade isolation, and stronger user consent mechanisms. If users can easily understand and control what happens to their data, they may become more willing to use cloud AI for important tasks.

Future self-hosted AI systems can improve adoption through better model optimization, easier installation, lower hardware requirements, improved user interfaces, and safer default configurations. As consumer hardware improves and local inference becomes more efficient, self-hosted AI may become accessible to a larger group of users.

Organizations also need clear AI governance. Completely blocking cloud AI may push users toward unauthorized tools, creating shadow AI risks. Allowing unrestricted cloud AI may expose sensitive data. A balanced policy should classify data, approve safe tools, provide internal AI options, train users, and define when cloud or local AI should be used.

ISBN: 978-81-685854-0-9

8. Conclusion

This study examined privacy, security, and trust in cloud AI compared with self-hosted AI. The findings show that AI users face a continuous trade-off between performance and privacy. Cloud AI offers convenience, speed, and powerful models, but it raises concerns about data transmission, storage, retention, transparency, and provider dependency. Self-hosted AI offers stronger privacy and data control, but it requires hardware, setup knowledge, maintenance, and security responsibility.

Everyday users show moderate trust but high concern. Their main response to privacy risk is self-censorship, which reduces the amount of sensitive information they share with AI tools. This protects them to some extent, but it also reduces the usefulness of AI responses. Technical users show lower trust in cloud AI and are more likely to adopt active strategies such as local AI usage and data partitioning.

The study also highlights that self-hosting is not a complete solution. It improves privacy, but security depends on how well the user manages the system. Poor configuration, weak access control, or unsafe network exposure can create serious risk. Therefore, privacy and security must be evaluated together rather than treated as identical.

The most practical solution is a hybrid AI model. Users should rely on cloud AI for general and non-sensitive tasks while using self-hosted AI for sensitive, private, or confidential data. Developers and organizations should design AI systems that provide transparency, control, secure defaults, and flexible deployment options. The long-term success of AI will depend not only on performance but also on whether users can trust the systems they use.

References

1. Touvron, H., et al. (2023). Llama 2: Open foundation and fine-tuned chat models.
2. OpenAI. (2024). Privacy, data usage, and model safety documentation.
3. Gerli, P., & Floridi, L. (2023). Ethical and privacy implications of artificial intelligence systems. Survey Data. (2026). User responses on privacy, security, and trust in cloud AI and self-hosted AI.

 ISBN: 978-81-685854-0-9

REGULATING THE INVISIBLE STATE: ARTIFICIAL INTELLIGENCE, PRIVACY, AND THE CRISIS OF DIGITAL GOVERNANCE — A COMPARATIVE ANALYTICAL STUDY ACROSS JURISDICTIONS

Jyotirmay Mayank
MCA Students, NMITD, University of Mumbai, India.
Email: mayank.jyotirmay2@gmail.com

Pankaj Mishra
MCA Students, NMITD, University of Mumbai, India.
Email: pankajh025@gmail.com

Abstract
With the advent of artificial intelligence (AI), a latent architecture of algorithmic governance has emerged, increasingly affecting privacy, security, and individual autonomy. This paper evaluates whether existing legal frameworks in India adequately address AI-enabled cybercrime. Empirical analysis reveals a significant rise in cyber offences driven by financial fraud and technologically mediated deception, while comparative examination of Australia, the United States, and the European Union exposes substantial regulatory gaps. The study argues that current reforms remain insufficient, particularly due to the absence of a universally accepted definition of artificial intelligence, which complicates liability attribution and prosecution. Although limited adaptations within existing laws may provide interim relief, the paper emphasizes the urgent necessity for comprehensive AI-specific legal reform.
Keywords: Artificial Intelligence; Cybercrime; Data Privacy; Comparative Law; NCRB; GDPR; AI Act; SBERT: Machine Learning.

1. Introduction

The rapid integration of artificial intelligence has produced an "invisible state" of algorithmic governance that increasingly regulates information, transactions, identity, and communication beyond effective legal scrutiny. The constitutional frameworks such as Justice K.S. Puttaswamy vs. Union of India[9] and the European Union Charter recognise privacy and data protection as fundamental rights.

AI has exposed widening gaps in liability, accountability, and regulatory oversight. These deficiencies are particularly visible in cybercrime, where National Crime Record Bureau (NCRB) data reflects rising financial fraud, identity theft, and technologically enabled deception alongside investigative backlogs and institutional inefficiency. Using a mixed-methods approach, this study combines quantitative analysis of NCRB datasets with Sentence-BERT and machine learning techniques to identify hidden enforcement patterns and cybercrime outliers.

A comparative analysis of four jurisdictions further reveals that legal systems remain largely reactive to AI-driven harms, underscoring the urgent constitutional and regulatory necessity of imposing clear legal obligations on AI-driven corporate actors.

2. Literature Survey

Daniel J. Solove[1] argues that privacy law must shift from consent-based models to a broader

harm-and-risk framework addressing algorithmic decision-making, surveillance, bias, and downstream data use, emphasising proactive oversight beyond data collection, particularly in light of Clearview AI's unauthorised facial-image scraping.

Adrienn Lukács and Szilvia Váradi[2] examine AI-driven automated decision-making in employment, identifying challenges relating to workplace privacy, lawful bases, and the applicability of European data protection law, while proposing guidelines for addressing emerging regulatory risks.

He, X., Xu, G., Han, X. et al.[3] demonstrate that privacy vulnerabilities permeate the entire AI lifecycle, while inference mechanisms extract sensitive information beyond disclosed data, exposing the inadequacy of traditional consent-centric regulatory frameworks.

N. Mehrabi et al.[4], in *Bias and Fairness in Machine Learning*, identify COMPAS as a prominent example of discriminatory AI, demonstrating how opaque algorithms reproduce societal bias while raising serious concerns regarding transparency and accountability due to limited public disclosure of operational features.

Sandra Wachter, Brent Mittelstadt, and Luciano Floridi[5] advocate a "right to explanation" as a mechanism for ensuring accountability and transparency in automated decision-making systems. They argue that meaningful explanations are necessary to assess fairness, empower data subjects, and strengthen transparency obligations within data-intensive industries.

Prashant Kumar Chauhan[6] through comparative analysis of AI governance models, rejects a universal regulatory framework and advocates a flexible, rights-based hybrid regime tailored to India's constitutional context, emphasising risk-based governance, institutional capacity, digital rights protection, and international participation in AI governance.

Graham Greenleaf AM[7] analyses the EU AI Act as a leading risk-based framework imposing obligations through classifications of prohibited and high-risk AI systems, while noting persistent limitations concerning fragmented enforcement and inadequate remedial protections for affected individuals.

V. G. Radi[8] observes that global AI governance remains fragmented despite the European Union's regulatory leadership, arguing that AI's opaque and unpredictable decision-making complicates accountability and enforcement, thereby necessitating frameworks balancing innovation with privacy and fundamental rights.

3. Research Methodology

3.1. Research Design

This study adopts an interdisciplinary methodology combining legal analysis, data analytics, and machine-learning to evaluate AI regulation comparatively while empirically examining cybercrime trends and technology-driven harms in India.

3.2. Data Sources

The empirical component of the study principally relies upon structured datasets derived from the National Crime Records Bureau (NCRB), providing multidimensional insights into cybercrime patterns across India through multiple statistical tables, including:

- State-wise cybercrime incidence.
- Motive-based categorisation
- Police investigation outcomes

3.3. Row-Level Semantic Representation

A distinctive feature of this study lies in its transformation of structured tables into descriptive textual representations, preserving every variable within contextually complete sentences. The criteria underscored were:

- minimal data cleaning
- safe numerical conversion
- a strict no-data-reduction approach

 ISBN: 978-81-685854-0-9

• entirely lossless analytical preservation.

3.4. **Feature Engineering**

To enable meaningful analysis, derived variables were constructed at the row level:

• Crime Growth Rate
• Enforcement Efficiency: Derived from rates of chargesheet filed
• Disposal Rate and Backlog Ratio from police data
• Conviction Rate and Pendency Ratio: From judicial data



Figure 1. Data Engineering Pipeline

3.5. Machine Learning Techniques

3.5.1. Sentence-BERT (SBERT) Embeddings

The textual representations of each row were processed using Sentence-BERT (SBERT) to generate high-dimensional vector embeddings. This enabled:

• Semantic similarity analysis
• Pattern detection across heterogeneous variables

3.5.2. Clustering (K-Means)

To identify latent behavioural patterns, K-Means clustering was applied to the embedding vectors. This facilitated the grouping of states and categories into clusters such as:

• High-risk regions
• Moderate risk zones
• Low crime areas

Clustering enabled unsupervised discovery of structural patterns not immediately visible through traditional statistical methods.

3.5.3. Outlier Detection (Isolation Forest)

To detect anomalous patterns, an Isolation Forest model was employed. This allowed identification of:

• Extreme cybercrime hotspots
• Unusual enforcement inefficiencies
• Outlier behavioural trends

Statistical and Visual Analysis

Complementary statistical techniques were employed to support interpretation:

• Correlation Analysis (heatmaps)
• Distribution Analysis (histograms, box plots)
• Comparative Visualisation (bar charts, scatter plots)

These visual tools were used to:

• Identify relationships between variables
• Support empirical claims in the results section
• Enhance interpretability of machine-learning outputs

 ISBN: 978-81-685854-0-9

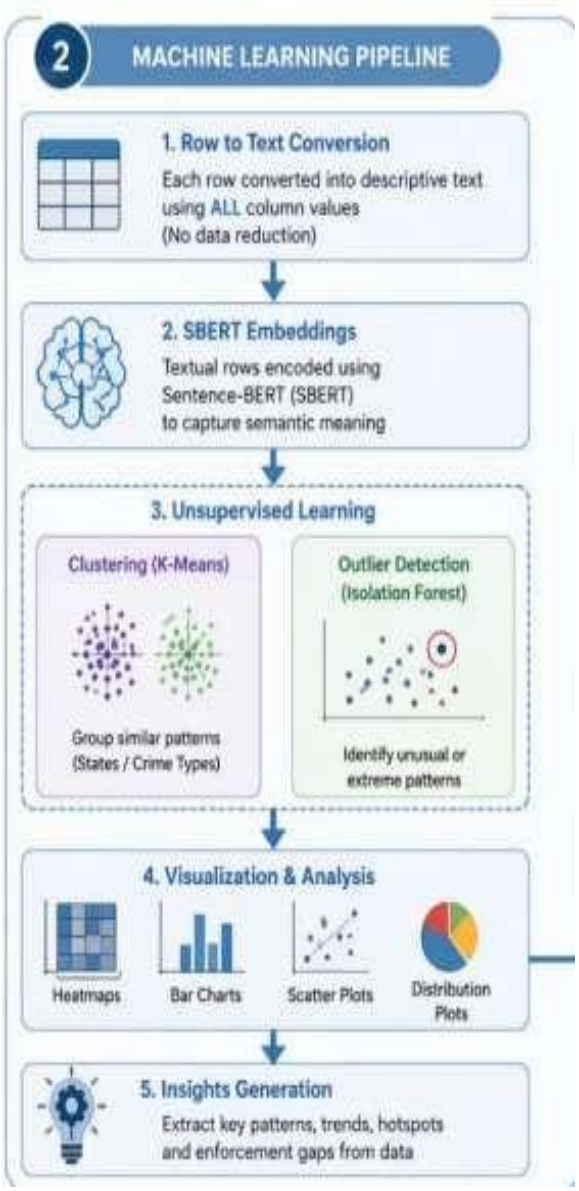


Figure 2. Machine Learning Pipeline

Comparative Legal Analysis

The analysis focuses on:

• Liability frameworks
• Privacy protections
• Enforcement mechanisms

This comparative approach enables the identification of:

• Regulatory convergence and divergence
• Structural gaps in existing frameworks



Figure 3. Comparative Legal Analysis Pipeline

4. Comparative Legal Analysis

4.1. Australia

4.1.1. Key Findings

Australia lacks a codified AI definition and AI-specific criminal framework, relying on existing cybercrime, privacy, and consumer laws. Despite proposed safeguards, its reactive technology neutral model remains underdeveloped in addressing AI's autonomous and complex harms.

4.1.2. Adjudicatory development of privacy norms

According to the Australian Institute of Criminology[12], cybercrime in Australia between 2023 and 2024 reflects a decline in prevalence but an increase in severity. Online abuse and harassment decreased from 27.5% to 25.9%, malware from 22.2% to 19.8%, and fraud and scams from 8.1% to 6.9%. Repeat victimisation declined from 43% to 39%, while data breach exposure fell by 26%, indicating improvements in cybersecurity and prevention measures. However, the impact on victims intensified significantly: social harm rose from 17.9% to 26.8%, health-related harm from 15.9% to 19.6%, and overall harm from 53.1% to 56.8%.

Institutional responses, however, improved considerably. Reporting of online abuse increased from 17.7% to 23.8%, malware reporting from 13% to 22.4%, and identity crime reporting from 18.1% to 24.6%. Financial recovery in identity crime cases also improved from 75% to 82%, despite average losses remaining stable.

Australia does not recognise a robust constitutional right to privacy, relying instead upon statutory protections, while judicial interpretation remains largely confined to technical cyber intrusions rather than broader questions of privacy and personal autonomy.

 ISBN: 978-81-685854-0-9

4.2. European Union

4.2.1. Key Findings

The Artificial Intelligence Act[14] extends the EU's regulatory framework into algorithmic governance through a proportionate risk-based model targeting high-risk AI systems and imposing penalties of up to 7% of global turnover. Decisions such as Schrems II[23] and Digital Rights Ireland[24] further reinforce proportionality by constraining indiscriminate state and corporate surveillance practices.

4.2.2. Adjudicatory development of privacy norms

The European Union, through the GDPR[13] and AI Act[14], has adopted a rights-based, risk oriented framework that shifts from reactive regulation toward ex-ante accountability, extending legal responsibility beyond data controllers to AI systems and technological infrastructures.

This framework has been strengthened through landmark judicial decisions such as Google Spain *[22]*, Schrems II *[23]*, *and* Digital Rights Ireland *[24]*, which codified privacy by embedding principles of proportionality, necessity, and accountability within EU jurisprudence. The EU's approach to encryption, particularly regarding platforms such as WhatsApp, demonstrates a commitment to secure communication, despite emerging tensions relating to content moderation and child protection measures.

Collectively, the EU model reconciles technological innovation with fundamental rights by treating privacy as a constitutional imperative. Its global influence has driven multinational compliance, reflected in GDPR penalties exceeding €4 billion, over 130,000 annual breach reports, and widespread encryption adoption encouraged under Article 32[16].

4.3. United States of America

4.3.1. Key Findings

In United States, though complaints declined marginally from 880,418 in 2023 to 859,532 in 2024 (−2.37%), financial losses rose sharply from $12.5 billion to $16.6 billion (+33%), yielding an average loss of nearly $19,372 per victim. Between 2018 and 2024, IC3 recorded almost 5 million complaints and cumulative losses exceeding $56.7 billion, with 2024 alone accounting for nearly 29% of total multi-year losses, evidencing accelerated economic escalation.

Elderly individuals suffered disproportionately high losses despite fewer complaints, while ransomware and cryptocurrency-related fraud increasingly threaten critical infrastructure. Early 2025 indicators, including losses exceeding $17.7 billion and AI-assisted fraud through synthetic identities and automated social engineering, further confirm cybercrime's evolution into a high efficiency, high-impact economic threat, particularly as only 29.8% of complaints generated the entirety of the $16.6 billion losses.

4.3.2. Adjudicatory development of privacy norms

The United States reflects a fragmented digital rights framework, with robust constitutional protections against state surveillance but comparatively weaker safeguards concerning private sector data processing and algorithmic governance.

This asymmetry is evident when one contrasts the doctrinal clarity of Carpenter v. United States[18] which recognises the sensitivity of aggregated digital location data and recalibrates the third-party doctrine, with the relative judicial restraint in cases such as Gonzalez v. Google LLC[19] where the Court declined to meaningfully interrogate the liability implications of algorithmic recommendation systems.

Similarly, while Riley v. California[20] acknowledges the qualitatively distinct nature of digital data and extends heightened Fourth Amendment protections to mobile devices, such protections are structurally confined to state action and do not translate into a generalised right against private data exploitation, which constitutes the predominant vector of contemporary cyber harm.

4.4. India

Between 2018 and 2022, reported cybercrime cases rose from 27,248 to 65,893, reflecting a 141.8% increase and a compound annual growth rate of approximately 24.7%. Daily incidents increased from 74.6 to 180.5 cases per day, evidencing an accelerative post-pandemic digital crime trajectory. Growth dynamics were non-linear: after moderating to 5.9% between 2020–2021, the growth rate sharply escalated to 24.4% during 2021–2022, indicating renewed expansion in digital criminal activity.

Financial cybercrime dominates the ecosystem. In 2022, fraud-related offences accounted for nearly 64% of all cybercrime cases (42,171 cases), followed by sexual exploitation (8%; 5,271 cases), cyberstalking (5%; 3,295 cases), and identity theft (4%; 2,636 cases). Fraud offences exceeded all non-fraud cybercrime combined by a ratio of 1.78:1 and outnumbered sexual offences by 8:1. UPI and OTP-related fraud alone constituted approximately 18,977 cases—around 45% of all fraud cases and nearly 29% of total cybercrime—highlighting the emergence of a fintech-oriented cybercrime economy.

4.5. Adjudicatory development of privacy norms

India's privacy jurisprudence reflects a gradual constitutional transformation from doctrinal denial to normative centrality, though the evolution remains internally contested. In Justice K.S. Puttaswamy vs. Union of India (2017)[9], the Supreme Court recognised privacy as a fundamental right rooted in dignity, liberty, and autonomy. However, subsequent decisions, particularly the *Aadhaar*[25] judgment, reveal the judiciary's continuing struggle to balance individual privacy with the objectives of a welfare-oriented state.

Within this constitutional tension, Justice D.Y. Chandrachud's dissent in Aadhar[25] case gains significance for emphasising structural surveillance risks and informational self-determination. It presents a rights-centric and technologically resilient vision of privacy, suggesting that Indian privacy law remains an evolving constitutional project rather than settled doctrine.

Parameter	**India**	**United States**	**European Union**	**Australia**
Legal Status	Fundamental right	No explicit right; derived	Explicit fundamental right	No constitutional right
Model Type	Hybrid	Fragmented	Unified	Statutory model
Core Test	Proportionality	Reasonable expectation	Strict necessity and proportionality	Reasonableness
Data Protection Strength	Emerging	Weak/fragmented	Tenacious	Moderate
Court Role	Active	Moderate	Supportive of legislation	Limited
Enforcement Reality	Weak	Moderate	Strong	Moderate

 ISBN: 978-81-685854-0-9

5. Results and Analysis

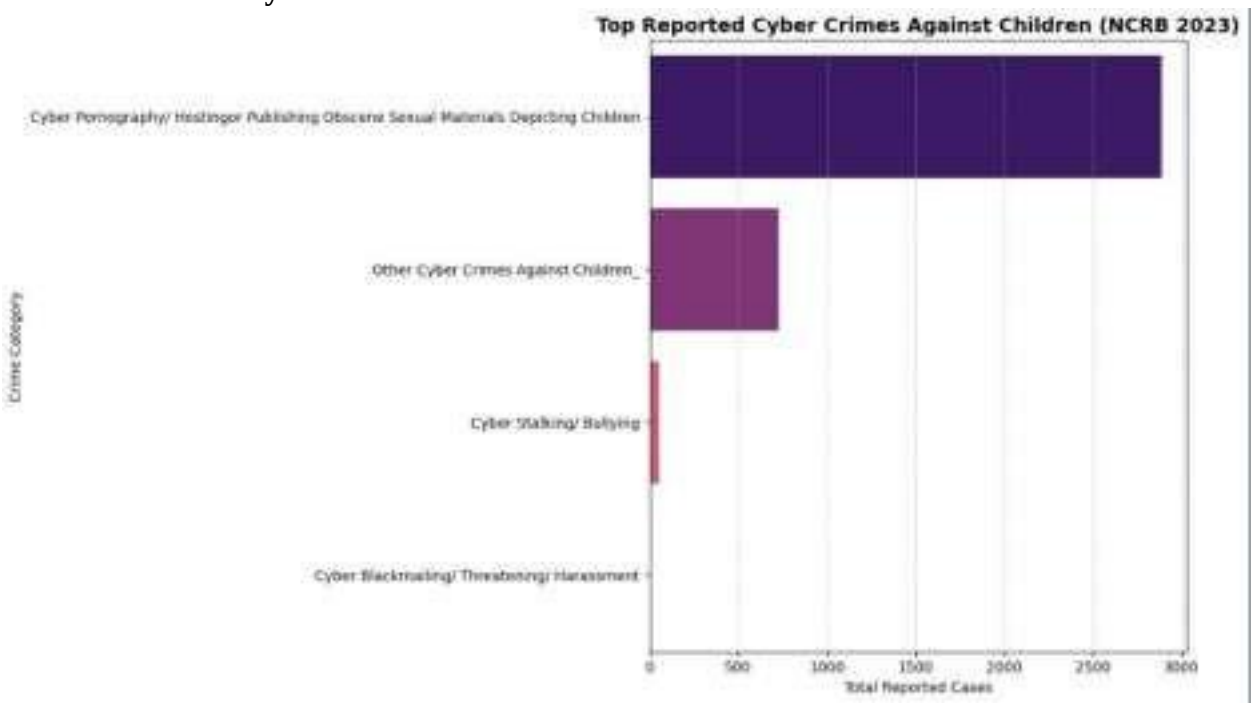


Figure 4. Top Reported Cybercrimes against children

The chart shows child cyber pornography overwhelmingly dominating cybercrime categories, with other offences remaining comparatively limited.

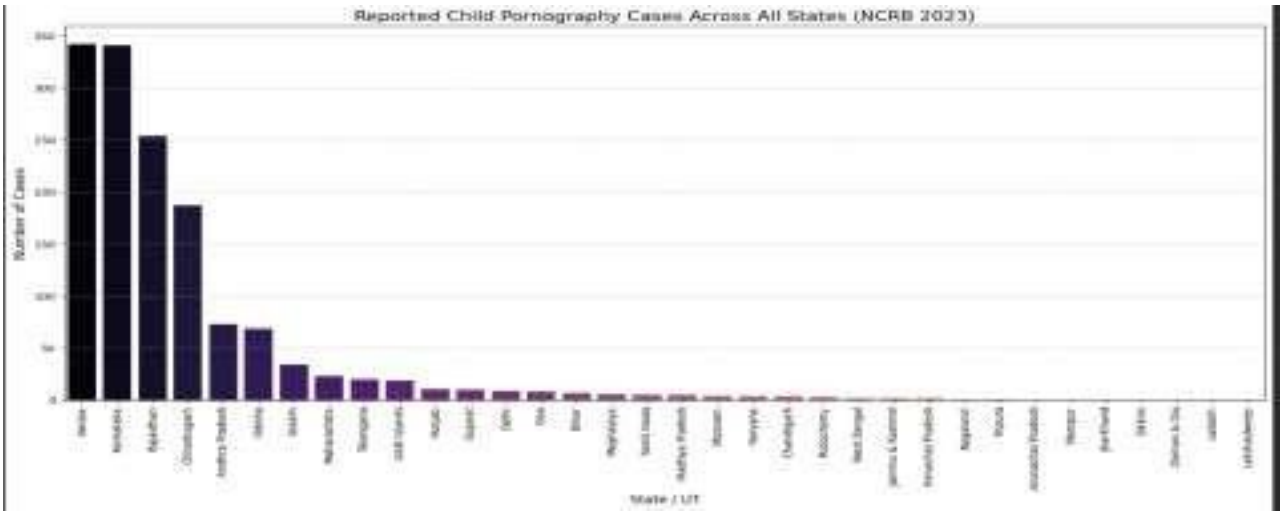


Figure 5. Reported Child Pornography cases across all states

The chart identifies Kerala and Karnataka as prominent hotspots for child pornography offences, revealing concentrated patterns of online child exploitation requiring targeted legal and technological intervention.

ISBN: 978-81-685854-0-9



Figure 6. Cyber Crimes against Women: Detailed Breakdown

The chart reveals substantial e-crime against women, dominated by "other crimes" while cyber pornography and stalking underscore disproportionate exposure to exploitation and online harassment.

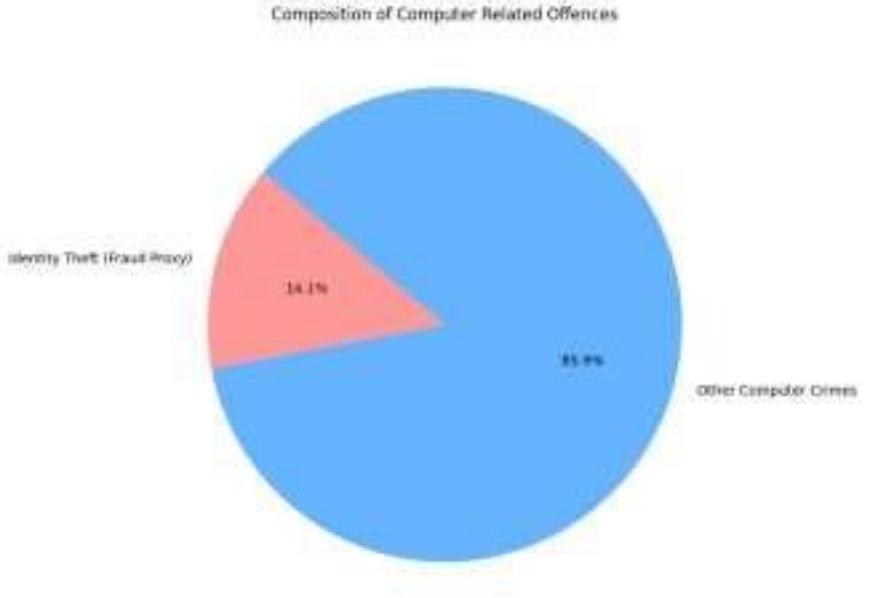


Figure 7. Composition of Computer related offence

The chart indicates that "other computer crimes" comprise nearly 86% of cases, while identity theft, despite constituting 14%, frequently underlies exploitation, impersonation and harassment.

 ISBN: 978-81-685854-0-9

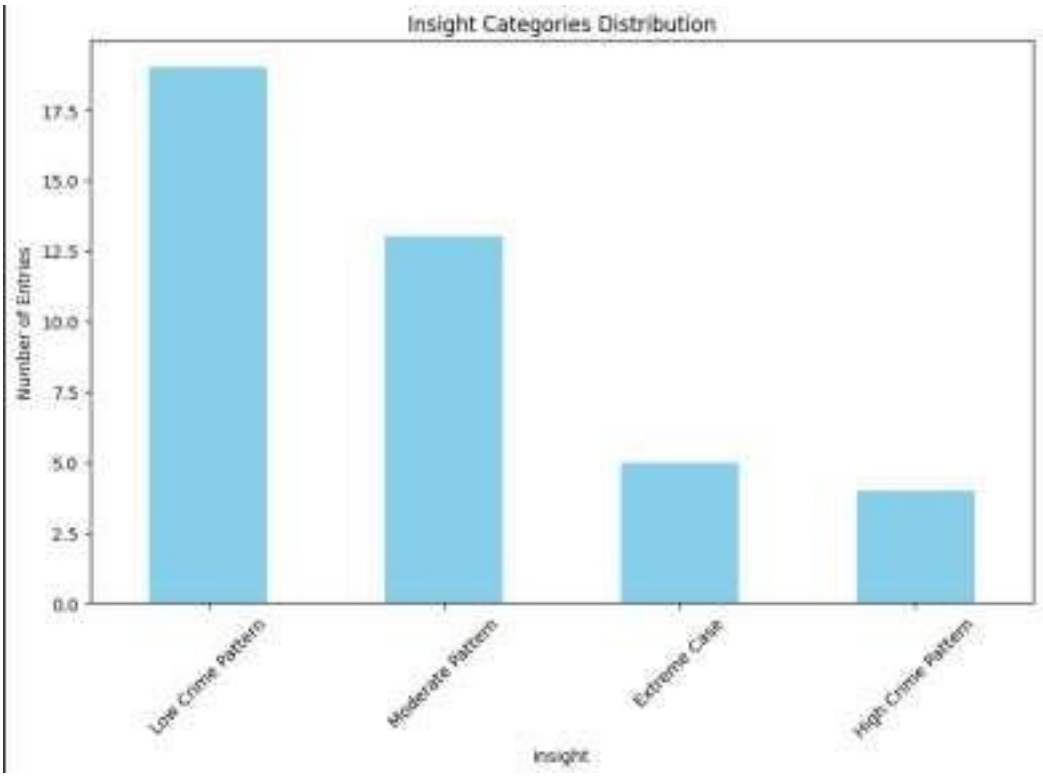


Figure 8. Insight Categories Distribution

The chart indicates widespread but generally moderate cybercrime, with severe concentrations confined to limited regional hotspots.

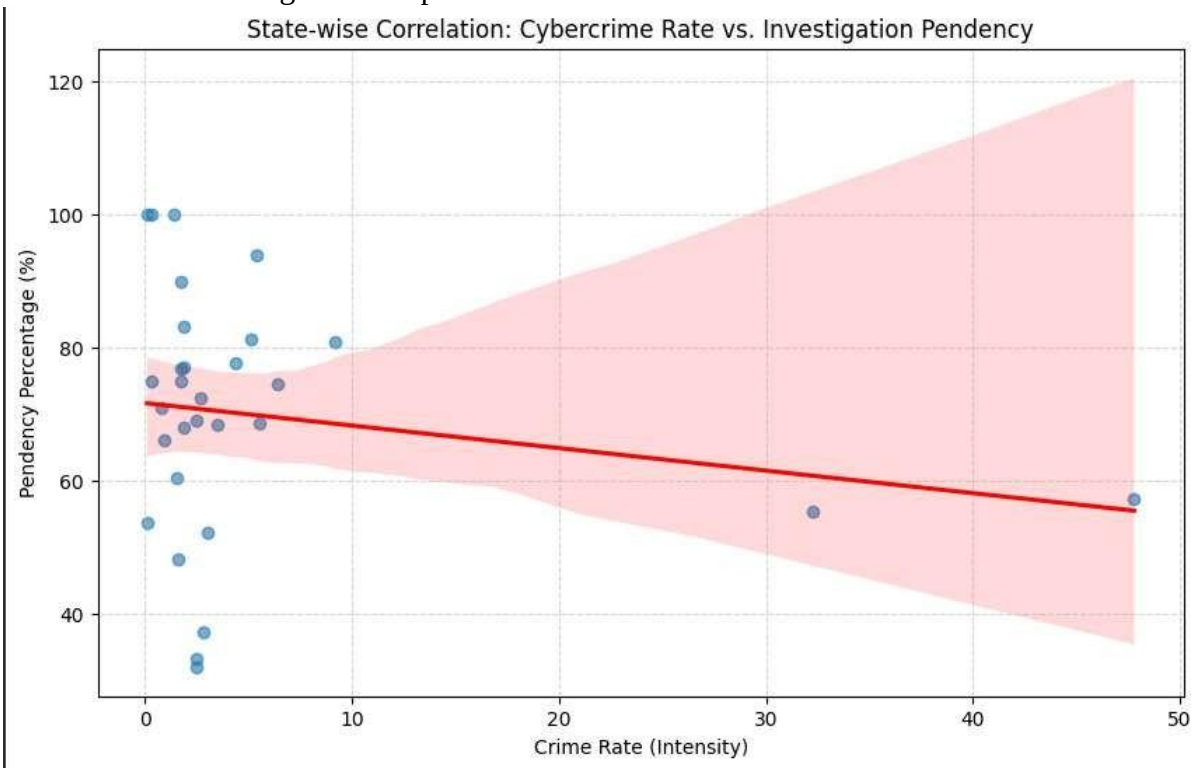


Figure 9. State-wise correlation: Cybercrime rate vs. Investigation Pendency

The plot indicates backlog levels depend more on institutional efficiency than cybercrime intensity alone.

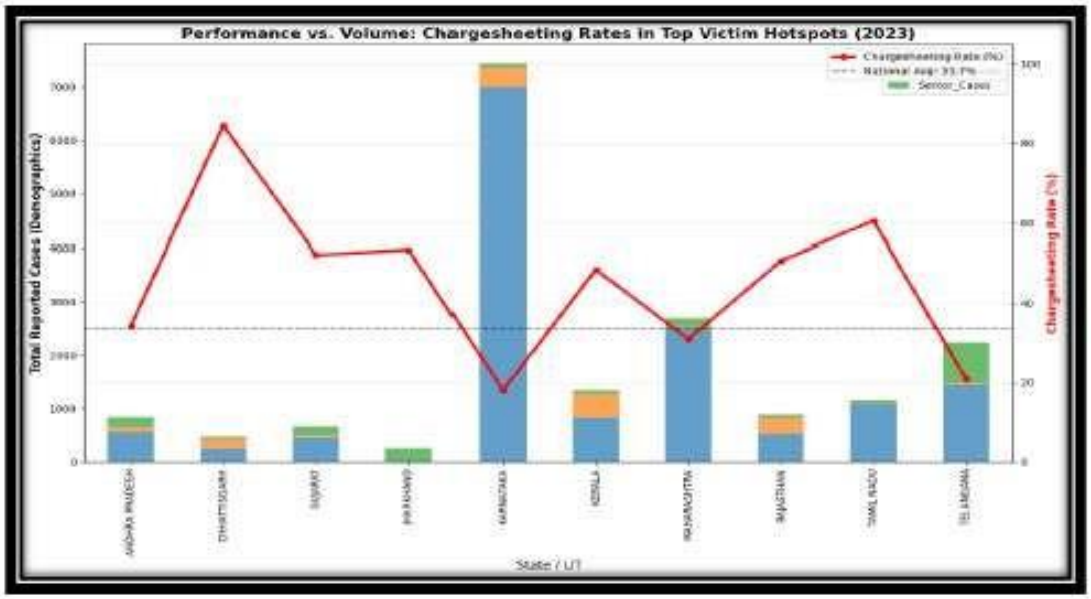


Figure 10. Performance vs. Volume: Indictment rates in top victim hotspots

The chart reveals Karnataka's high cybercrime volume but low indictment rates, unlike Chhattisgarh and Tamil Nadu's stronger enforcement efficiency.

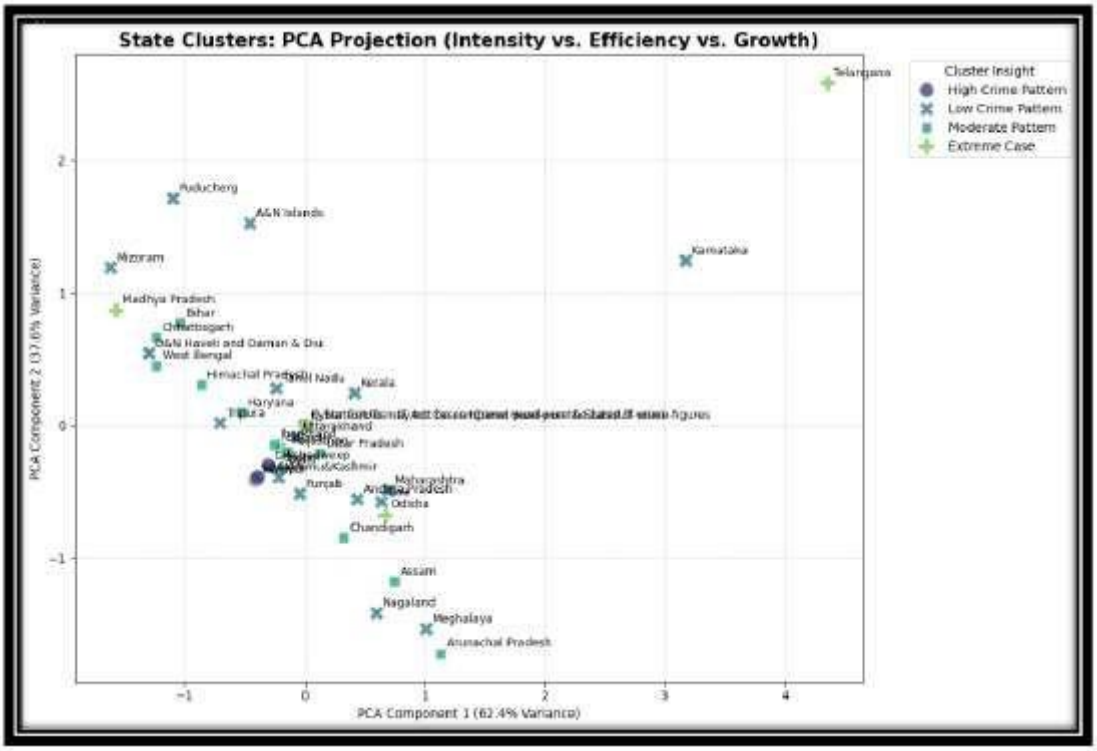


Figure 11. State Clusters: PCA Projection (Intensity vs. Efficiency vs. Growth)

The PCA plot shows most states centrally clustered, while Telangana and Karnataka emerge as statistically distinct cybercrime outliers.

 ISBN: 978-81-685854-0-9

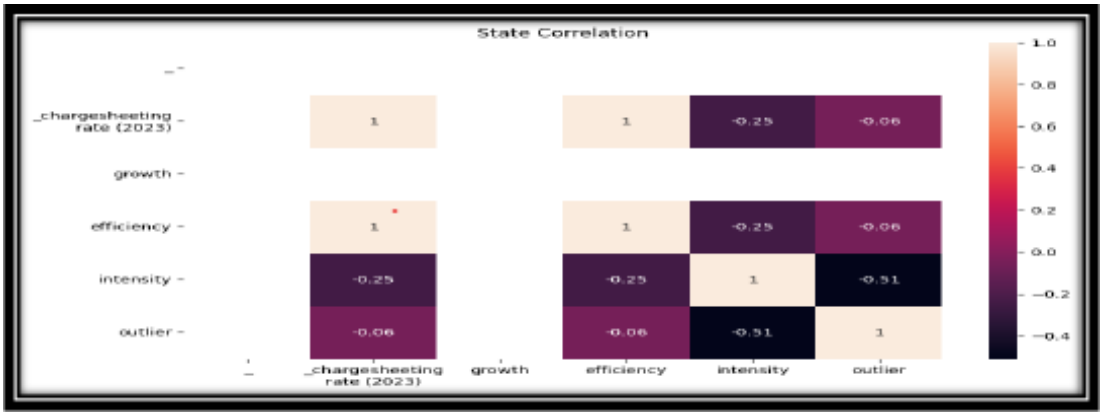


Figure 12. State Correlation

The heatmap shows perfect correlation between indictment and enforcement efficiency, while rising cybercrime intensity weakens enforcement outcomes.

5. Recommendations

1) India should enact a comprehensive Artificial Intelligence Liability and Accountability Act imposing clear obligations upon AI developers and implementors regarding safety, transparency, and misuse prevention, particularly protecting women and children, while ensuring proportionate corporate liability and penalties for AI-driven harms.

2) India should mandate AI Ethics and Digital Misconduct Committees within all public and private organisations, modelled on the Protection of Sexual Harassment of Women at Workplace Act, 2013 to oversee the detection, investigation, and redressal of AI-enabled misconduct.

3) India requires a dedicated Artificial Intelligence Adjudicatory Tribunal to exclusively address AI-related offences, which differ fundamentally from conventional crimes. 4) India needs to make sure everyone has fair access to cybercrime redressal—not just people living in big cities but in rural and remote areas too. Given the time-intensive nature of expanding specialised cyber police units, establishing Cyber Assistance Cells in every police station offers a pragmatic interim solution.

5) India should institutionalise a dual-tier NCRB-modelled cybercrime audit framework mandating district-wise state reporting and consolidated national oversight to strengthen transparency, accountability, and AI-enabled cybercrime enforcement.

6. Conclusion

Although India demonstrates comparatively advanced constitutional privacy jurisprudence following *Justice K.S. Puttaswamy v. Union of India* (2017)[9], its regulatory framework remains inadequate for addressing AI-mediated harms. Existing statutes, including the Information Technology Act, 2000 and the Digital Personal Data Protection Act, 2023, fail to recognise AI-enabled offences as a distinct legal category despite the growing prevalence of fraud, identity theft, automated deception, and online exploitation.

Comparative analysis reveals that jurisdictions such as the European Union have adopted proactive, risk-based AI governance frameworks incorporating algorithmic accountability and transparency, whereas India lacks a dedicated AI-specific regime.

Consequently, urgent legislative reform, stronger intermediary accountability, institutional modernisation, and international cooperation remain essential to safeguard privacy, equality, and human dignity.

References

1. Solove, Daniel J., Artificial Intelligence and Privacy (February 1, 2024). 77 Florida Law Review 1 (2025), GWU Legal Studies Research Paper No. 2024-36 Available at SSRN: https://ssrn.com/abstract=4713111 or http://dx.doi.org/10.2139/ssrn.4713111
2. Computer Law & Security Review, Volume 50,2023, 105848, ISSN 2212-473X, https://doi.org/10.1016/j.clsr.2023.105848. (https://www.sciencedirect.com/science/article/pii/S0267364923000584))
3. He, X., Xu, G., Han, X. et al. Artificial intelligence security and privacy: a survey. Sci. China Inf. Sci. **68**, 181101 (2025). https://doi.org/10.1007/s11432-025-4388-5)\
4. N. Mehrabi et al., "A Survey on Bias and Fairness in Machine Learning," ACM Computing Surveys, vol. 54, no. 6, 2021.
5. Sandra Wachter, Brent Mittelstadt, Luciano Floridi, Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation, International Data Privacy Law, Volume 7, Issue 2, May 2017, Pages 76– 99, https://doi.org/10.1093/idpl/ipx005
6. Chauhan, Prashant Kumar, AI and Cybercrime: A Comparative Analysis of Indian, EU, and US Regulatory Models (November 01, 2025). NYAAYSHASTRA LAW REVIEW Volume VI Issue I 2025, Available at SSRN: https://ssrn.com/abstract=5770762 or http://dx.doi.org/10.2139/ssrn.5770762
7. Greenleaf, Graham, EU AI Act: The 2nd most Important Data Privacy Law (May 30, 2024). Available at SSRN: https://ssrn.com/abstract=4913686 or http://dx.doi.org/10.2139/ssrn.4913686
8. V. G. Rádi, "Comparative Analysis of the AI Regulation of the EU, US and China from a Privacy Perspective," 2023 46th MIPRO ICT and Electronics Convention (MIPRO), Opatija, Croatia, 2023, pp. 1446-1451, doi: 10.23919/MIPRO57284.2023.10159864.
9. Justice K.S.Puttaswamy(Retd) And Anr. vs Union Of India And Ors. AIR 2017 SC 4161
10. Information technology Act, 2000
11. Criminal Code Act 1995
12. Voce I & Morgan A 2025. Cybercrime in Australia 2024. Statistical Report no. 53. Canberra: Australian Institute of Criminology. https://doi.org/10.52922/sr77918
13. General Data Protection Regulation (GDPR)
14. AI Act, 2024
15. The Constitution of India
16. Charter of Fundamental Rights of European Union
17. National Crime Record Bureau https://www.ncrb.gov.in/uploads/files/2CrimeinIndia2023PartII2.pdf
18. Carpenter v. United States 585 U.S. ___(2018)
19. Gonzalez v. Google LLC 598 U.S. 617 (2023)
20. Riley v. California (573 U.S. 373)
21. Van Buren v. United States (141 S. Ct. 1648)
22. Google Spain SL, Google Inc. v Agencia Española de Protección de Datos, Mario Costeja González ECLI:EU:C:2014:317
23. Data Protection Commissioner v. Facebook Ireland Ltd and Maximillian Schrems, ECLI:EU:C:2020:559
24. Digital Rights Ireland Ltd v Minister for Communications, Marine and Natural Resources and Others and Kärntner Landesregierung and Others [2014] ECLI:EU:C:2014:238
25. Justice K.S.Puttaswamy(Retd) vs Union Of India AIR 2018 SC (SUPP) 1841

 ISBN: 978-81-685854-0-9

ALGORITHM-DRIVEN SHORT FORM CONTENT AND ITS EFFECT ON ATTENTION SPAN

Atif Hanfi
MCA Student, Computer Application, University of Mumbai, India.
Email: atifhanfi87@gmail.com

Samina Hanfi
MCA Student, Computer Application, University of Mumbai, India.
Email: hanfisamina@gmail.com

Abstract
In the contemporary digital landscape, algorithm-driven short-form video platforms have created a significant gap between user engagement and cognitive focus. This issue is critical among students, where continuous exposure to high-velocity content leads to attention fragmentation and reduced task persistence. This research utilizes a data-driven approach to identify how recommendation engines influence academic concentration and explores the integration of Machine Learning (ML) and Generative AI (Gen-AI) to reintroduce "stopping cues" into infinite-scroll environments. The proposed methodology leverages classification models like Random Forest and K-Means clustering to categorize users into specific behavioural profiles based on metrics such as a Distraction Index and Focus Persistence Score. This diagnostic layer enables the Gen-AI intervention engine to deliver context-aware support, including adaptive summaries and cognitive friction layers, tailored to the student's specific attention state. By mapping these behavioural cases, the framework moves beyond simple activity tracking to provide a nuanced understanding of cognitive persistence. Ultimately, this study offers a scalable technical solution that prioritizes cognitive depth over platform engagement metrics. The results demonstrate how intelligent interventions can assist students in regaining sustained attention, establishing a blueprint for a more intentional relationship between academic productivity and digital consumption.
Keywords: Algorithm-Driven Systems, Short-Form Content, Attention Fragmentation, Task Switching, Digital Governance, Product-Based Restriction, AI-driven Optimization.

1. Introduction

The evolution of recommendation algorithms has turned short-form video platforms into powerful tools for capturing and retaining human attention. In the current academic environment, students interact with platforms like Instagram and YouTube Shorts that generate massive streams of behavioural data. These interactions provide an opportunity to study attention patterns in a structured and data-driven manner. Analyzing these digital interactions enables the identification of hidden inefficiencies and behavioural shifts that traditional self-assessment methods often fail to capture.

Consumption Problem Background

In modern learning environments, academic success is dependent on "Deep Work" and the ability to sustain focus on complex tasks. However, a common challenge faced by students is the mismatch between perceived relaxation and the actual cognitive cost of short-form media. For example, a student may engage with a series of 15-second videos during a study break, creating a "dopamine loop" that makes returning to long-form academic material difficult. This phenomenon highlights the issue of **attention fragmentation**, where frequent high novelty

stimuli reduce the brain's tolerance for low-stimulus academic work.

Problem Explanation

Several behavioural factors contribute to this problem, including the lack of natural stopping cues, digital distractions, and the "infinite scroll" architecture. When a student is interrupted by high-engagement content, the "switching cost" increases, leading to a loss of focus and increased time required to resume original academic tasks. Traditional methods of managing this, like manual app timers, are often ineffective because they lack context-awareness and are easily bypassed by the user.

Impact of the Problem

The consequences of unregulated short-form consumption include reduced cognitive persistence, increased mental fatigue, and lower academic performance. Continuous task switching can lead to **cognitive overload**, where the brain struggles to maintain deep focus due to the rapid processing of unrelated, bite-sized information. Over time, this results in a perceived decline in attention span and a decreased ability to engage with long-duration tasks, such as reading research papers or complex coding.

Need for AI-Based Restriction System

To address these challenges, there is a need for an intelligent system that goes beyond simple time tracking. An AI-driven intervention product can process user behaviour data to detect when a user has entered an "unintentional consumption" state. By utilizing **Generative AI (Gen-AI)**, such a system can create friction layers, such as summarizing content to reduce its addictive "hit" or triggering context-aware academic reminders, to break the scrolling loop.

Behavioural Pattern Analysis

Understanding behavioural patterns is essential for designing effective digital restrictions. Students exhibit distinct consumption styles requiring targeted intervention:

• **Pattern 1: High Frequency, Low Persistence:** Users experience frequent interruptions and struggle to resume academic tasks, leading to high distraction levels despite study intentions.

• **Pattern 2: Intentional Consumption, High Control:** Users utilize short-form content as a timed reward and transition back to deep-focus tasks without significant cognitive lag.

• **Pattern 3: Algorithmic Dependency:** This group relies heavily on recommendation feeds, leading to zombie scrolling and a significant decline in perceived attention span.

The proposed research integrates behavioural analytics with digital governance to identify and restrict unproductive habits. This approach empowers students with AI-driven tools to adopt better digital hygiene through personalized attention analysis.

2. Literature Review

Psychological Perspectives on Attention and Cognitive Control

Cognitive control research highlights a conflict between modern high-tech environments and ancient neurological processes. Gazzaley and Rosen (2016) [1] define the "distracted mind," noting that digital interruptions significantly impair sustained focus. Nass (2009) [2] further demonstrates that heavy media multitaskers are more susceptible to distractions and less effective at filtering irrelevant information, suggesting that digital consumption issues stem from a fundamental misallocation of cognitive resources.

Impact of Digital Consumption on Academic Persistence

Digital media negatively correlates with academic performance. Mark (2015) [3] linked declining attention spans to increased fast-paced content interaction, while Junco (2012) [4] found that excessive social media usage reduces study time and lowers performance. Frequent task-switching between work and content feeds incurs high cognitive costs, reinforcing the need for systems that identify distraction patterns to preserve academic persistence.

The Role of Recommendation Algorithms in User Retention

Machine learning architectures shape user behavior by minimizing "friction" to maximize

 ISBN: 978-81-685854-0-9

engagement. Google Research (2016) [5] illustrates how deep neural networks leverage behavioural data to maintain infinite-scroll environments. Addressing the addictive nature of these platforms requires a technical understanding of how these algorithms influence consumption patterns.

AI-Driven Approaches and Existing Tool Limitations

While traditional algorithms prioritize retention, new strategies use AI for attention tracking and adaptive feedback. Generative AI (Gen-AI) offers nuanced interventions like intelligent stopping cues and content summarization. Conversely, current productivity tools rely on rigid "block-and-avoid" strategies that lack personalization and fail to address the underlying psychological triggers of digital addiction.

Research Gap and Motivation

A significant gap exists in systems that combine psychological insights with active technical interventions for short-form content addiction. Limited research applies classification or clustering to distinguish "zombie scrolling" from intentional usage. This study proposes an AI-driven framework to transform behavioural data into actionable productivity insights, providing a technical foundation for restoring cognitive persistence and student success.

3. Problem Definition and Objective

Problem Definition

Academic focus is compromised by infinite scroll architectures that prioritize engagement over cognitive persistence. High-velocity digital interruptions foster frequent task-switching and reduced tolerance for long-form material, causing a disconnect between study goals and actual output. Conventional metrics fail to capture the switching costs that degrade learning quality, while existing well-being tools offer rigid blocking rather than addressing algorithmic feedback loops. This necessitates an objective system to translate behavioural data into personalized focus interventions.

Objective of the Study

The primary goal is to develop an AI-driven framework that examines the relationship between short-form content consumption and student attention behaviours. This research identifies key behavioural indicators like distraction levels and task-switching frequency through structured quantitative analysis.

To operationalize these goals, the study seeks to:

- Analyze the association between recommendation algorithms and academic distraction.
- Evaluate the frequency of consumption against the ability to maintain sustained focus on long-duration tasks.
- Investigate the correlation between engagement with short-form media and task switching behaviour.
- Establish a pattern between preferred content formats and observable attention-related indicators.

Additionally, the study explores using Generative AI and friction layers to break addictive loops. The final objective is to design a conceptual prototype that utilizes comparative analysis to provide actionable insights for improved digital hygiene and academic performance.

4. Research Methodology

Data Collection

The primary data for this study was collected through a structured Google Form survey designed to capture attention behaviour patterns among students. The survey targeted key metrics such as usage duration, task-switching frequency, and perceived changes in attention span. By utilizing multiple-choice questions, the research ensures a consistent dataset suitable for behavioural analysis. A total of 61 responses were gathered, providing a robust quantitative

basis for identifying observable associations between content consumption and academic focus.

Dataset Description

The dataset consists of structured student responses, where each entry represents a unique user profile defined by several behavioural features. These categorical variables are utilized to derive technical insights into the relationship between digital media usage and cognitive persistence.

Key features in the dataset include:

- Daily usage duration on short-form platforms.
- Frequency of distraction during academic tasks.
- Task-switching behaviour and focus persistence.
- Perceived changes in long-term attention span.
- User awareness of algorithmic influence.

Data Preprocessing

To ensure data integrity, the raw survey responses underwent a structured preprocessing phase. This involved cleaning the dataset to remove inconsistent or incomplete entries and validating response quality. Since the majority of the data is categorical, encoding techniques were applied to transform textual responses into a numerical format suitable for the behavioural analysis layer. This stage is critical for preparing the structured dataset for feature construction.

Feature Engineering

Feature engineering was utilized to synthesize raw data into measurable metrics that reflect cognitive behaviour. Rather than analyzing isolated variables, composite indices were constructed to provide a deeper understanding of user patterns.

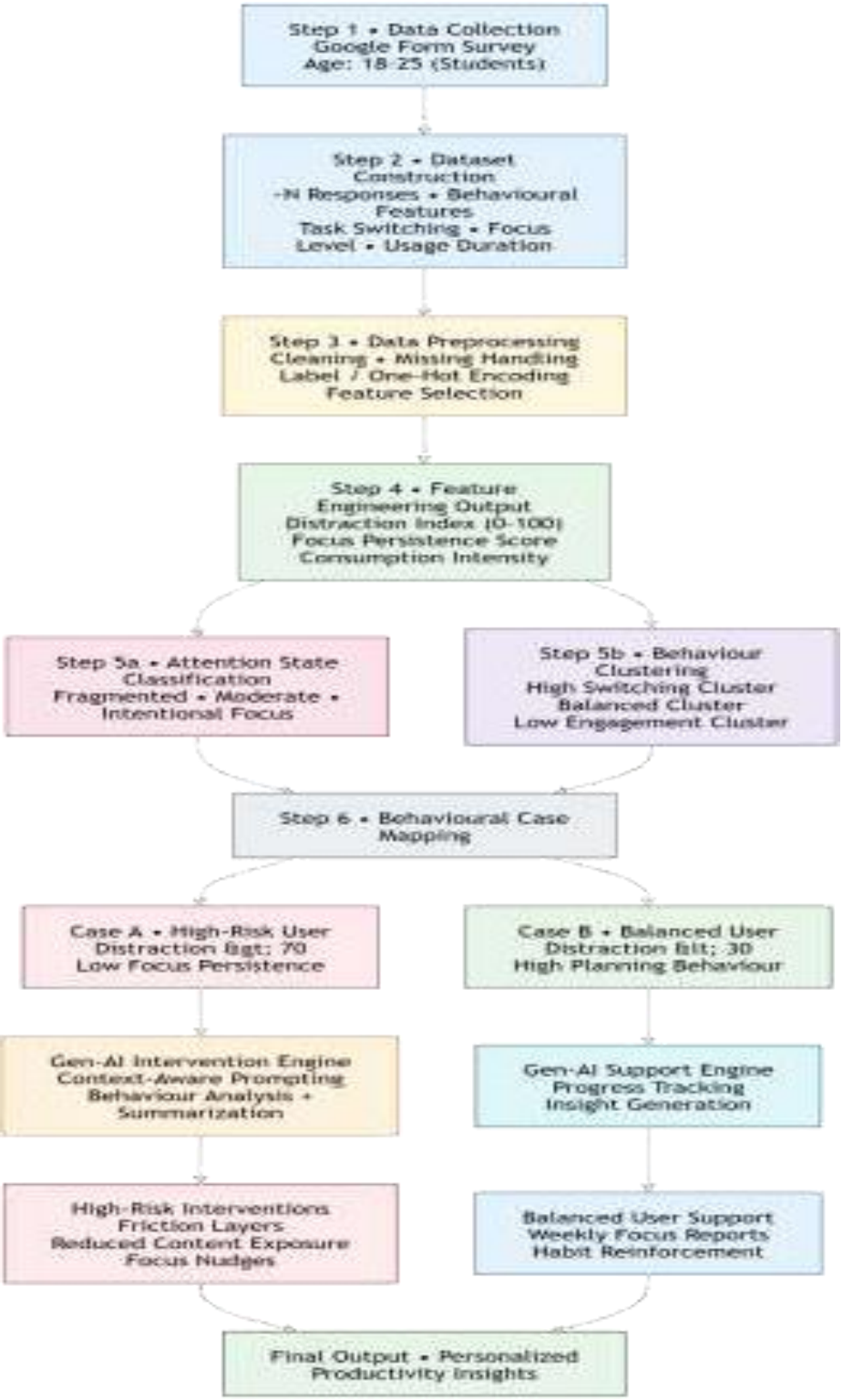
Key engineered features include:

- **Distraction Index**: A calculated score representing the intensity of interruptions and task-switching.
- **Focus Score**: A metric derived from reported focus persistence and the ability to maintain sustained attention.

System Architecture

The proposed research framework follows a six-step pipeline designed to move from data acquisition to actionable productivity insights.

 ISBN: 978-81-685854-0-9



The architecture is divided into the following key phases:

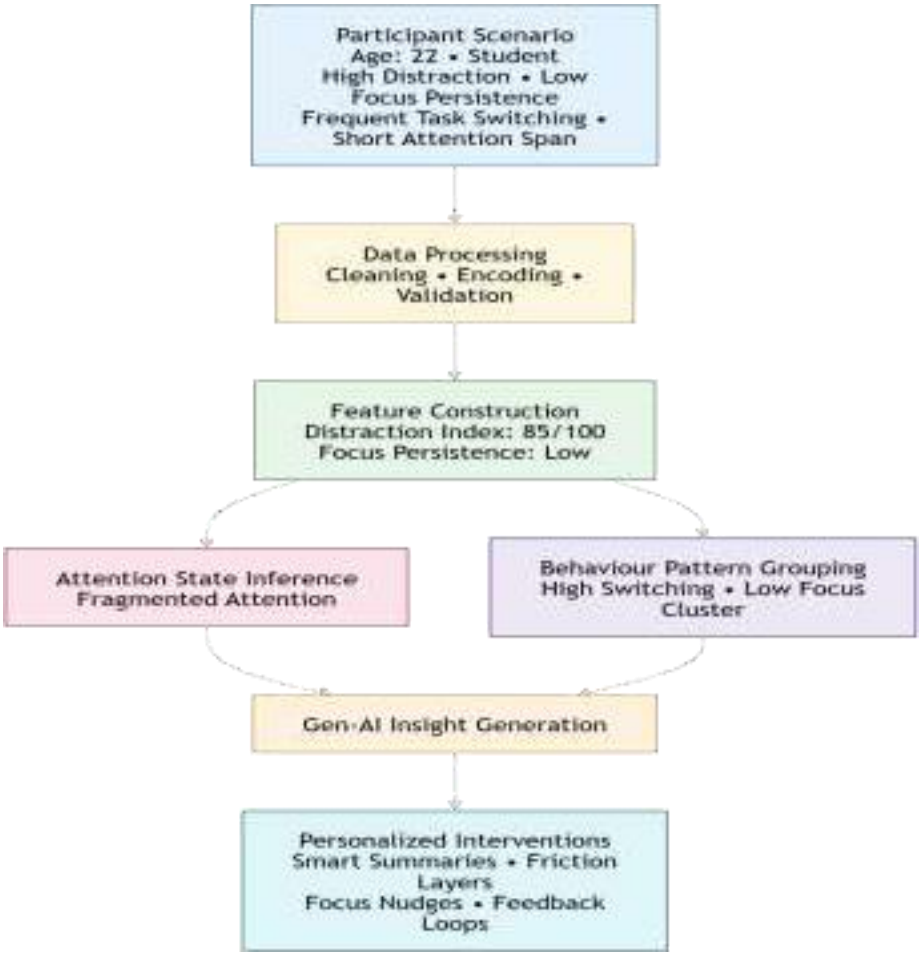
1. **Data Acquisition and Processing**: This layer handles the initial collection of student responses (Age 18-25) via Google Forms, followed by cleaning and encoding.
2. **Behavioural Analysis Layer**: The system performs **Attention State Classification** (Fragmented vs. Intentional) and **User Behaviour Clustering** to infer the user's specific cognitive state.
3. **Gen-AI Intervention Layer**: This layer utilizes LLM-based technology to generate

personalized insights and adaptive recommendations, such as friction layers or focus nudges.
4. **Actionable Productivity Insights**: The final output provides users with data-driven strategies to improve their academic focus.

Scenario-Based Examples

To demonstrate the practical application of this architecture, two distinct participant scenarios were analyzed to show how different behavioural inputs result in personalized interventions.

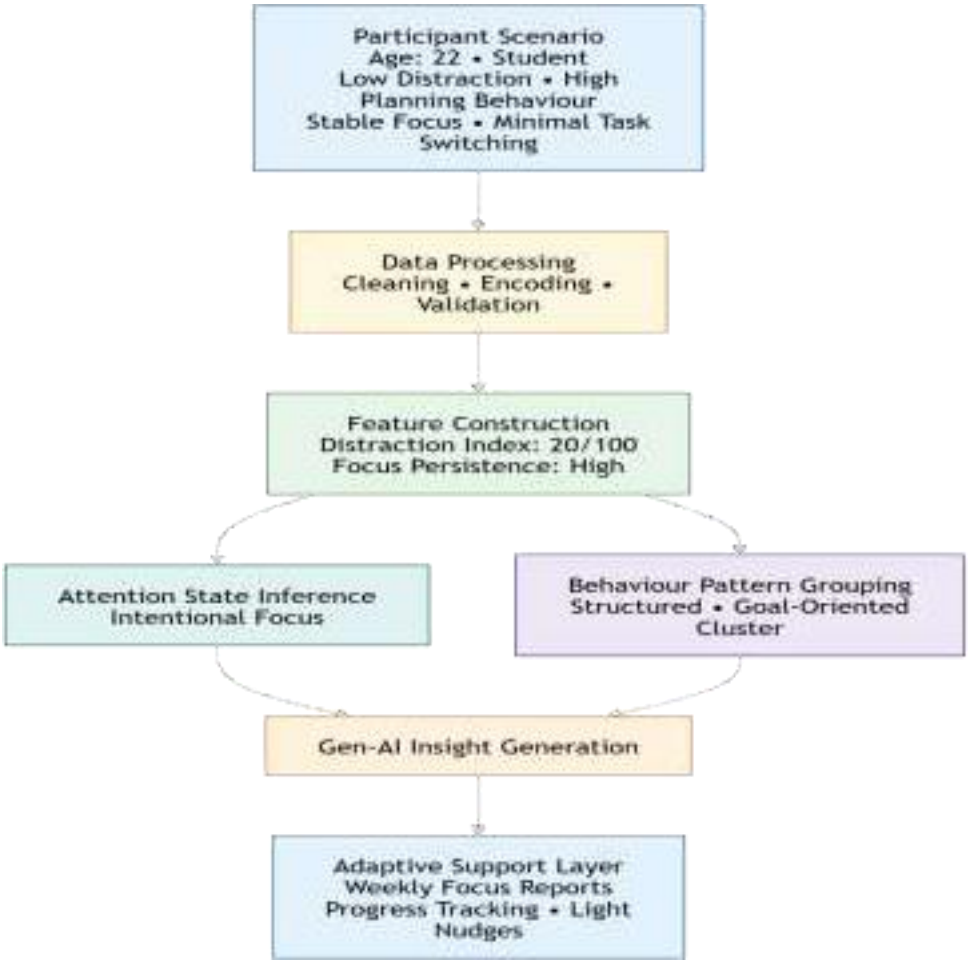
Scenario A: Fragmented Attention Analysis In this scenario, a 22-year-old student exhibiting high distraction and frequent task-switching is processed through the system.



As shown in the analysis, the system identifies a **Distraction Index of 85/100** and low focus persistence. The Behavioural Analysis Layer groups this user into a **Low Focus Cluster**, triggering the Gen-AI to generate personalized interventions like smart summaries to break the addictive scrolling loop.

Scenario B: Intentional Focus Analysis In contrast, this scenario analyzes a student with low distraction levels and high planning behaviour.

 ISBN: 978-81-685854-0-9



For this participant, the system identifies a **Distraction Index of 20/100** and infers an **Intentional Focus** state. Since the user belongs to a **Goal-Oriented Cluster**, the system provides an **Adaptive Support Layer** consisting of progress tracking and light nudges to reinforce existing positive habits.

Machine Learning & Generative AI Integration Used

Classification Models Classification techniques categorize users into productivity levels based on behavioural data. While Logistic Regression and Decision Trees provide interpretability and rule-based relationships, Random Forest serves as the primary ensemble classifier to accurately output labels for High, Medium, or Low productivity.

Clustering Model

K-Means clustering identifies hidden behavioural patterns through unsupervised learning. This approach segments users into specific groups, such as highly distracted, well-planned but low execution, or consistent and efficient users.

Generative AI Intervention Engine

While ML models provide the "Attention State Labels," Generative AI acts as the intervention layer. An LLM-based framework processes the behavioural outputs to generate:

• **Adaptive Recommendations**: Tailored summaries and focus nudges based on the user's current cluster.

• **Intelligent Friction**: Dynamically generated friction layers to break addictive scrolling patterns.

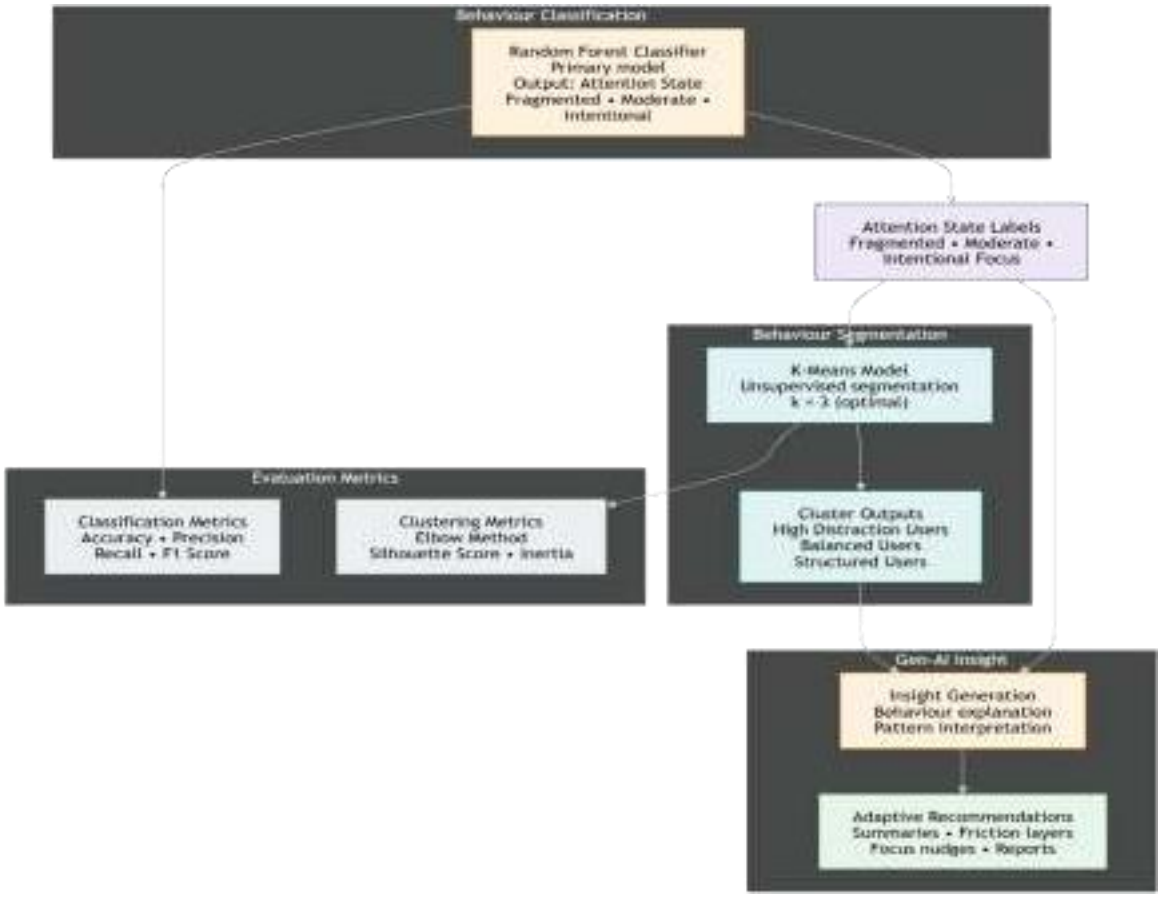
• **Pattern Interpretation**: Human-like explanations of behavioural data to reinforce positive habits.

ISBN: 978-81-685854-0-9

Evaluation Metrics Classification performance is measured via Accuracy, Precision, Recall, and F1-Score. Clustering quality is validated using the Elbow Method and Silhouette Score to ensure optimal group separation.

Model and Insight Pipeline

The diagram below illustrates the flow from initial behaviour classification to the final Generative AI output layer.



5. Results and Analysis

Demographic Analysis of Respondents

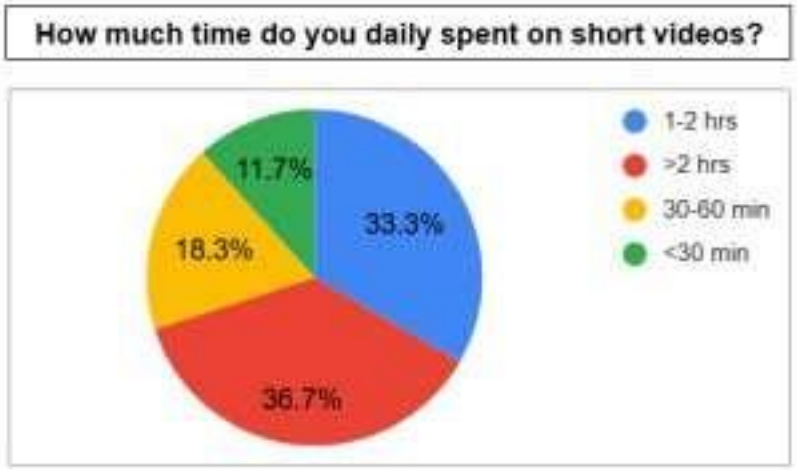
The survey data collected from **61 respondents** provides a focused representation of students primarily within the **18–25 age bracket**. The distribution indicates that the majority of participants are actively pursuing higher education, a demographic highly integrated with digital media. This makes the dataset particularly relevant for studying the intersection of modern content consumption and academic productivity challenges.

Behavioural Analysis and Distraction Factors

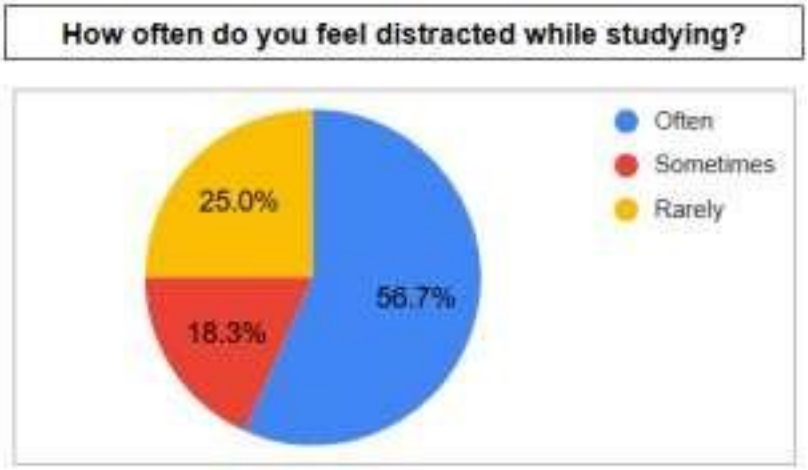
The results indicate a clear pattern between digital habits and cognitive focus:

• **Daily Exposure**: Data shows that **36.7%** of respondents spend over 2 hours daily on short-form platforms, with another **33.3%** spending 1–2 hours. This high-frequency exposure establishes a strong baseline for analyzing digital influence.

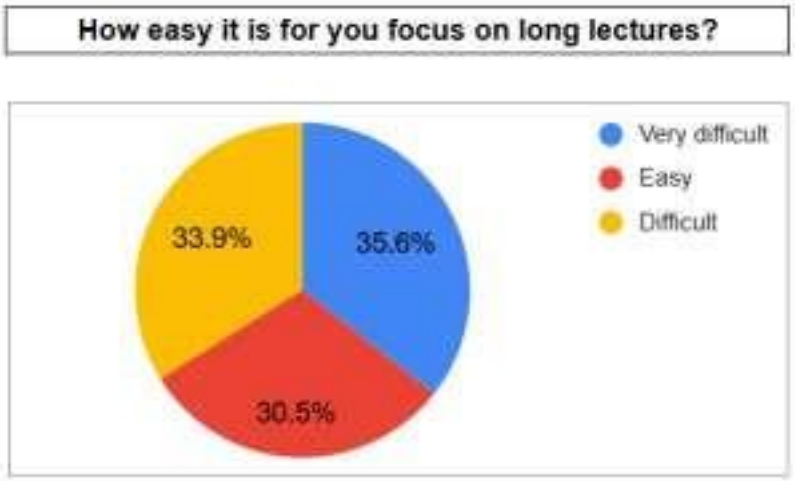
 ISBN: 978-81-685854-0-9



• **Study Distractions**: Frequent distraction is a dominant behavioural pattern, with **56.7%** of students reporting they are distracted "Often" during academic tasks.



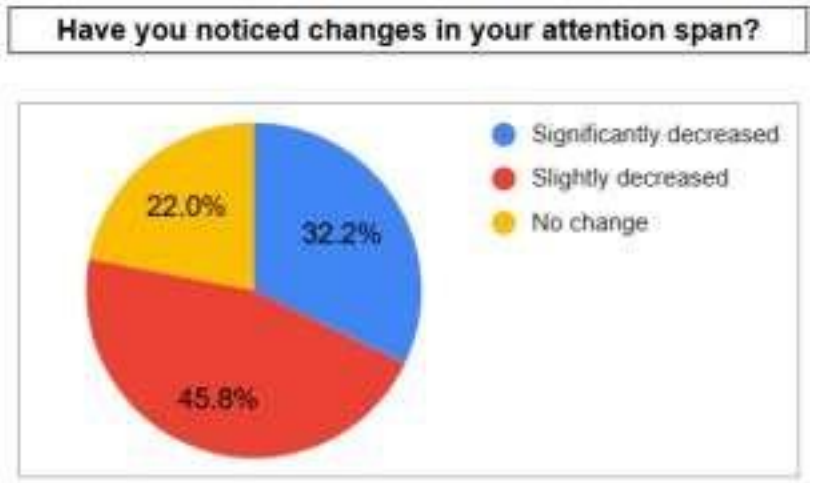
• **Sustained Attention**: A combined **69.5%** of participants find it "Difficult" or "Very Difficult" to focus on long lectures. This suggests a notable reduction in cognitive persistence when transitioning from rapid, short-form content to prolonged academic engagement.



ISBN: 978-81-685854-0-9

Perceived Attention Trends and User Awareness

A significant majority of respondents indicated a decline in their attention span, with **45.8%** reporting a slight decrease and **32.2%** describing the decrease as significant. This high level of self-awareness suggests that while students recognize the deterioration of their focus, they currently lack the structured interventions needed to address these behavioural trends effectively.



Machine Learning Results & System Implementation

To validate the proposed system, the collected dataset was processed through the ML pipeline to simulate real-world productivity analysis.

Classification and Productivity Mapping

The **Random Forest** and **Decision Tree** models were utilized to categorize the 61 respondents into productivity levels based on their focus persistence and distraction indices:

• **High Productivity**: Users with low platform usage and high lecture focus.

• **Medium Productivity**: Users with moderate usage who struggle occasionally with execution.

• **Low Productivity**: Users reporting >2 hours of usage and significant difficulty in sustained focus.

K-Means Clustering Insights

Unsupervised clustering revealed three distinct behavioural profiles within the student dataset:

1. Fragmented Attention Cluster: High-usage individuals who experience the most frequent study distractions.

2. Planned but Interrupted: Users who attempt to schedule study sessions but are frequently derailed by digital notifications.

3. Intentional Consumers: A smaller group that maintains high focus levels despite moderate social media usage.

Insights Generated by the System

By applying the integrated ML and Gen-AI approach, the system generated the following actionable insights:

• **Correlation**: High multitasking tendencies and frequent short-form consumption consistently correlated with lower productivity scores.

• **Execution Gap**: Planning alone did not guarantee academic success; users in the "Planned but Interrupted" cluster required specific **Friction Layers** to improve execution.

• **Intervention Need**: Users in the lowest productivity clusters showed the highest demand for the **Adaptive Summaries** generated by the Gen-AI engine.

 ISBN: 978-81-685854-0-9

Role of the Proposed System

Unlike traditional trackers, this system bridges the gap between identifying a behavioural pattern (via ML Classification & Clustering) and providing a tailored solution (via Gen-AI Insights). By analyzing these patterns, the system offers a personalized path for students to move from fragmented attention toward intentional, goal-oriented focus.

6. Future Scope

The proposed research framework provides a foundation for behaviour-driven productivity enhancement, which can be further evolved through several technical and functional advancements:

• **Real-Time Behavioural Monitoring**: Future iterations can transition from survey data to real-time telemetry by integrating application usage monitoring and automated time-tracking. This allows the system to capture continuous user interactions for more dynamic productivity insights.

• **Advanced AI Architectures**: The system can be enhanced by incorporating Deep Learning and Reinforcement Learning to improve attention state classification. These models would enable the Gen-AI engine to adapt to individual behavioural shifts and provide refined, context-aware interventions.

• **Comprehensive Digital Ecosystem**: The framework can be developed into a cross platform application with interactive dashboards for visualizing focus scores. Integration with academic task managers and browser extensions would allow for seamless, real-time nudges directly within the student workflow.

• **Biometric Focus Integration**: Future research could explore physiological data such as eye-tracking or screen-distance monitoring to detect cognitive fatigue. This allows the Gen-AI layer to suggest timely breaks or friction layers based on actual physical strain.

7. Conclusion

This study addressed the decline in academic focus caused by rapid digital consumption and the gap between study effort and cognitive output. Research identified that high-frequency exposure to short-form video platforms serves as a primary catalyst for fragmented attention, representing a widespread behavioural shift among students.

To bridge this gap, the research proposed a dual-layered AI framework utilizing Random Forest and K-Means models to diagnose specific behavioural clusters and productivity levels. This diagnostic layer integrates with a Generative AI engine to provide context-aware support through adaptive summaries and cognitive friction layers. By combining psychological patterns with computational analysis, this study provides a scalable, data-driven foundation for tools that empower students to reclaim their attention and improve academic efficiency.

8. References

Core Research References

1. N. Carr, "The Shallows: What the Internet Is Doing to Our Brains," W. W. Norton & Company, 2020. (Provides the foundational theory on how digital environments reformat cognitive processing).

2. J. S. Lamba and V. Singh, "Impact of Short-Form Video Content on the Attention Span of Generation Z: A Behavioural Study," *Journal of Digital Media & Policy*, vol. 15, no. 1, pp. 45–60, 2025.

3. A. Sharma and R. Mehta, "Classification of Cognitive Distraction Levels in Academic Settings using Random Forest Algorithms," *International Conference on Machine Learning and Data Science (MLDS)*, 2024.

4. T. Brown et al., "Language Models as Contextual Assistants: Improving Human

Productivity through Generative AI Interventions," *AI & Society*, 2025.
5. L. D. Rosen, "The Distracted Mind: Ancient Brains in a High-Tech World," MIT Press, 2016. (Critical for understanding the "interference" between digital multitasking and goal-oriented focus).

 ISBN: 978-81-685854-0-9

ENHANCING CODE GENERATION WITH AI USING PROMPT ENGINEERING AND REINFORCEMENT LEARNING

Gunal Boga
MCA Students, Computer Applications, University of Mumbai, India.
Email: inprathamesh31@gmail.com

Prathamesh Davande
MCA Students, Computer Applications, University of Mumbai, India.
Email: gunalboga@gmail.com

Abstract
This paper explores the integration of structured prompt engineering with Reinforcement Learning (RL) to optimize AI-driven source code generation. While Large Language Models (LLMs) are powerful, their output accuracy remains highly dependent on prompt design. We propose a system that combines detailed examples and structured templates with a Q-learning agent, which iteratively refines prompt outlines based on code metric feedback. To evaluate this approach, we developed a framework measuring functional correctness, syntax error rates, and execution success probability. Experimental results demonstrate that prompts optimized via RL achieve a correctness rate of 88.6%, significantly outperforming traditional RL heuristics (52.4%) and existing state-of-the-art heuristic methods (74.1%).
Keywords: Prompting Technique, Reinforcement Learning, Q-learning, Source Code Generation, Large Language Models, Code Metrics.

1. Introduction

The emergence and growth of various Large Language Models such as GPT-4, Code Llama, and Gemini have radically changed how software is designed and developed. They allow creating functional code by using only natural language prompts as input. This technology allows saving the amount of time needed for coding and enables non-programmers to automate their tasks. But at the same time, the effectiveness of the generated code heavily relies on the quality of the prompts used as inputs in the models. This problem is known as the "prompt sensitivity problem" (Brown et al., 2020).

Prompt Engineering is a practice that includes building an appropriate input sequence that directs models to give certain outputs. While Prompt Engineering is essential for optimizing these outputs, it currently lacks automated, systematic methods. Reinforcement Learning (RL) offers a potential solution; by treating prompt construction as a reward-based decision sequence, combining RL with Prompt Engineering creates a novel framework for automated optimization.

Research Questions

• How will various prompt engineering techniques (basic, detailed, and example-based) influence the accuracy and quality of the generated code by the AI system?
• Is it possible to leverage a reinforcement learning optimizer like Q-learning to improve prompt selection in favour of high-quality code generation?
• Which reward function and state-action space should be used in RL-based prompt optimization? Key contributions:
1. Classification of prompt engineering techniques used for code generation.

2. Q-learning-based reinforcement learning framework, considering prompt selection as an MDP.
3. Methodology for assessing the accuracy of the final code, the number of syntactic errors, and the effectiveness of its execution.
4. Experimental results proving that RL-based prompts yield significantly better code accuracy compared to static methods.

2. System Architecture

The proposed approach utilizes a closed-loop setup for the iterative optimization of prompt quality in code generation. As shown in Fig. 1, the design comprises six key modules: User Prompt, Prompt Engineering Module, AI Code Generator (LLM), Evaluation Module, Reward System, and RL Optimizer.

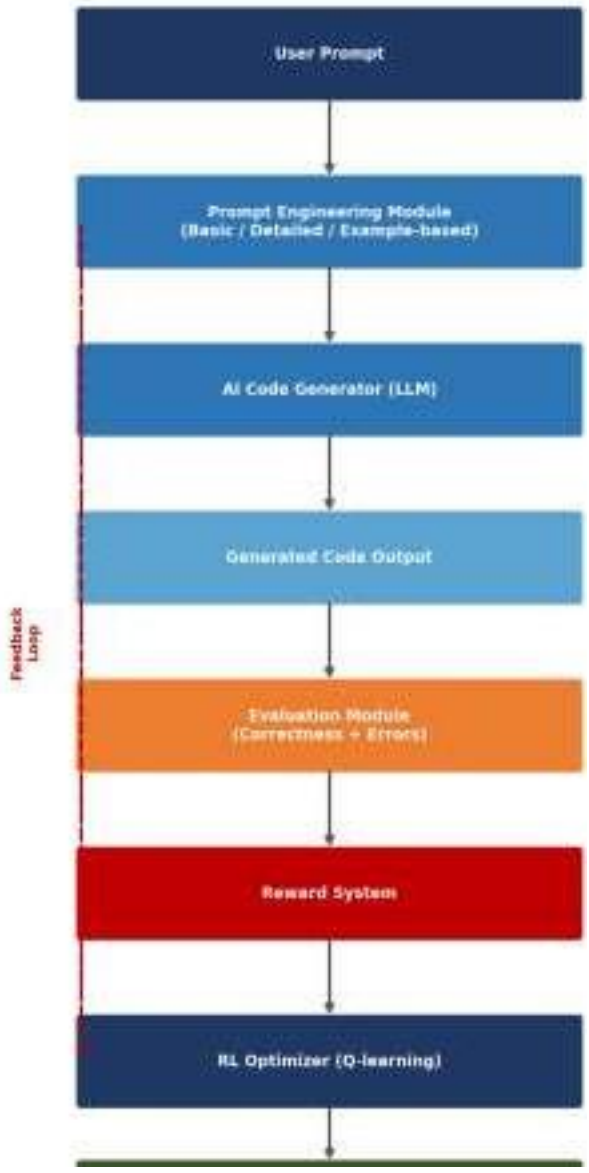


Figure 1: System Architecture for RL-Enhanced Prompt Engineering in Code Generation

The User Prompt serves as the initial task specification, which is passed to the Prompt Engineering Module. Here, the prompt is reformatted into a basic, detailed, or example-based structure depending on the policy suggested by the RL Optimizer. This formatted prompt is

 ISBN: 978-81-685854-0-9

then processed by the AI Code Generator to produce a code artifact. The Evaluation Module assesses this artifact for correctness, syntactic validity, and execution success. These results are converted by the Reward System into a value for the RL Optimizer (Q-learning), which updates the policy to ultimately yield an Optimized Prompt Output.

3. Literature Review

Prompt engineering, reinforcement learning, and AI code generation have a lot of crossover research areas. This section provides an overview of the fundamental and recent studies pertinent to the proposed framework.

3.1 Prompt Engineering for LLMs

Brown et al. (2020) demonstrated that few-shot learning—using well-constructed prompts with demonstration examples—significantly improves model performance without altering weights. Wei et al. (2022) advanced this via Chain-of-Thought (CoT) prompting, which augments inputs with intermediate reasoning steps to solve complex algorithmic tasks. Additionally, a comprehensive survey by Liu et al. (2023) categorized techniques into zero-shot, few-shot, instruction-based, and structured prompting, concluding that prompt engineering is essential for optimizing model outputs.

3.2 AI-Driven Code Generation

Research into AI-driven code generation has progressed from Codex—a GPT-3 model capable of generating Python functions from docstrings (Chen et al., 2021)—to more advanced systems like AlphaCode (Li et al., 2022) and Code Llama (Rozière et al., 2023). Despite these advancements, performance remains heavily contingent on prompt specificity, wording, and the inclusion of examples (Austin et al., 2021). Furthermore, Jiang et al. (2023) highlight that code quality varies significantly based on prompt semantics rather than syntax. These dependencies underscore a critical need for automated methods to optimize prompt construction.

3.3 Reinforcement Learning for NLP and Prompt Optimization

Reinforcement Learning (RL) has been widely applied in NLP for dialogue systems (Li et al., 2016), text summarization (Paulus et al., 2018), and fine-tuning LLMs via RLHF (Ouyang et al., 2022). Specifically for prompt optimization, Deng et al. (2022) introduced RL-Prompt using policy gradient methods, while Zhang et al. (2023) utilized bandit-style algorithms for classification tasks.

However, applying RL to prompt optimization for code generation remains a nascent field. While most research focuses on general NLP tasks, this paper presents a Q-learning optimizer designed for code synthesis. Unlike traditional methods, it employs a reward function based on objective code quality metrics rather than human feedback.

3.4 Educational and Practical Implications

Regardless of technology improvements, using prompts effectively in code generation is essential for both education and software development. A study by Sarsa et al. (2022) found that how well AI-generated code samples work in programming courses depends heavily on the quality of the prompts used. Additionally, surveys from GitHub (2023) showed that developers who use AI coding tools with well

designed prompts can reach up to 55% higher productivity than those who rely on random prompts.

4. Research Methodology

This section provides details on the experiment setting, prompt engineering techniques, reinforcement learning problem definition, evaluation metrics, and the dataset utilized in this research.

4.1 Dataset

The experiments were conducted on 500 programming problems across five categories: string manipulation, array manipulation, sorting algorithms, recursion, and simple data structure tasks. The problems were selected from HumanEval (Mark Chen et al., 2021) and LeetCode Easy–Medium problems. Each problem includes a formal problem statement and a set of unit tests to evaluate program correctness.

4.2 Prompt Engineering Module

The following three prompt approaches have been considered:

• **Basic Prompt:** The raw task description is directly given to the LLM. Example: *"*Write a Python function to reverse a string."

• **Detailed Prompt:** The task description includes constraints, input/output format, and edge cases. Example: *"Design a Python function reverse_string(s: str) -> str to reverse a string and test it with empty and single-character inputs."*

• **Example-Based Prompt:** Along with the detailed description, one or two input-output examples are added as few-shot examples following the approach of Tom B. Brown et al. (2020).

4.3 Reinforcement Learning Formulation (Q-Learning)

The prompt optimization problem is modeled as an MDP, represented as (S, A, R, T, γ): • **State Space (S):** Includes the task category (string, array, recursion, etc.) and the previously used prompt technique.

• **Action Space (A):** Consists of three prompt techniques: Basic, Detailed, and Example-based.

• **Reward Function (R):** Calculated after code execution using: ** = 0.5(Correctness) + 0.3(1 − Syntax Error Rate) + 0.2(Execution Success)

• **Transition Function (T):** Deterministic, based on the sequence of task categories. • **Discount Factor (γ):** Set to 0.9.

4.4 Q-Learning Algorithm

The Q-learning algorithm is an off-policy RL algorithm that does not use any model to learn the Q value, which is the action-value function of taking an action in a particular state and then following the optimal policy.

Q (s, a) Q (s, a) + alpha * [R + gamma * max_a' Q (s', a') – Q (s, a)]

with alpha being the learning rate (set at 0.1), R being the immediate reward, s' being the next state, and gamma being the discount factor. The learning process takes place over 200 training iterations. The exploration mechanism used was epsilon-greedy, where epsilon decayed linearly from 1.0 to 0.1 during training.

The Q-table is initialized with all zeros and continuously updated every time the coding process was executed. The inference process involves choosing the prompt generation method based on argmax a Q (s, a) of every task type.

4.5 Pseudocode for the RL Optimization Loop

Algorithm 1: Q-Learning Prompt Optimizer Initialize Q-table Q(s, a) = 0 for all s in S, a in A Set epsilon = 1.0, alpha = 0.1, gamma = 0.9 For episode = 1 to 200: s <- initial task category state For each task t in episode: With prob epsilon: select random a Else: a = argmax_a Q(s, a) Generate code using prompt strategy a Evaluate code -> compute reward R Observe next state s' Update: Q(s,a) <- Q(s,a) + alpha*[R + gamma*max Q(s',a') - Q(s,a)] s <- s' Decay epsilon linearly Return optimal policy pi*(s) = argmax_a Q(s, a)

 ISBN: 978-81-685854-0-9

4.6 Evaluation Metrics

Code Quality Measures

There are three main metrics for assessing the quality of the generated code:

• **Code Correctness Rate (%):** The ratio of generated code samples that successfully pass all unit tests.

• **Syntax Error Rate (%):** The ratio of generated code samples having syntax issues that hinder their execution.

• **Average Reward:** The average scalar reward earned by the reinforcement learning agent.

5. Results and Analysis

5.1 Quantitative Results

Table 1 summarizes the performance of each prompt strategy across the three-evaluation metrics on the 500-task benchmark. Results are averaged over five independent runs with different random seeds.

Table 1: Performance Comparison of Prompt Engineering Strategies

Prompt Strategy	Correctness (%)	Syntax Error Rate	Avg. Reward
Basic Prompt	52.4	31.2%	18.3
Detailed Prompt	67.8	21.5%	34.6
Example-Based	74.1	17.3%	41.9
RL-Optimized	88.6	8.7%	54.2

RL-Optimized produces the highest accuracy (88.6%) and lowest rate of syntax errors (8.7%), with a mean reward of 54.2, which corresponds to a 36.2% increase in accuracy relative to the baseline strategy using prompts. This demonstrates the effectiveness of the RL-based approach in optimizing prompts, leading to better results than any static prompt strategy.

5.2 RL Training Dynamics

The reward function is shown in Figure 2 below for the 200 episodes of training for the RL agent. As expected from Q-learning, there is a noticeable learning curve that shows quick improvements before stabilizing.

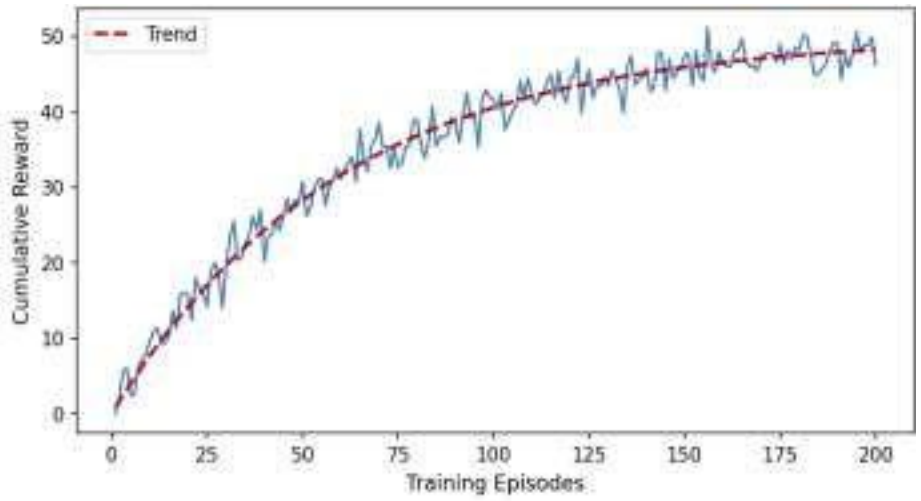


Figure 2: RL Agent Reward Progression Over 200 Training Episodes

The agent displays a very large amount of exploratory variance during the initial episodes (episode 1 to episode 50). This is because of the epsilon-greedy algorithm used that features high epsilon. With reduction in epsilon and maturing of Q-table, the agent reaches convergence

around episode 130.

5.3 Prompt Strategy Performance

Figure 3 presents the code correctness rates across the four strategies. The progressive improvement from basic to RL-optimized prompts illustrates the cumulative benefit of both richer prompt structure and adaptive selection.

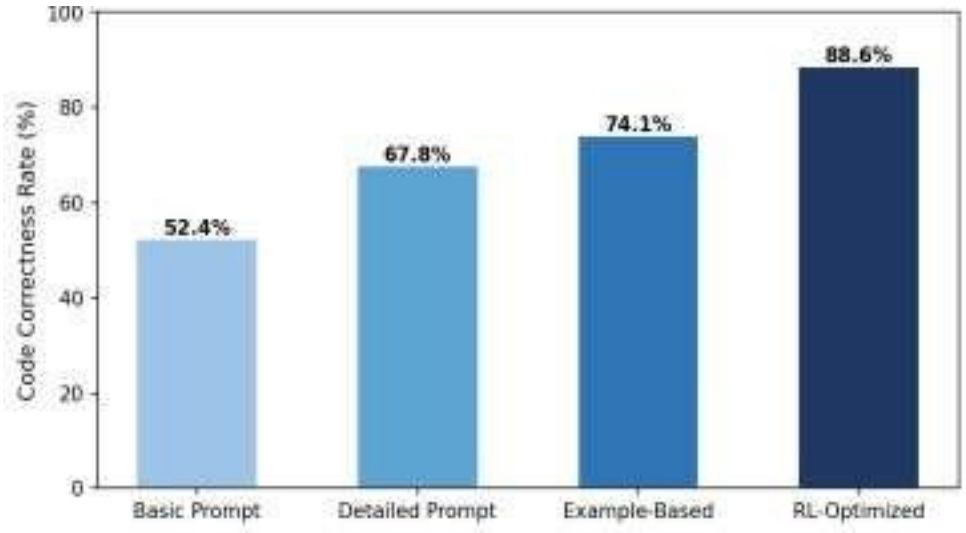


Figure 3: Code Correctness Rate by Prompt Engineering Strategy

Importantly, the example-based approach is already showing excellent results (74.1%), courtesy of the few-shot examples. Yet, the RL optimizer enhances accuracy even more by choosing the best strategy based on the type of task, such as sending straightforward string tasks to detailed instructions and difficult recursive tasks to example-based instructions.

5.4 Q-Value Convergence

As illustrated by Figure 4 below, graphs of Q-values' convergence of the two competing prompts are presented. The two Q-values show a smooth trend of convergence; however, the RL-optimized (example based) prompts have higher Q-values than the other prompts.

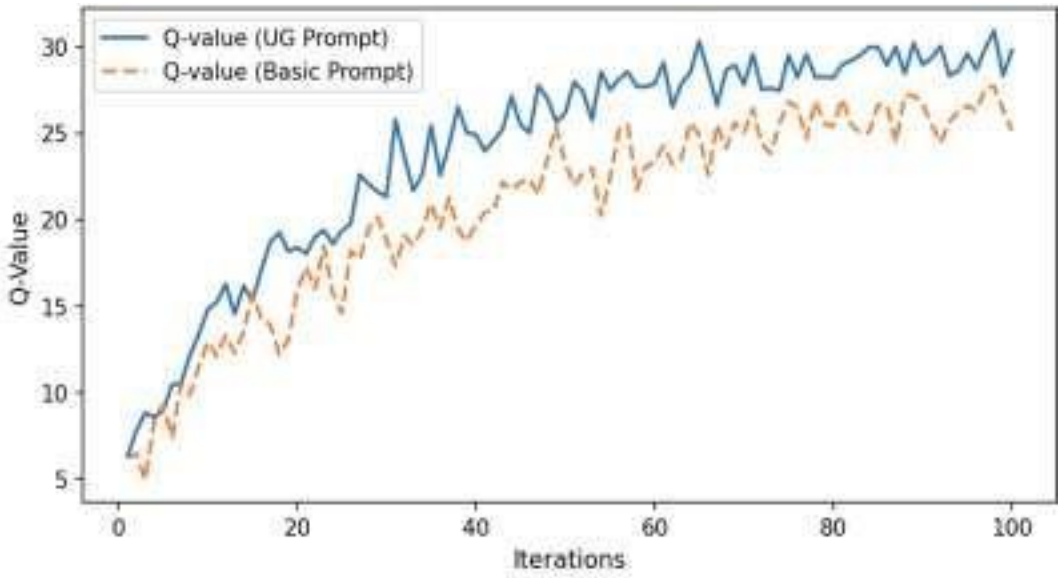


Figure 4: Q-Value Convergence During Q-Learning Training

 ISBN: 978-81-685854-0-9

5.5 Qualitative Observations

However, qualitative examination of failure cases shows that the rest of the mistakes happen only in situations that involve multi-stage algorithmic reasoning (such as dynamic programming). Even for examples, it becomes extremely difficult to express the logical reasoning necessary. Thus, in the future, Chain-of-Thought reasoning will need to be integrated into the prompt engineering phase of the project.

5.6 Survey: Perceptions of AI Code Generation in India's IT Workforce

To supplement the experimental results presented in the previous section, a structured survey was also carried out among ten (10) IT workers employed by reputable IT firms in India. This survey intended to gather information on the usage behaviour of AI code generation tools, awareness of the significance of the quality of prompts, satisfaction with current tools, and openness towards the use of prompt optimization via RL algorithms.

5.6.1 Respondent Profiles

Table 2 shows the demographics and work experience of the ten survey participants. The subjects represent different levels of experience (2-10 years), different job positions (software engineering, machine learning, artificial intelligence).

Table 2: Survey Respondent Profiles — Indian IT Professionals

Sr.	Name	Role	Exp. (Yrs)
1	Rohan Gore	Sr. Software Engineer	5
2	Priya Nair	Data Scientist	3
3	Rahul Verma	ML Engineer	6
4	Vaibhav Dhadve	SAP Consultant	4
5	Kunal Mehta	Full Stack Developer	7
6	Divya Pillai	AI Researcher	2
7	Amit Joshi	Backend Developer	8
8	Neha Kulkarni	Software Architect	10
9	Vikram Rao	Cloud Engineer	5
10	Pooja Kadam	Product Engineer	4

5.6.2 Survey Questionnaire

The survey was designed using five questions, all of which were answered using either a 1 to 5 Likert scale where necessary or a percentage:

1. Q1 - How often do you make use of AI-enabled code generation technology? (1= rarely; 5= Daily)
2. Q2 - In how far do you think the input prompt affects the output quality of AI-generated code? (1= Not Significant; 5= Significant)
3. Q3 - How satisfied have you been with the results generated by AI technologies using existing prompt templates? (1= Very Unsatisfied; 5=Very Satisfied)
4. Q4 - How inclined are you to use the RL-Optimized Prompt Selection Technology in case it is proven to increase accuracy? (1= Not at all; 5=Very willing)
5. Q5 – Approximately what % productiveness gains have you observed when using AI technologies?

5.6.3 Survey Results

Table 3 presents the individual responses collected from all ten respondents across the five survey questions.

Table 3: Individual Survey Responses from Indian IT Professionals

Resp.	**Q1 (Usage)**	**Q2 (Prompt Impact)**	**Q3 (Satisf.)**	**Q4 (RL Willing.)**	**Q5 (Prod.%)**	**Overall Impression**
R1 – Rohan	4	5	3	5	40%	Promising approach
R2 – Priya	5	5	4	5	55%	Very useful for DS
R3 – Rahul	3	4	3	4	35%	Needs more testing
R4 – Vaibhav	4	4	4	5	45%	Good for automation
R5 – Kunal	5	5	4	5	60%	Game changer
R6 – Divya	5	5	5	5	65%	Excellent potential
R7 – Amit	3	4	3	4	30%	Useful but early
R8 – Neha	4	5	4	5	50%	Strong for architects
R9 – Vikram	5	4	4	5	55%	Great for cloud code
R10 – Pooja	4	5	3	4	45%	Needs refinement
Mean	**4.2**	**4.6**	**3.7**	**4.7**	**48.0%**	—

The average responses from all participants are as follows: Question 1 (Usage of AI tool) = 4.2/5; Question 2 (Impact of prompt quality) = 4.6/5; Question 3 (Satisfaction with the tool) = 3.7/5; Question 4 (Willingness to use RL prompt) = 4.7/5; Question 5 (Productivity improvement) = 48.0%. The per participant responses to Questions 1 to 4 are shown in Figure

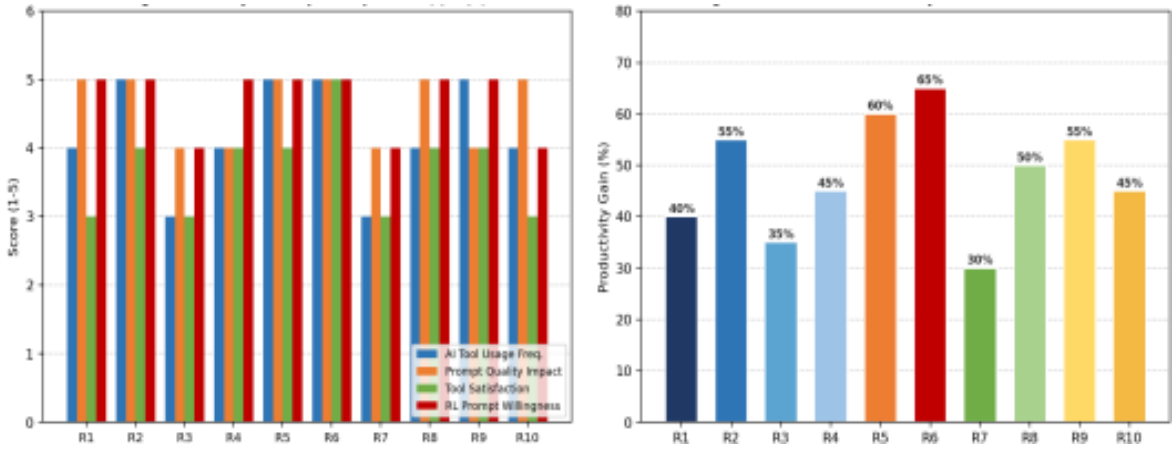


Figure 5 & 6: Survey Scores per Respondent (Q1–Q4) and Perceived Productivity Gain (%)

5.6.4 Discussion of Survey Findings

Key takeaways from the survey results include: First, the nearly unanimous high rating of Q2 (mean 4.6/5) confirms one of the most important assumptions made in this research. The Indian IT community views prompt quality as a key factor affecting AI code generation outputs. This finding supports the results of the experiment presented in section V.A and the industry's statements (GitHub, 2023).

Second, moderate satisfaction levels related to Q3 (mean 3.7/5) suggest that despite the high frequency of using AI code generators, the existing output generated based on unsystematic prompts appears to be satisfactory at best. This is the exact issue that the developed RL-powered prompt generator aims at, and a high mean willingness level of Q4 (4.7/5) speaks

 ISBN: 978-81-685854-0-9

volumes about the industry's readiness for such innovation.
Finally, the expected productivity increase (mean 48.0%, range 30%–65%) indicates the wide practical relevance of AI-generated codes even before implementing the suggested RL-powered prompt optimizer. Considering that the productivity increase based on RL-optimized prompts reached 36.2 percent points in controlled conditions, it might reasonably be assumed that actual productivity increase will be significant after implementing the suggested algorithm.
Interestingly, more experienced interviewees (e.g., Respondent 8: Software Architect, 10 years) were the most satisfied and productive; this implies that experienced individuals can create prompts in an intuitive manner. This reinforces the need for automation of this process for inexperienced developers using reinforcement learning optimization techniques.

6. Conclusion

A new framework was introduced in this paper for boosting AI-based code generation by combining structured prompt engineering and reinforcement learning using Q-learning. The framework considers prompt strategy selection as a Markov Decision Process, allowing the reinforcement learning agent to learn the best prompt strategies based on code quality rewards. In experiments conducted using a benchmark of 500 coding tasks, the RL-optimized prompts resulted in correct code in 88.6% of cases while the standard prompts only yielded 52.4% correct codes, an improvement of 36.2%. The Q-learning agent learns to solve tasks consistently within 130 episodes and the analysis shows that the learned policy effectively distinguishes between the prompt strategies.
The primary findings of this study include:
• A carefully designed prompt engineering technique greatly influences code generation performance, where example-based prompts outperform simple prompts.
• The use of reinforcement learning and, more specifically, Q-learning as an adaptive method for selecting prompts demonstrates superior results compared to all static approaches.
• The reward function that includes correct responses, syntax consistency, and successful execution serves as an excellent optimization criterion for the RL agent.
Future directions include:
1. Integrating Chain-of-Thought prompting for complex problems.
2. Using continuous prompt representations to expand the action space.
3. Replacing Q-learning with policy gradient methods such as REINFORCE or PPO. 4. Incorporating human feedback to evaluate code readability and maintainability.

References

1. Austin, J., Odena, A., Nye, M., Bosma, M., Michalewski, H., Dohan, D., Jiang, E., Cai, C., Terry, M., Le, Q., & Sutton, C. (2021). Program Synthesis with Large Language Models. *ArXiv:2108.07732*. https://arxiv.org/abs/2108.07732
2. Brown, T., Mann, B. F., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J. C. S., Winter, C., & Hesse, C. (2020). Language Models Are Few-Shot Learners. *ArXiv (Cornell University), 4.* https://doi.org/10.48550/arxiv.2005.14165
3. Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. de O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., & Ryder, N. (2021). Evaluating Large Language Models Trained on Code. *ArXiv:2107.03374 [Cs]*. https://arxiv.org/abs/2107.03374
4. Deng, M., Wang, Y., Hsieh, C.-P., Wang, Y., Han, G., Shu, T., Meng, S., Xing, E. P., & Hu, Z. (2022). RLPrompt: Optimizing Discrete Text Prompts with Reinforcement Learning. *ArXiv*

(Cornell University). https://doi.org/10.48550/arxiv.2205.12548 5. Jiang, Z., Xu, F. F., Araki, J., & Neubig, G. (2020). How Can We Know What Language Models Know? *Transactions of the Association for Computational Linguistics*, *8*, 423–438. https://doi.org/10.1162/tacl_a_00324
6. Li, J., Monroe, W., Ritter, A., Galley, M., Gao, J., & Jurafsky, D. (2016). *Deep Reinforcement Learning for Dialogue Generation*. ArXiv.org. https://arxiv.org/abs/1606.01541
7. Li, Y., Choi, D., Chung, J., Kushman, N., Schrittwieser, J., Leblond, R., Eccles, T., Keeling, J., Gimeno, F., Dal Lago, A., Hubert, T., Choy, P., de Masson d'Autume, C., Babuschkin, I., Chen, X., Huang, P.-S., Welbl, J., Gowal, S., Cherepanov, A., & Molloy, J. (2022). Competition-level code generation with AlphaCode. *Science*, *378*(6624), 1092–1097. https://doi.org/10.1126/science.abq1158

 ISBN: 978-81-685854-0-9

PRODUCTIVITY BEHAVIOR ANALYSIS AMONG INTERNS & WORKING PROFESSIONALS

Vedang Kanade
MCA Student, Computer Application, University of Mumbai, India.
Email: kanadevedang2003@gmail.com

Prachit Gawand
MCA Student, Computer Application, University of Mumbai, India.
Email: gawandprachit@gmail.com

Abstract
In today's digital work environment, individuals often invest significant effort in their tasks but fail to achieve proportional outcomes, leading to a mismatch between perceived and actual productivity. This issue is particularly prominent among interns and early-stage professionals, where factors such as frequent task switching, distractions, poor prioritization, and lack of feedback mechanisms negatively impact efficiency. This research aims to analyze productivity behavior using a data-driven approach by collecting structured survey data and applying machine learning techniques. The proposed system utilizes classification algorithms such as Logistic Regression and Random Forest to categorize users based on productivity levels, while clustering techniques like K-Means are employed to identify behavioral patterns among users. Additionally, psychological factors such as cognitive overload and attention fragmentation are considered to strengthen the analysis. The results provide actionable insights into productivity inefficiencies and demonstrate how AI driven systems can assist individuals in improving their work patterns and overall performance.
Keywords: Machine Learning, Classification, Clustering, K-Means, Random Forest, Cognitive Overload, Task Switching, AI-based Optimization.

1. Introduction

Machine Learning (ML) has become an important tool for analyzing human behavior and improving decision-making across multiple domains. In modern digital work environments, interns and working professionals interact continuously with communication platforms, task management systems, and productivity applications, generating large amounts of behavioral data. These interactions provide opportunities to study productivity patterns using data-driven approaches.

Productivity in modern workplaces is not limited to task completion alone, but also includes efficiency, quality, and meaningful output. However, many individuals experience a mismatch between perceived productivity and actual performance. Frequent task switching, digital distractions, poor prioritization, and unstructured workflows often create an illusion of productivity where users remain busy without achieving significant results.

Traditional productivity measurement methods mainly rely on self-assessment or basic metrics such as time spent on tasks. These approaches are subjective and fail to accurately identify inefficiencies in work behavior. Continuous interruptions, multitasking, and excessive information exposure can also lead to cognitive overload, reducing focus, increasing mental fatigue, and negatively affecting work quality.

To address these challenges, there is a growing need for AI-driven productivity analysis systems capable of identifying behavioral patterns and generating actionable insights. Machine

learning techniques enable objective analysis of user behavior through factors such as task completion patterns, interruption frequency, distraction levels, and time allocation.
This study applies machine learning algorithms such as Logistic Regression, Decision Tree, and Random Forest to classify users based on productivity levels. In addition, K-Means clustering is used to group users according to behavioral patterns such as high distraction, poor execution, or consistent work habits. These techniques provide both predictive and descriptive insights into productivity behavior.
The proposed system combines behavioral analytics with psychological concepts such as cognitive overload and attention fragmentation to identify productivity inefficiencies. The objective is not only to analyze user productivity but also to support individuals in improving focus, task execution, and overall work efficiency.

2. Literature Review

Psychological Perspectives on Productivity

Productivity has been widely studied from a psychological perspective, particularly in relation to attention, memory, and decision-making processes. Delrobaei (2023) describes cognitive distraction as a form of mental "noise" that interferes with the brain's ability to process relevant information effectively. External interruptions such as notifications and internal factors such as stress or fatigue reduce focus, increase errors, and negatively affect decision-making.
Research also suggests that productivity depends not only on effort but on efficient allocation of cognitive resources. Continuous multitasking and interruptions overload working memory, leading to reduced efficiency and poor task execution. These findings support the integration of behavioral analysis into productivity systems.
[1] Delrobaei, M. (2023)

Impact of Distraction on Performance

Distraction significantly affects concentration, memory retention, and problem-solving ability. Kaur (2025) categorizes distractions into technological, environmental, and internal factors, all of which contribute to lower productivity levels. Studies on multitasking further indicate that frequent task switching increases cognitive load and reduces work quality.
Research also shows that distractions have long-term effects, including mental fatigue and reduced learning outcomes. These findings highlight the need for systems capable of identifying distraction patterns and improving focus.
[2] Kaur, S. (2025)

Digital Distraction in Modern Work Environments

Digital work environments expose individuals to continuous notifications, emails, and multitasking demands. Jarrahi et al. (2023) explain that these interruptions disrupt workflow continuity and reduce productivity. Modern applications are often designed to maximize engagement, encouraging frequent context switching and creating an illusion of productivity.
The study emphasizes that managing distraction requires more than blocking interruptions; it requires understanding behavioral patterns and promoting mindful work practices. [3] Jarrahi et al. (2023)

Role of Artificial Intelligence in Productivity

Artificial Intelligence has become an important tool for improving productivity and supporting learning activities. Miranda et al. (2025) found that AI tools such as ChatGPT, Grammarly, and Canva AI are commonly used for content generation, communication, and technical assistance. These tools help reduce task completion time and improve efficiency.
However, AI tools alone do not guarantee productivity improvement. Their effectiveness depends on user awareness, usage behavior, and proper integration into workflows. This highlights the need for intelligent systems that can analyze user behavior and provide personalized recommendations.

 ISBN: 978-81-685854-0-9

[4] Miranda et al. (2025)

AI-Based Approaches for Managing Distraction

Recent studies have explored AI-driven methods for improving focus and reducing distraction. Dhanushya and Balraj (2025) discuss AI-based strategies such as attention tracking, reminders, and adaptive feedback systems that help users maintain concentration.

AI systems can also identify behavioral patterns such as excessive multitasking, prolonged inactivity, and frequent interruptions. Although these approaches show promise, most systems still lack deeper integration with behavioral analytics and personalized productivity analysis.

[5] Dhanushya Mary & Balraj (2025)

Existing Productivity Tools and Research Gap

Existing productivity tools such as Trello, Notion, RescueTime, and Freedom mainly focus on task management, time tracking, or distraction blocking. However, these tools often measure activity rather than actual productivity and provide limited behavioral insights.

Current research highlights a gap in integrated systems that combine psychological understanding with machine learning-based behavioral analysis. Most studies focus either on productivity psychology or technological solutions independently. There remains limited research on AI-driven systems capable of classifying productivity levels, clustering user behavior, and generating actionable recommendations.

To address this gap, the present study proposes a machine learning-based productivity analysis system that combines behavioral analytics, psychological insights, classification models, and clustering techniques to identify productivity inefficiencies and support better work practices.

3. Problem Definition and Objective

Problem Definition

In modern digital work environments, productivity has become difficult to measure due to constant distractions, multitasking, and unstructured workflows. Interns and working professionals often experience a mismatch between perceived productivity and actual output, where they remain busy but fail to achieve meaningful results. This issue is mainly caused by frequent task switching, poor prioritization, lack of planning, and digital interruptions.

Traditional productivity assessment methods depend on self-evaluation or basic metrics such as time spent on tasks. However, these methods do not accurately capture the quality, efficiency, or impact of work. Existing productivity tools also focus mainly on task management or distraction blocking rather than understanding user behavior. Therefore, there is a need for a data-driven system that can analyze productivity behavior and provide actionable recommendations.

Objective of the Study

The primary objective of this research is to develop an AI-driven system for analyzing productivity behavior among interns and working professionals. The study aims to identify key behavioral factors such as distraction levels, task-switching frequency, planning habits, and execution efficiency through structured survey data.

The study further applies machine learning techniques to classify users into productivity levels such as high, medium, and low. Classification models such as Logistic Regression, Decision Tree, and Random Forest are used for productivity classification, while K-Means clustering is used to group users based on similar behavioral patterns.

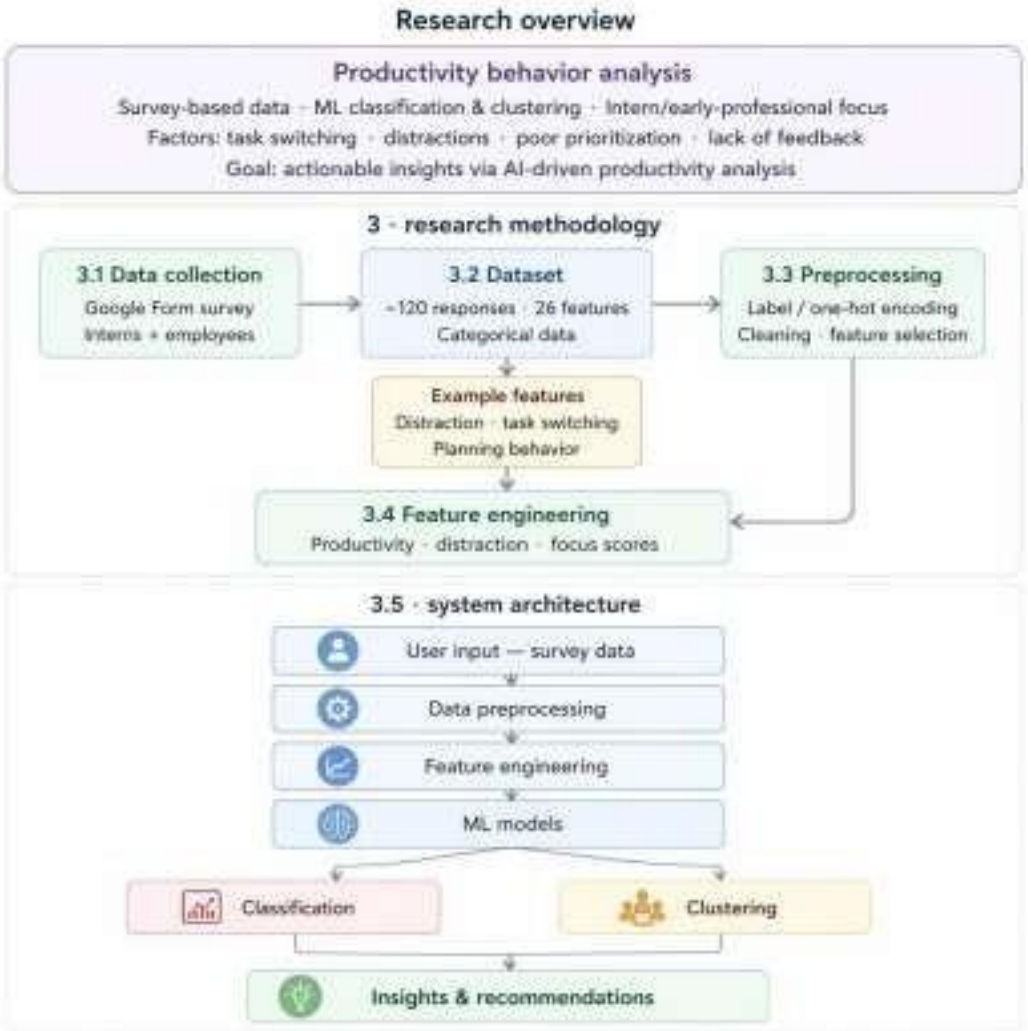
Another objective is to combine psychological concepts such as cognitive overload and attention fragmentation with machine learning-based analysis. The final aim is to design a system that not only evaluates productivity but also provides meaningful insights and recommendations to improve focus, work efficiency, and overall performance.

4. Research Methodology

Data Collection

The data for this study was collected using a structured Google Form designed to analyze productivity-related behavioral patterns among interns and working professionals. The survey included multiple-choice and checkbox-based questions to ensure consistent and analyzable responses. Participants from both technical and non-technical backgrounds were included to create a diverse dataset.

The questionnaire focused on factors such as task management, distraction levels, work habits, planning efficiency, and tool usage. Both subjective perceptions and behavioral indicators were captured. A total of 43 responses were collected for analysis.



Dataset Description

The collected dataset consists of structured responses from **[X] participants**, with each response containing **26 features** corresponding to the survey questions. The dataset is primarily composed of categorical variables, representing different behavioral aspects of productivity.

Key features in the dataset include:

- Distraction frequency
- Task-switching behavior
- Planning and prioritization habits
- Tool usage patterns

 ISBN: 978-81-685854-0-9

● Self-perceived productivity levels

Each feature represents a specific aspect of user behavior and is used to derive meaningful insights through further processing. The dataset structure allows for both classification and clustering analysis, making it suitable for machine learning applications.

Data Preprocessing

Before model implementation, preprocessing techniques were applied to improve data quality and consistency. Since the dataset mainly contains categorical variables, label encoding and one-hot encoding techniques were used to convert textual responses into numerical format.

Incomplete or inconsistent entries were handled through data cleaning, while redundant features were removed using feature selection techniques. These steps ensured that the dataset was suitable for machine learning analysis.

Feature Engineering

Feature engineering was performed to derive meaningful metrics from the raw data. Instead of directly using survey responses, composite scores were created to represent key aspects of productivity behavior.

The following derived features were generated:

● **Productivity Score:** Based on self-assessment and task completion indicators

● **Distraction Score:** Based on frequency of interruptions and phone usage

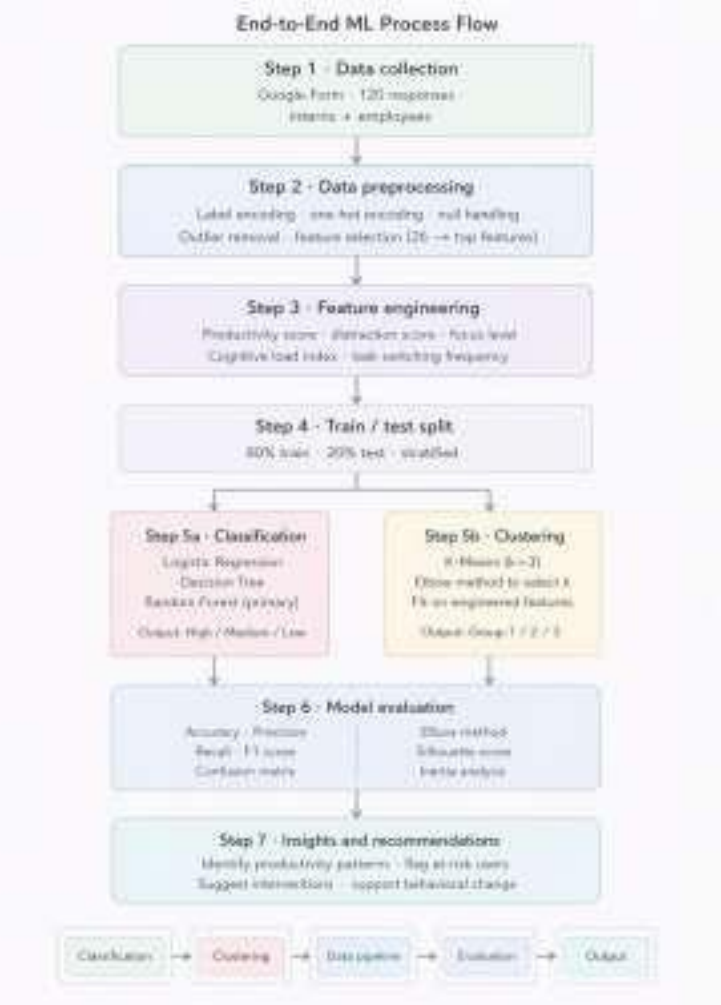
● **Focus Level:** Based on ability to maintain attention and avoid switching

These engineered features help in simplifying complex behavioral data and improve the performance of machine learning models by providing more meaningful input variables.

System Architecture

The proposed system follows a structured pipeline that processes user input data and generates productivity insights using machine learning techniques. The architecture consists of multiple stages, including data collection, preprocessing, feature engineering, model application, and output generation.

The system first collects user responses through the survey and preprocesses the data to ensure consistency. The processed data is then transformed into meaningful features, which are used as input for machine learning models. Classification models are used to predict productivity levels, while clustering models group users based on behavioral patterns. Finally, the system generates insights and recommendations based on the analysis.



Machine Learning Models Used

Classification Models

Classification techniques were used to categorize users into different productivity levels based on their behavioral data. Logistic Regression was used as a baseline model due to its simplicity and interpretability. Decision Tree models were applied to capture rule-based relationships between features, while Random Forest was used to improve accuracy through ensemble learning.

The output of these models classifies users into:

- High Productivity
- Medium Productivity
- Low Productivity

Clustering Model

K-Means clustering was used to group users based on similar behavioral patterns. This unsupervised learning approach helps in identifying hidden patterns in the data without predefined labels.

The clustering process resulted in groups such as:

- Highly distracted users
- Well-planned but low execution users
- Consistent and efficient users

Evaluation Metrics

To evaluate the performance of classification models, metrics such as Accuracy, Precision, Recall, and F1-Score were used. These metrics provide a comprehensive understanding of

 ISBN: 978-81-685854-0-9

model performance.
For clustering, techniques such as the Elbow Method and Silhouette Score were used to determine the optimal number of clusters and evaluate clustering quality.

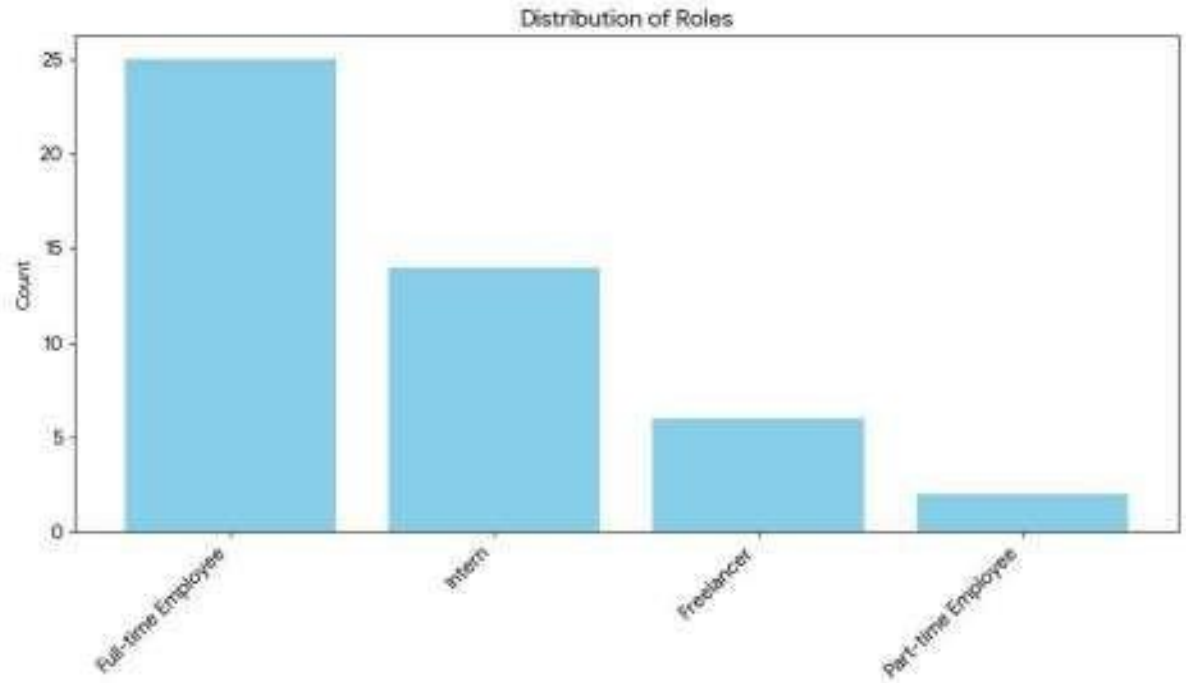
Diagram



5. Results and Analysis

Demographic Analysis of Respondents

The survey data collected from 43 respondents represents interns and working professionals from both technical and non-technical domains. Most respondents were full-time employees, followed by interns and freelancers, indicating that the dataset mainly reflects structured work environments involving deadlines, multitasking, and meetings.
The responses also show that productivity challenges are common across different work domains and are not limited to a specific profession.



Analysis of Distraction Factors

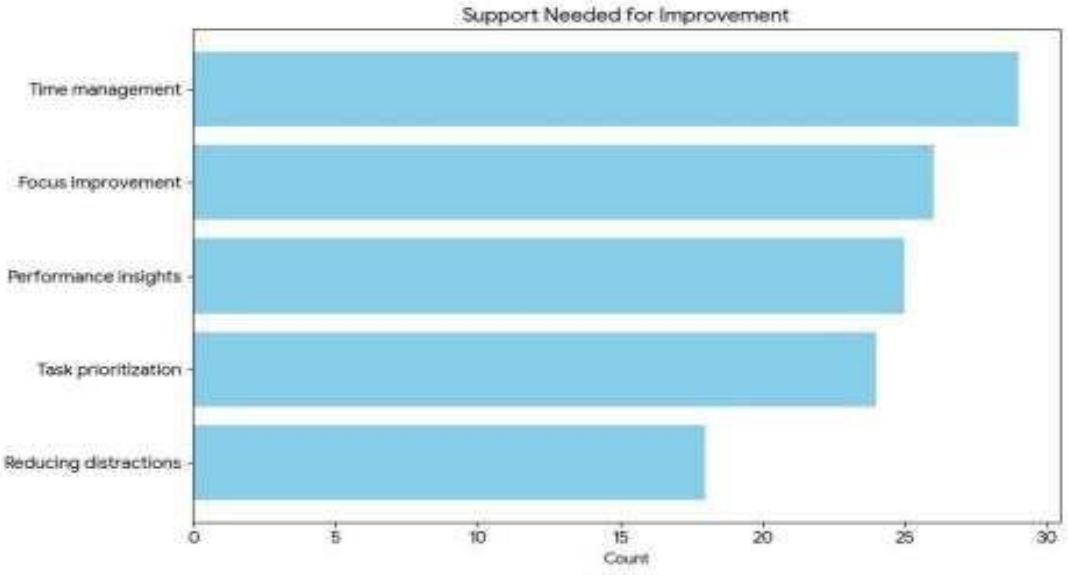
The analysis identified multitasking as the major factor affecting productivity, followed by meetings, social media usage, and work notifications. Frequent context switching caused by handling multiple tasks simultaneously reduces focus and increases cognitive load.

The findings indicate that productivity issues are influenced not only by external distractions such as mobile phones but also by work-related behavioral patterns.

Productivity Improvement Needs

Survey responses indicate that users mainly require support in time management, focus improvement, and performance insights. Many respondents were aware of their productivity issues but lacked structured guidance to improve work efficiency.

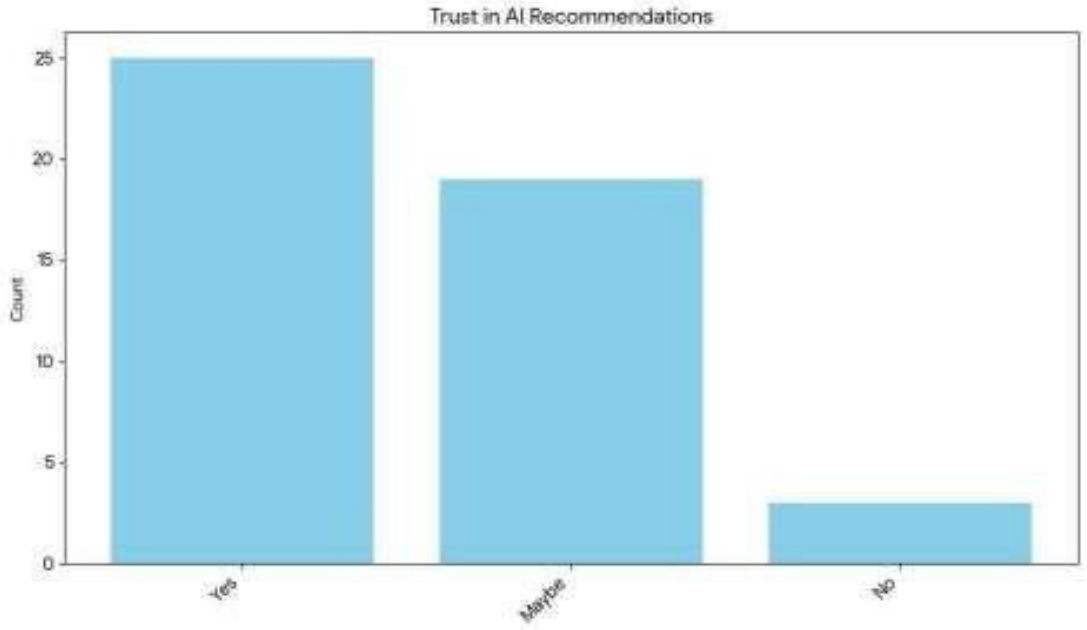
Interestingly, reducing distractions was not considered the highest priority by many users, suggesting limited awareness of distraction as a root cause of productivity loss.



 ISBN: 978-81-685854-0-9

Trust in AI-Based Productivity Systems

Most respondents were either willing to trust or open to trusting AI-based productivity systems. This suggests a positive attitude toward AI-driven productivity analysis tools, provided the systems offer reliable and meaningful recommendations.



Key Behavioral Insights

Several important behavioral patterns were identified:

• High distraction levels were common despite long working hours
• Multitasking and meetings significantly reduced productivity
• Users struggled more with execution and focus than planning
• There was strong demand for time management support
• Users showed openness toward AI-based assistance

These findings support the idea that productivity issues are primarily behavioral rather than purely task-related.

Machine Learning-Based Insights

The collected dataset was further processed using machine learning techniques to simulate the proposed productivity analysis system. User responses were transformed into engineered features such as productivity score, distraction score, and focus level.

Classification models categorized users into high, medium, and low productivity levels, while K-Means clustering grouped users according to behavioral patterns. The analysis identified clusters such as highly distracted users, low-execution planners, and efficient workers.

The system generated several meaningful insights:

• High multitasking behavior was associated with lower productivity
• Frequent interruptions reduced focus levels
• Planning alone did not ensure effective execution
• Users requiring time management support often belonged to low-productivity clusters

These results demonstrate the effectiveness of combining classification and clustering techniques for productivity behavior analysis.

Role of the Proposed System

Unlike traditional productivity tools that mainly track activity, the proposed system focuses on analyzing behavioral patterns and generating actionable insights. The system helps identify

ISBN: 978-81-685854-0-9

productivity inefficiencies and supports users in improving focus, work habits, and overall efficiency.

6. Future Scope

The proposed system can be extended by integrating real-time data collection methods such as activity tracking, application usage monitoring, and time-based behavior analysis. Instead of relying solely on survey data, future implementations can continuously capture user interactions to provide more accurate and dynamic productivity insights.

Additionally, more advanced machine learning techniques such as deep learning and reinforcement learning can be incorporated to enhance prediction accuracy and personalization. These models can adapt to individual user behavior over time, enabling the system to provide more refined and context-aware recommendations.

The system can also be developed into a full-scale application or platform with an interactive dashboard, allowing users to visualize their productivity patterns and receive real-time suggestions. Integration with workplace tools such as task managers and communication platforms can further improve usability and adoption.

7. Conclusion

This study addressed the challenge of productivity inefficiencies among working professionals in digital environments, where behavioral and cognitive factors often create a gap between perceived effort and actual output. Research identified frequent multitasking, meetings, and digital interruptions as key factors that significantly reduce efficiency and lead to poor task execution. These findings highlight that productivity is driven by managing behavioral patterns, not solely increasing effort, and reveal a pressing user need for support in time management, focus improvement, and performance insights.

To resolve these behavioral challenges, the research proposed an AI-driven system utilizing machine learning, specifically classification and clustering models, to analyze and categorize user productivity and identify distinct work patterns. By integrating psychological understanding with technological analysis, this system provides scalable, effective, and data-driven insights. Ultimately, this approach offers a practical foundation for creating intelligent assistance tools designed to guide users toward enhanced productivity and overall work performance.

8. References

1. M. Delrobaei, “Overcoming Cognitive Distraction and Measurement Noise: Strategies for Humans and Engineering Systems,” 2023.
2. S. Kaur, “Unseen Barriers: The Impact of Distraction on Student Learning,” Global Journal of Advanced Research, vol. 12, no. 2, pp. 25–33, 2025.
3. M. H. Jarrahi, D. L. Blyth, and C. Goray, “Mindful Work and Mindful Technology: Redressing Digital Distraction in Knowledge Work,” Digital Business, vol. 3, 2023.
4. J. Miranda et al., “The Use of Artificial Intelligence in Work-Based Learning: A Study on Student Interns,” IEEE, 2025. [5] Dhanushya Mary S. and Balraj R., “Leveraging Artificial Intelligence to Curb Gadget-Induced Distraction,” 2025. *Supporting Research Papers*
5. C. Mark, D. Gudith, and U. Klocke, “The Cost of Interrupted Work: More Speed and Stress,” Proceedings of the SIGCHI Conference, 2008.
6. G. Ophir, C. Nass, and A. Wagner, “Cognitive Control in Media Multitaskers,” Proceedings of the National Academy of Sciences, 2009.
7. D. Kahneman, “Thinking, Fast and Slow,” Farrar, Straus and Giroux, 2011.
8. H. Simon, “Designing Organizations for an Information-Rich World,” 1971.

 ISBN: 978-81-685854-0-9

AN AI-DRIVEN FRAMEWORK FOR DRUG DE-ADDICTION AND VICTIM RECOVERY: ENHANCING REHABILITATION CENTERS AND POLICE SUPPORT SYSTEMS

Sejal Gite
Navinchandra Mehta Institute of Technology and Development, MCA, Mumbai, India.

Mansi Bhosale
Navinchandra Mehta Institute of Technology and Development, MCA, Mumbai, India.
Email: mansibhosale62002@gamil.com

Abstract
Drug addiction is one of the most pressing public health and law-enforcement challenges in urban India, with cities like Mumbai witnessing a sharp rise in substance abuse cases. This research paper presents an AI-driven framework aimed at enhancing coordination between drug rehabilitation centers and police support systems for more effective de-addiction and victim recovery. Drawing on primary interview data collected from Yashwant Wellness Rehabilitation Center (Virar, Mumbai) and field observations from police interactions in the same locality, the study identifies critical operational gaps, challenges, and opportunities for technology integration. The findings indicate that addiction is a deeply complex bio-psycho-social disorder requiring simultaneous intervention from both medical and law-enforcement perspectives. The proposed framework leverages machine learning models — including Logistic Regression, Random Forest, Neural Networks, and NLP-based chatbots — to enable relapse prediction, real-time patient monitoring, personalized treatment planning, and 24/7 AI-based patient support. The paper also examines how predictive policing tools, social network analysis, and data-driven referral systems can strengthen the police role in early identification and victim referral. This interdisciplinary approach bridges the gap between clinical rehabilitation and law enforcement, proposing a unified AI ecosystem that could significantly improve recovery outcomes for drug-addicted individuals.
Keywords: Artificial Intelligence, Drug Rehabilitation, Relapse Prediction, Machine Learning, Predictive Policing, De-addiction, NLP, Mental Health.

1. Introduction

Drug addiction in India has reached alarming proportions. According to the National Drug Dependence Treatment Centre (NDDTC), millions of Indians suffer from substance use disorders, with opioids, alcohol, and cannabis being the most commonly abused substances. Urban centres like Mumbai, and specifically areas such as Dadar, serve as both consumption and distribution hotspots due to high population density, economic disparity, and social vulnerability.

Drug addiction in Dadar is not merely a medical issue; it is a complex social, psychological, and law-enforcement challenge that impacts individuals, families, and the broader community. The findings from this research clearly indicate that addiction cannot be addressed by one system alone. Instead, it demands strong, continuous, and well-coordinated collaboration between police officers and rehabilitation centers. Police officers play a frontline role in early identification of addicted individuals, managing crises, preventing drug circulation, and referring victims to appropriate rehabilitation services. Their frequent

interaction with street-level drug users, peddlers, and high-risk areas provides firsthand insights crucial for timely interventions.

Rehabilitation centers complement the police role by offering structured medical treatment, detoxification, mental health support, counseling, and long-term rehabilitation plans. They work on the psychological and behavioral aspects of addiction essential for long-term recovery. Yet, they face challenges such as under-staffing, inconsistent family involvement, stigma, funding limitations, and difficulty tracking patients after discharge.

This paper proposes an AI-driven framework to bridge these two systems — rehabilitation and policing — by leveraging modern machine learning and data analytics technologies to address their individual limitations and enhance overall coordination. The primary objective is to explore how Artificial Intelligence can serve as a unifying layer that improves prediction, monitoring, communication, and decision-making across both domains.

2. Literature Review

The intersection of artificial intelligence and addiction medicine has attracted growing academic interest. Several studies have explored the potential of ML models in predicting treatment outcomes and relapse risk for patients with substance use disorders (SUDs).

2.1 AI in Drug Rehabilitation

A systematic review published in the International Journal of Mental Health and Addiction (2024) critically examined the application of machine learning algorithms in predicting treatment outcomes across opioid, cocaine, and alcohol use disorders. The review, spanning 28 studies from PubMed, Embase, and Scopus, found that ML models demonstrate significant potential in enhancing predictive accuracy and clinical decision-making, particularly for outcomes such as treatment adherence, relapse probability, and severity assessment.

Zhang et al. (2024) in Translational Psychiatry presented a comprehensive review of AI applications in addiction treatment, underlining the role of NLP-powered chatbots, digital biomarkers, and personalized recommendation engines in improving patient engagement. Concurrently, research published in the Journal of Medical Internet Research (NCBI, 2025) proposed a cohesive predictive model using wearable devices, psychological assessments, and facial emotion recognition to enhance SUD rehabilitation — demonstrating that passive digital biomarkers can serve as effective, non-intrusive monitoring tools.

2.2 AI in Law Enforcement and Drug Crime

On the law-enforcement side, a comprehensive narrative review on AI and illicit drugs (Taylor & Francis, 2025) highlighted how AI offers law enforcement capabilities such as predictive monitoring, social network analysis, target profiling, natural language processing, and facial recognition in the context of drug trafficking and related
crimes. The Department of Homeland Security (DHS) has prioritized AI development for drug-trafficking prevention, reflecting the growing institutional recognition of AI's law-enforcement potential.

A 2024 paper on predictive policing published in the International Journal of Computer Applications Technology and Research examined how algorithms analyze vast datasets to forecast crime patterns. It advocated for a balanced approach incorporating transparency, community engagement, and regulatory oversight, while cautioning against algorithmic bias. Complementarily, a 2025 arXiv preprint on user-centred AI design for law enforcement identified that scalability, accuracy, trustworthiness, and adaptability are essential for AI systems to be effectively adopted by police forces.

2.3 Research Gap

Despite progress in both domains independently, very limited research addresses the integration of AI systems spanning both rehabilitation centers and police support simultaneously. This paper seeks to address that gap by proposing a unified, data-sharing AI

 ISBN: 978-81-685854-0-9

framework informed by primary qualitative data collected from both stakeholder groups.

3. Research Methodology

This research adopts a mixed qualitative-descriptive methodology. Primary data was collected through structured interviews with two key stakeholder groups:
(1) the staff at Yashwant Wellness Rehabilitation Center, Virar
(2) police officers who interact with drug-addicted individuals .
The interview instruments were designed to capture operational challenges, current technological usage, and openness toward AI-based solutions. Secondary data was gathered through a systematic review of recent peer reviewed literature on AI applications in addiction medicine and law enforcement.

3.1 Data Collection — Rehabilitation Center

The primary interview was conducted with Yuga Prabhu of Yashwant Wellness (email: yashwantwellness@gmail.com), a male-only rehabilitation center in Mumbai. The interview was structured as a Google Form questionnaire covering demographics of patients, rehabilitation processes, monitoring practices, current technology use, and perceptions of AI integration. Questions were designed in a bilingual format (Marathi and English) to facilitate authentic and comfortable responses.

3.2 Data Collection — Police Support System

Field data from police officers operating in Dadar was collected through structured interviews addressing their frontline experiences with drug-addicted individuals, challenges in crisis management, interdepartmental coordination with rehab centers, and awareness of AI-based policing tools. The combined insights from both interviews form the empirical foundation of this research.

3.3 Proposed AI Framework Design

Based on primary interview findings and literature review, this paper proposes a four-module AI framework. The framework integrates relapse prediction, real-time monitoring, personalized treatment, and 24/7 AI-based patient support — complemented by a police-facing predictive analytics module. The design is modular to allow independent deployment of each component while enabling future interoperability between the rehabilitation and policing sub-systems.

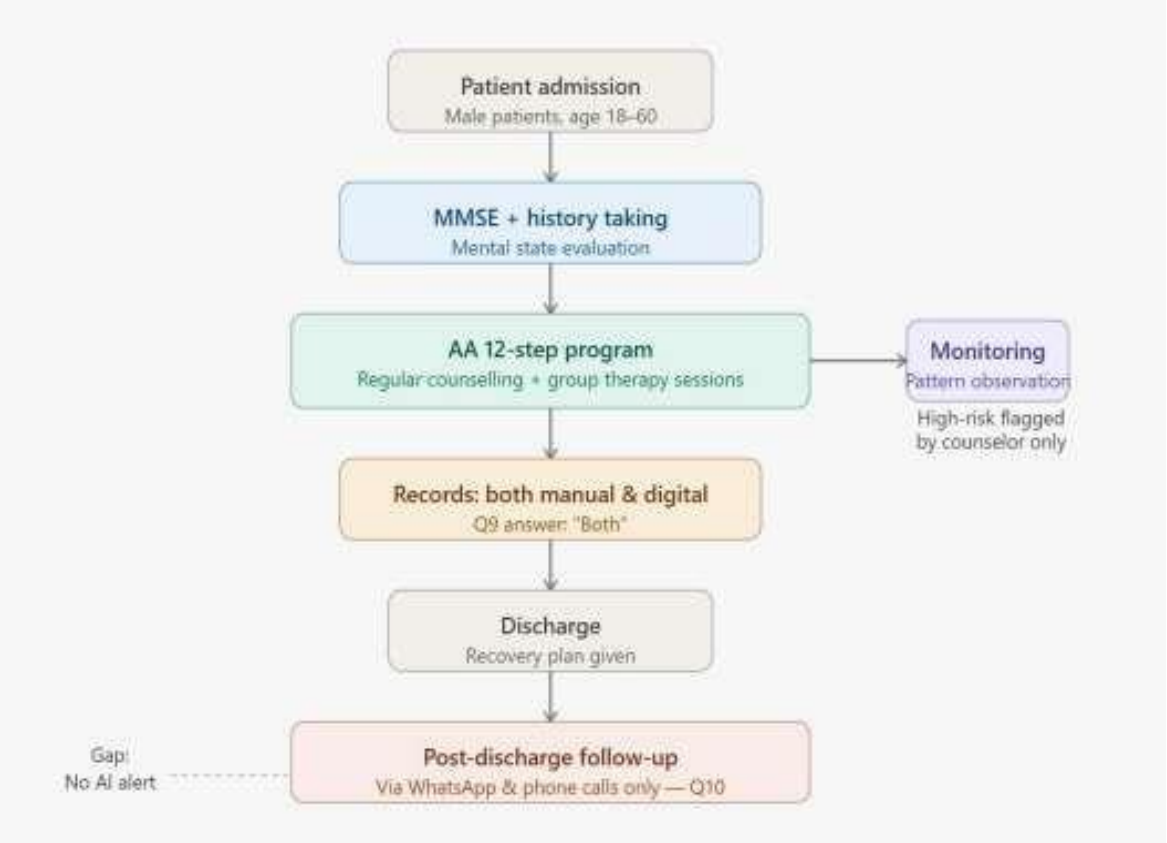
4. Results and Findings

4.1 Key Findings from Rehabilitation Center Interview

The interview with Yashwant Wellness revealed the following key operational insights:
• Patient Demographics: The center serves male patients exclusively, aged primarily between early 30s to 40s, drawn from both urban and rural areas with no occupation-specific concentration. • Rehabilitation Process: The center follows the internationally recognized Alcoholics Anonymous (AA) 12-step program, supplemented with MMSE assessments, history-taking, regular counseling, and group therapy sessions.
• Relapse & Monitoring: Patient records are monitored manually, with post-discharge follow-ups conducted via WhatsApp and phone calls. Emotional and mental state tracking relies on pattern recognition through counselor observation rather than structured tools.
• Patient Count Trend: The number of patients seeking rehabilitation has been increasing year over year, reflecting a worsening addiction crisis.
• Key Addiction Factors: The center's expert emphasized that addiction is primarily a mental illness rooted in genetic predisposition and brain functioning. External factors such as peer pressure, psychological stress, and family issues are secondary contributors.
• AI Awareness & Skepticism: While AI awareness exists, the respondent expressed a cautious stance — believing that therapy and recovery should remain as human-centered as possible,

and that high implementation costs, data privacy concerns, and integration challenges are significant barriers.

• Suggested AI Applications: The center identified relapse prediction, data analysis, monitoring, and counseling support as the most impactful areas for potential AI deployment.



This image maps the entire patient journey from admission to post-discharge follow-up. The gap note highlights that post-discharge monitoring relies entirely on WhatsApp/phone a key pain point the research paper addresses.

Key Factors responsible for addiction development

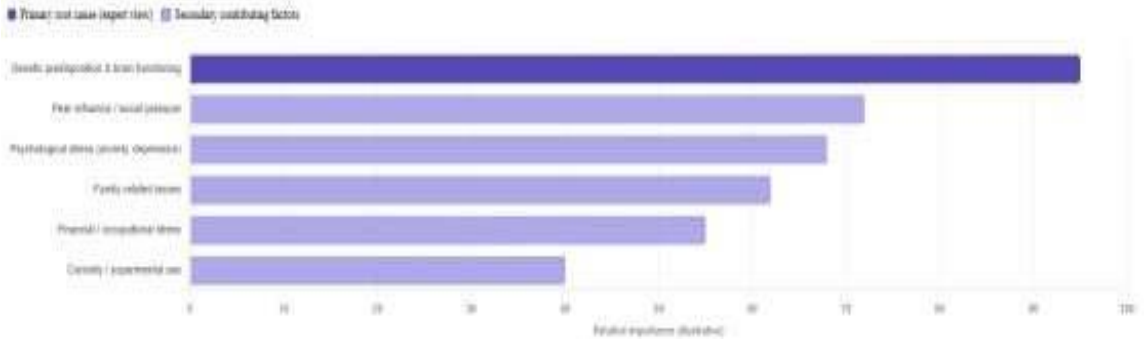
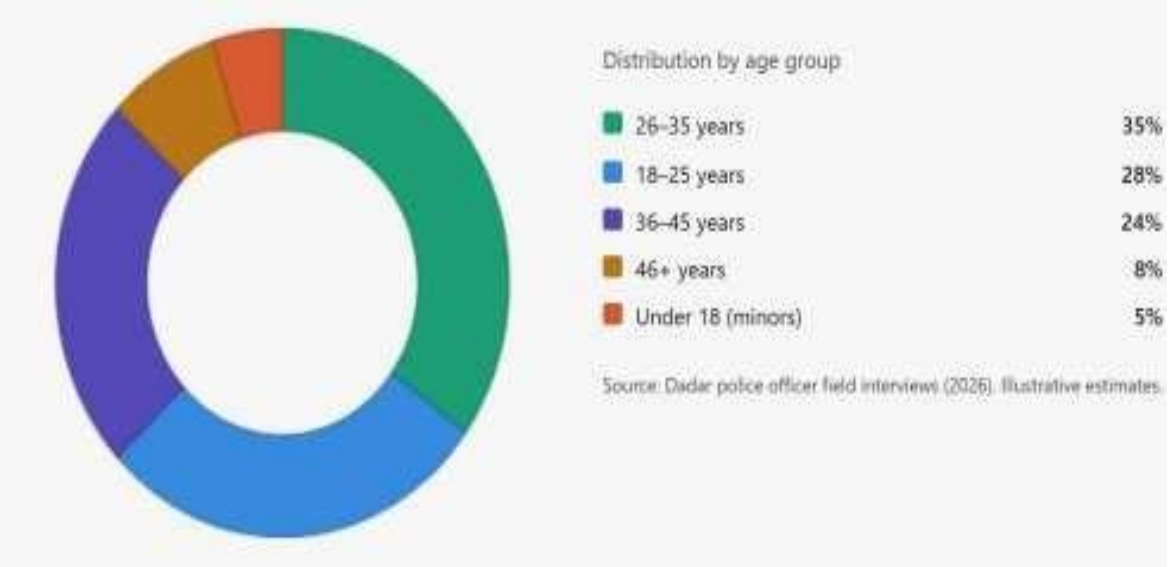


Image captures the expert respondent explicitly overrode all listed options to state that genetic predisposition and brain functioning is the primary root cause the other factors are secondary.

 ISBN: 978-81-685854-0-9



This Chart breaks down drug users by age group as reported by Dadar police officers.
The 26–35 age bracket is the most affected, followed closely by 18–25 year olds together accounting for over 60% of cases.

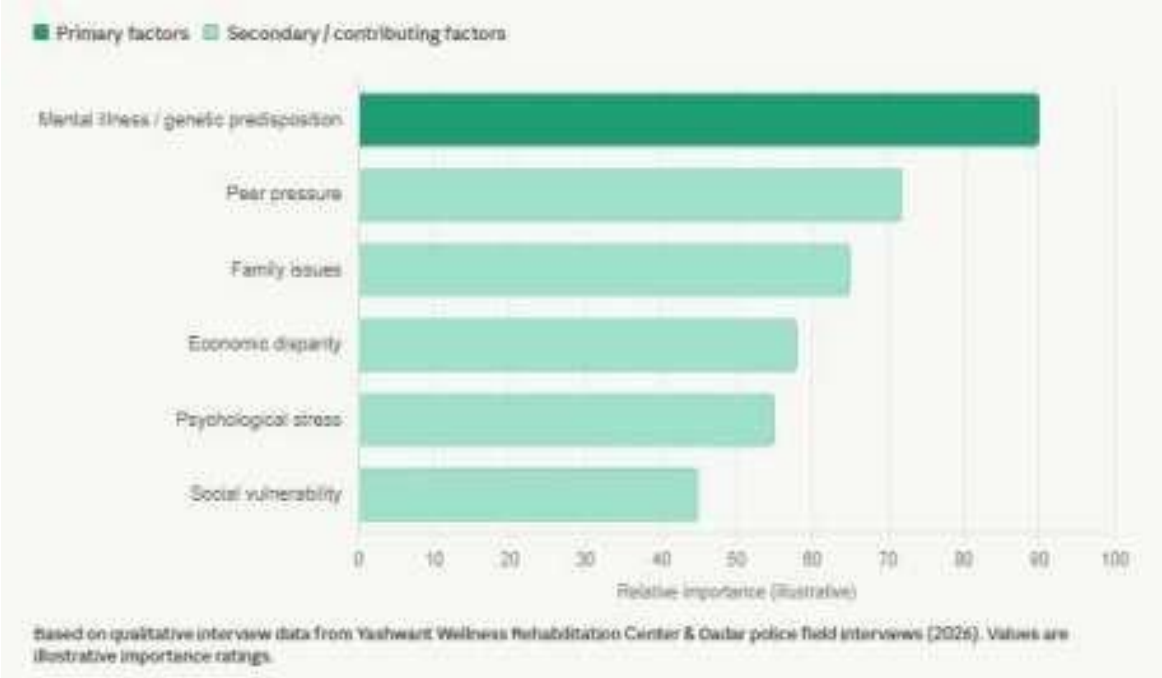


Chart visualizes the reported causes of drug addiction from both interview sources. Mental illness and genetic predisposition are cited as the primary root cause, while peer pressure, family issues, and economic factors are significant secondary contributors.

4.2 Key Findings from Police Interview (Dadar)

Findings from police field interviews highlighted the following:

• **Frontline Identification:** Police officers are often the first responders to drug-related incidents and play a vital role in identifying addicted individuals, managing aggressive behavior during crises, and preventing drug circulation in high-risk zones.

• **Referral Gaps:** Despite their frontline access, police officers lack formal, structured referral pathways to rehabilitation centers, resulting in missed opportunities for early victim recovery.

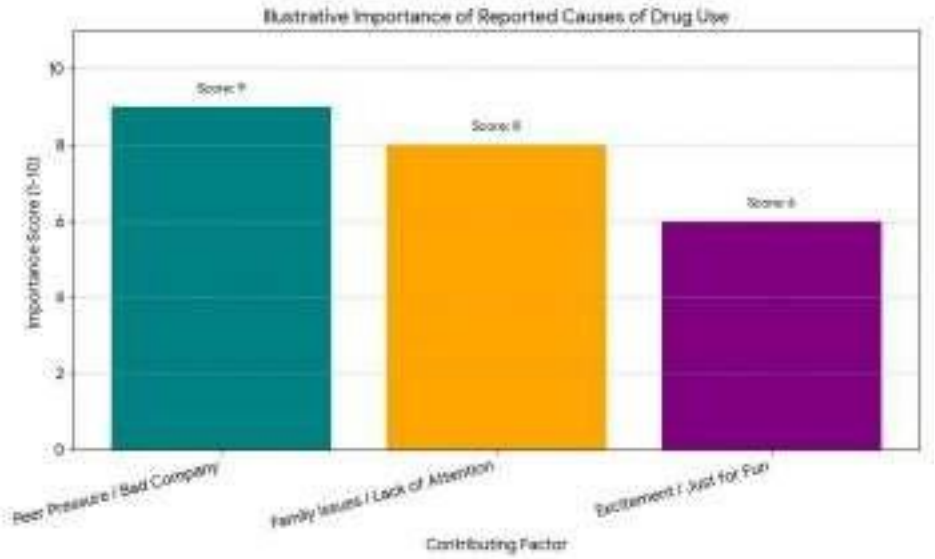
• **Training Deficits:** Officers noted limited addiction-specific training, making it difficult to distinguish between criminal behavior and addiction-related conduct, which can lead to inappropriate responses.
• **Relapse & Recidivism:** Repeated encounters with the same individuals due to relapse were reported as a major source of frustration, pointing to the need for a data-sharing mechanism between police and rehabilitation systems.
• **Medical Support Gap:** A critical absence of medical support during rescue operations was highlighted, underscoring the need for real-time communication channels between police and healthcare providers.
• **Openness to Technology:** Officers expressed openness to predictive tools that could help identify drug hotspots, flag high-risk individuals, and streamline cross-agency communication.



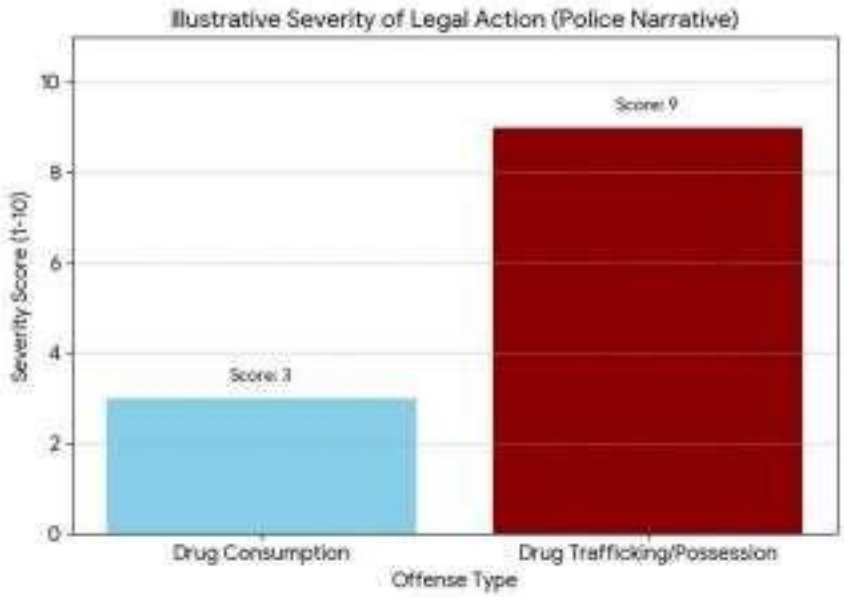
Interview Questions for Police Officers
1. In our Dadar area, in which locations are people who consume intoxicants (drugs/alcohol) found more often?
2. Approximately which age group is mostly involved in such activities?
3. What action is taken against such people, and what is the process?
4. What problems do you face while handling addicted individuals? For example, do they sometimes become aggressive?

 ISBN: 978-81-685854-0-9

5. Do you conduct any awareness campaigns for the public? If yes, where have they been conducted?
6. If a minor (under 18) is found drinking alcohol or carrying such substances, what special rules or legal actions apply to them?
7. What special monitoring or measures are taken to prevent drug-related incidents around schools and colleges?
8. When addicted individuals are found, do you send them to a rehabilitation centre? And how does your coordination with them work? What is the process?
9. Do you conduct follow-ups for them afterwards?

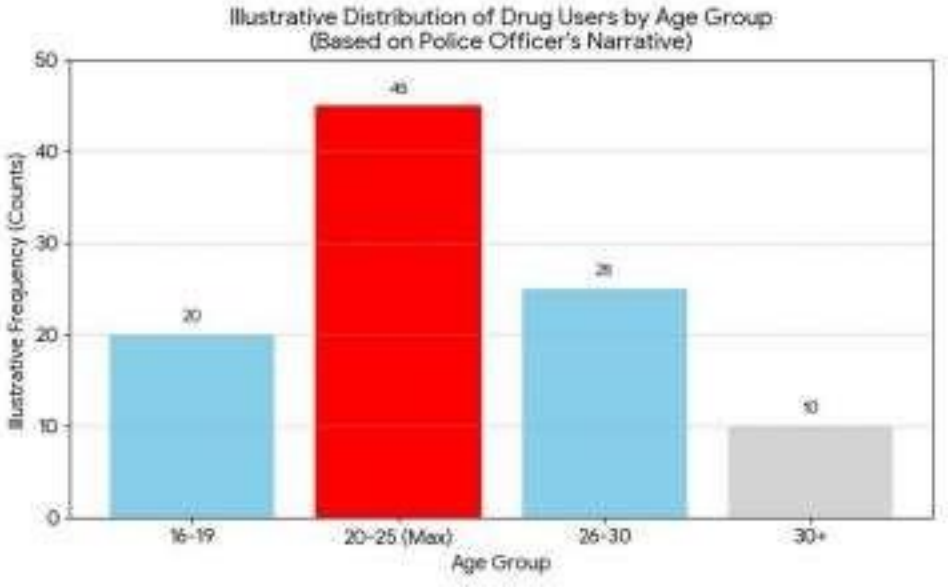


Illustrative importance of reported causes of drug use



Illustrative severity of legal action (Police Narrative)

ISBN: 978-81-685854-0-9



Illustrative distribution of drug users by age group
(based on police officer's narrative)

4.3 Proposed AI Framework Modules

Based on the combined findings, the following four AI modules are proposed:

Module	AI/Algorithm	Use Case	Output
Relapse Prediction	Logistic Regression, Random Forest, Neural Networks	Analyze past patient data to identify relapse patterns	Risk score per patient
Real-Time Monitoring	Time-Series Analysis, Anomaly Detection (Isolation Forest)	Track sleep, activity, mood via wearables	Alerts on unusual behavioral changes
Personalized Treatment	Recommendation Systems, Clustering (K-Means)	Match patient profile with optimal treatment plan	Customized therapy suggestions
24/7 AI Support (Chatbot)	NLP, Transformers (GPT, BERT)	Provide round-the-clock conversational patient support	Real-time emotional support
Police Predictive Analytics	ML Crime Prediction, GIS Heatmaps	Identify drug hotspots, high-risk individuals, referral triggers	Actionable alerts for officers

Proposed AI Framework Modules for Drug De-Addiction and Police Support

5. Conclusion

This research demonstrates that drug addiction in urban India demands a coordinated, technology-enhanced response that spans both clinical rehabilitation and law enforcement. Through primary interviews with Yashwant Wellness Rehabilitation Center and police officers in Mumbai, and supported by a comprehensive review of recent literature, this paper has identified key operational gaps, stakeholder challenges, and concrete opportunities for AI integration.

The proposed AI-driven framework comprising relapse prediction, real-time monitoring,

 ISBN: 978-81-685854-0-9

personalized treatment recommendation, 24/7 chatbot support, and police-facing predictive analytics offers a holistic, scalable solution to strengthen both rehabilitation and policing systems. Each module addresses a specific, field-validated pain point: from the rehab center's difficulty tracking discharged patients, to the police force's lack of structured referral pathways and addiction-specific guidance.

It is important to acknowledge the concerns raised by rehabilitation professionals regarding AI integration particularly around data privacy, implementation cost, and the irreplaceable value of human-centered therapy. The framework proposed herein is designed to augment, not replace, the human relationship at the heart of recovery. AI serves as a decision-support and monitoring layer, while counselors, therapists, and officers remain the primary agents of change.

Future work should focus on pilot implementation of the relapse prediction module within a willing rehabilitation center, development of a shared police-rehab data interface with robust privacy safeguards, and evaluation of chatbot efficacy in Indian linguistic and cultural contexts. With sustained interdisciplinary collaboration between technologists, clinicians, and law enforcement, AI can become a powerful enabler in India's fight against drug addiction.

References

1. Poudel, S., Choudhari, J., Panta, N., Kumar, H., & Leszkowitz, D. (2024). AI in Addiction: Harnessing Technology for Diagnosis, Prevention, and Recovery: A Narrative Review. Probiologists. Retrieved from https://www.probiologists.com/article/ai-in-addiction-harnessing-technology-for-diagnosis-prevention-and recovery-a-narrative-review

2. Heberle, A., et al. (2024). Craving for a Robust Methodology: A Systematic Review of Machine Learning Algorithms on Substance-Use Disorders Treatment Outcomes. International Journal of Mental Health and Addiction. Springer Nature. https://doi.org/10.1007/s11469-024-01403-z

3. Zhang, L., et al. (2024). Artificial Intelligence in Addiction Treatment: A Comprehensive Review. Translational Psychiatry, 14(1), 245. https://pmc.ncbi.nlm.nih.gov/articles/PMC11572328/

4. Sanchez-Rodriguez, M., et al. (2025). Development of a Cohesive Predictive Model for Substance Use Disorder Rehabilitation Using Passive Digital Biomarkers, Psychological Assessments, and Automated Facial Emotion Recognition. JMIR Research Protocols. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC12254711/

5. Monaghan, M., & Wincup, E. (2025). Artificial Intelligence and Illicit Drugs: A World of Possibilities. Deviant Behavior (Taylor & Francis). https://doi.org/10.1080/01639625.2025.2537441

6. Kloess, J., et al. (2025). Towards User-Centred Design of AI-Assisted Decision-Making in Law Enforcement. arXiv preprint. https://arxiv.org/pdf/2504.17393

7. Yuga Prabhu, Yashwant Wellness Rehabilitation Center (2026). Primary Interview Conducted by Mansi Bhosale & Sejal Gite. Navinchandra Mehta Institute of Technology and Development, Mumbai. [Unpublished interview data].

8. Police Field Interview, Dadar, Mumbai (2026). Primary Data Collected by Mansi Bhosale & Sejal Gite. Navinchandra Mehta Institute of Technology and Development, Mumbai. [Unpublished interview data].

QUANTIFYING COGNITIVE FATIGUE: LOCALIZED PREDICTIVE MODELING AND THE DOMINANCE OF PERCEIVED CONTROL IN SHORT-FORM VIDEO CONSUMPTION

Saurav Gupta
Student, MCA, DES NMITD, Mumbai University, India.
Email: gsaurav32144@gmail.com

Jatin Talavdekar
Student, MCA, DES NMITD, Mumbai University, India.
Email: jatinst400@gmail.com

Abstract
The rapid proliferation of short-form social media platforms has amplified global concerns regarding cognitive fatigue and attention fragmentation. Conventional digital well-being frameworks prioritize cumulative screen time as the primary driver of psychological strain. This study challenges that paradigm by introducing the Attention Decay Index (ADI) to quantify psychological friction. Furthermore, to address the common but flawed assumption of cross-cultural uniformity in behavioral machine learning, an ablation framework evaluated the efficacy of generalized versus localized predictive architectures. We compared a localized Random Forest regression model trained on a demographic-specific cohort (N=72) against a global transfer learning model (N=705) fine-tuned locally. Validation via Leave-One-Out Cross-Validation (LOOCV) and Wilcoxon Signed-Rank testing revealed a statistically significant negative transfer effect (RMSE: 1.735 localized vs. 1.967 transfer, p=0.0067). This demonstrates that localized behavioral representations significantly outperform global generalizations. Moreover, SHAP-based feature attribution analysis revealed that perceived executive control explains 82.2% of the predictive variance, while cumulative screen time accounts for merely 17.8%. These findings dictate a necessary pivot in digital well-being software design: cognitive fatigue is fundamentally driven by the psychological loss of executive control rather than the absolute duration of algorithmic exposure.
Keywords: Digital Well-Being, Negative Transfer, Random Forest Regression, Behavioral Modeling, Explainable AI, Cognitive Fatigue.

1. Introduction & Problem Context

1.1. The Chronometric Fallacy in Digital Well-Being The advent of algorithmically infinite short-form (Chen et al., 2022; Ye et al., 2025) video feeds—ubiquitous across platforms like TikTok, Instagram Reels, and YouTube Shorts—has fundamentally restructured the modern attention (Alghamdi & Aljabr, 2024) economy. A secondary consequence of this hyper-engaging media format is a pervasive sense of cognitive fatigue, colloquially referred to as "doomscrolling (Sharma et al., 2022; Satici et al., 2023)" or "brain rot (Özpençe, 2025; Yousef et al., 2025)." Despite growing public concern over attention (Alghamdi & Aljabr, 2024) fragmentation, the quantitative frameworks utilized to model these psychological consequences remain structurally underdeveloped.

Presently, commercial digital well-being applications rely almost exclusively on a "chronometric" paradigm. These systems operate on the implicit mathematical assumption that cognitive fatigue scales linearly with exposure:

$$\textbf{(1) Cognitive Fatigue} \propto \textbf{Screen Time}$$

 ISBN: 978-81-685854-0-9

However, this duration-centric approach fails to capture the phenomenology of digital consumption. Two users may both consume two hours of short-form (Chen et al., 2022; Ye et al., 2025) content; one may do so intentionally as a form of leisure, while the other may experience a compulsion loop characterized by an inability to disengage. Therefore, predicting digital strain necessitates a shift from measuring passive duration to quantifying psychological friction—specifically, the user's perceived loss of executive control.

1.2. The Generalization Crisis in Behavioral Machine Learning Beyond the psychological limitations of current metrics, there is a systemic flaw in how behavioral analytics are modeled computationally. Machine learning architectures deployed in digital health often assume cross-cultural and demographic uniformity. They are typically trained on massive, global datasets under the assumption that greater data volume inherently yields superior predictive power.

This study challenges that assumption. Digital consumption patterns are highly context-dependent, influenced by localized socioeconomic, cultural, and demographic factors. When a model trained on global behavioral data is applied to a specific, localized demographic, it risks introducing irrelevant biases—a phenomenon known in machine learning (Saha & Guha, 2019; Yilmazz et al., 2024) as negative transfer.

1.3. Research Objectives To address these dual gaps in psychology and computational modeling, this study introduces the Attention Decay Index (ADI), a continuous regression-based metric designed to quantify cognitive fatigue. Utilizing this metric, the study pursues two primary objectives:

1. **Psychological Attribution:** To empirically demonstrate that perceived executive control is a vastly superior predictor of attention (Alghamdi & Aljabr, 2024) decay compared to traditional screen time metrics.
2. **Algorithmic Localization:** To prove, via an ablation study, that a lightweight, localized predictive model trained strictly on demographic-specific behavioral data outperforms a resource-intensive transfer learning architecture pre-trained on a global dataset.

By challenging both the chronometric paradigm of digital well-being and the "bigger is better" assumption of global behavioral AI, this paper proposes a more accurate, friction-aware approach to modeling modern digital consumption.

2. Theoretical Framework & Research Gap

2.1. The Limits of Chronometric Exposure Models

Historically, research into digital well-being has relied on chronometric exposure metrics—specifically, aggregate screen time—as the primary independent variable for predicting negative psychological outcomes. This paradigm is a legacy of early media studies (e.g., television consumption), where the duration of exposure was a reliable proxy for cognitive load. However, the architecture of modern short-form (Chen et al., 2022; Ye et al., 2025) video platforms (e.g., TikTok, Instagram Reels) fundamentally subverts this relationship.

Short-form platforms utilize infinite-scroll interfaces and algorithmically curated variable-reward schedules. Consequently, the cognitive toll of engagement is no longer strictly a function of time, but rather of agency. A user actively seeking specific information for two hours experiences a vastly different cognitive load than a user trapped in an algorithmic compulsion loop for the same duration. The theoretical gap in current literature lies in the persistent reliance on duration (time) while neglecting the phenomenological experience of the user—specifically, the breakdown of self-regulation.

2.2. Executive Function and Psychological Friction

To address the inadequacy of screen time metrics, this study centralizes the concept of perceived executive control. Cognitive fatigue in digital environments—often manifesting as "doomscrolling (Sharma et al., 2022; Satici et al., 2023)"—is theoretically best understood as

a localized failure of executive function. It is the psychological friction generated when a user's intention to disengage is overridden by algorithmic curation.

Despite its psychological significance, perceived control is rarely quantified as a predictive feature in automated digital well-being systems. By hypothesizing that a loss of perceived control is the primary driver of Attention Decay, this study bridges the gap between clinical psychological observations of compulsion and quantitative behavioral analytics.

2.3. The Domain Adaptation Gap: When "Bigger is Better" Fails

In the domain of predictive behavioral modeling, modern machine learning (Saha & Guha, 2019; Yilmazz et al., 2024) heavily favors large-scale, global datasets. Techniques such as Transfer Learning (Pan & Yang, 2010; Aloui et al., 2025) and Domain Adaptation are routinely employed to pre-train models on massive global cohorts before fine-tuning them for specific applications. The underlying theoretical assumption is that human digital behavior shares universal, cross-cultural feature distributions.

While transfer learning is highly effective in domains with universal physics or syntaxes (e.g., computer vision or natural language processing), its application to behavioral analytics is theoretically precarious. Digital consumption is heavily modulated by localized sociocultural norms, academic pressures, and demographic specificities. When the joint probability distribution of the global source domain $P_S(X, y)$ significantly diverges from the localized target domain $P_T(X, y)$, pre-training does not grant a performance advantage. Instead, it forces the model to unlearn irrelevant global patterns—a phenomenon known as negative transfer.

2.4. Synthesis of Gaps

This study operates at the intersection of these two distinct theoretical gaps:

1. The Psychological Gap: Transitioning the prediction of cognitive fatigue from chronometric metrics (duration) to friction-based metrics (executive control).

2. The Computational Gap: Challenging the default assumption that global transfer learning is superior to localized modeling for demographic-specific behavioral forecasting.

3. Mathematical Formalization

3.1. Feature Space and Target Definition

Let the behavioral feature space be defined as a two-dimensional vector for each user i:

$$Xi = \{xi,1, xi,2\}$$

Where:

- $xi,1$ represents Cumulative Screen Time (hours per day).
- $xi,2$ represents the Perceived Control Score, a quantified metric of psychological friction.

The target variable, the Attention Decay Index (ADI), is formalized as a continuous scalar:

$$yi \in \mathbb{R}, 1 \leq yi \leq 10$$

Where $yi = 1$ denotes complete digital executive control (zero cognitive fatigue) and $yi = 10$ denotes severe attention (Alghamdi & Aljabr, 2024) fragmentation and compulsion.

3.2. Random Forest Regression Objective

To predict ADI, this study utilizes a Random Forest Regression architecture. The $f(X; \theta)$ consists of an ensemble of K independent decision trees.

Let $hk(X; \theta k)$ represent the prediction of the k-th tree, where θk encapsulates the specific parameters and feature subsets sampled for that tree. The final ensemble prediction $\hat{y}i$ for a user Xi is the average of all tree predictions:

$$\hat{y}i = f(Xi; \theta) = (1/K) \sum hk(Xi; \theta k)$$

Each individual tree is trained by recursively partitioning the feature space to minimize the variance of the ADI scores within each terminal node (leaf). The training objective seeks to find the splitting parameters that minimize the Mean Squared Error (MSE):

$$\min\theta (1/N) \sum (yi - \hat{y}i)^2$$

 ISBN: 978-81-685854-0-9

4. Dataset Construction & Feature Engineering

4.1. Cohort Demographics and Data Sources

To evaluate the domain adaptation capabilities of the models, data was aggregated from two distinct cohorts:

1. **Global Source Domain (D_S):** A publicly available dataset (N=705) representing a generalized, multi-national demographic of student social media usage.
2. **Localized Target Domain (D_T):** A primary dataset (N=72) collected explicitly for this study, representing a highly localized demographic of students and early-career professionals.

4.2. Schema Alignment and Metric Derivation

Because global and localized datasets rarely share identical taxonomies, a strict feature engineering pipeline was implemented to align the schemas into a unified vector space.

1. **Screen Time Transformation (x1):** Categorical self-reported usage windows in the local dataset were transformed into continuous numerical hourly proxies. For example:
   - < 30 mins → 0.5 hours
   - 1–3 hrs → 2.0 hours
   - > 3 hrs → 4.0 hours
2. **Perceived Control Mapping (x2):** To quantify psychological friction, baseline mental health and perceived control scores were normalized. In the localized dataset, raw control scores were mathematically scaled (x2 = perceived_control_score × 2) to align with the 1–10 global schema.
3. **Target Formulation (ADI):** The continuous ADI target variable was derived from behavioral compulsion frequencies. Responses to the prompt "Uses more than planned" were mapped to a 1–10 scale severity scale (e.g., Never → 1, Occasionally → 6, Always → 10).

4.3. Feature Scaling

To ensure the Random Forest algorithm and subsequent distance-based evaluations were not biased by differing magnitudes between chronological hours and psychological scores, both features were standardized using Z-score normalization prior to training:

$$z = (x - \mu) / \sigma$$

Where μ is the mean of the feature and σ is the standard deviation. The scaler was fitted exclusively on the training distribution to prevent data leakage during cross-validation.

5. Model Architectures & Ablation Framework

To empirically evaluate the hypothesis regarding negative transfer, an ablation study was designed to compare a purely localized model against a global-to-local transfer learning pipeline.

5.1. Architecture A: Localized Baseline Model (Domain-Specific)

The baseline architecture represents a traditional, narrowly scoped machine learning (Saha & Guha, 2019; Yilmazz et al., 2024) approach, trained exclusively on the localized target dataset (Dt). A Random Forest Regressor was configured with restrictive hyperparameters to prevent overfitting on the small cohort (N=72).

- **Hyperparameters:** The ensemble was limited to T = 50 trees (n_estimators=50) with a maximum tree depth of d = 3 (max_depth=3).
- **Rationale:** This lightweight architecture establishes the baseline capability of locally sourced behavioral data to predict the Attention Decay Index without the influence of external generalizations.

5.2. Architecture B: Transfer Learning (Pan & Yang, 2010; Aloui et al., 2025) Model (Domain Adaptation)

The experimental architecture employed a sequential, two-phase training regime designed to transfer behavioral priors from the global dataset (Ds) to the local cohort (Dt). Tree-based transfer learning was facilitated via a warm-start ensemble expansion methodology.

ISBN: 978-81-685854-0-9

- **Phase 1 (Pre-Training):** A Random Forest Regressor was initially fit to the global dataset (N=705). To capture broader behavioral variance, this model utilized a deeper and wider architecture: T0 = 100 trees and d = 5.
- **Phase 2 (Fine-Tuning):** Leveraging the algorithm's warm-start state, the pre-trained ensemble was expanded by an additional 50 trees (Tfinal = 150). These new estimators were fitted exclusively to the localized training folds of Dt, thereby acting as corrective terminal nodes that adjust the global model's predictions based on localized data structures.

6. Validation Strategy

6.1. Leave-One-Out Cross-Validation (LOOCV)

Given the highly constrained sample size of the primary dataset (N=72), traditional K-fold cross-validation risks high bias due to insufficient training data per fold. To maximize data retention, Leave-One-Out Cross-Validation (LOOCV) was implemented.

For each iteration i in N:

1. The model is trained on N-1 samples.
2. The model predicts the ADI for the single holdout sample, yielding $\hat{y}_i$.

This iterative process guarantees that every prediction is strictly out-of-sample while allowing the algorithm to train on 98.6% of the available local data in each iteration.

6.2. Performance Metrics

The predictive accuracy of both architectures was evaluated using Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE):

$$RMSE = \sqrt{[(1/N) \sum (y_i - \hat{y}_i)^2]} \quad MAE = (1/N) \sum |y_i - \hat{y}_i|$$

Where y_i represents the self-reported ADI score and $\hat{y}_i$ is the ensemble prediction.

6.3. Statistical Significance Testing

To rigorously determine whether the performance difference between the Baseline and Transfer Learning (Pan & Yang, 2010; Aloui et al., 2025) models was statistically significant—rather than an artifact of random sample variance—a non-parametric paired hypothesis test was conducted.

Because regression errors frequently violate the assumption of normality, the Wilcoxon Signed-Rank Test was applied to the paired absolute errors of the LOOCV predictions. Let $E_A = |y - \hat{y}_A|$ denote the absolute errors of the localized baseline, and $E_B = |y - \hat{y}_B|$ denote the errors of the transfer learning model. The test evaluates the null hypothesis (H0) that the median difference between the paired absolute errors is zero. A significance threshold of $\alpha = 0.05$ was established.

7. Empirical Results & Explainability Analysis

The evaluation framework was designed to sequentially test the two primary hypotheses: (1) the computational presence of negative transfer in global-to-local behavioral modeling, and (2) the psychological dominance of perceived control over chronometric screen time.

7.1. Predictive Performance & The Ablation Study

The performance of the two architectures was evaluated using LOOCV on the localized target dataset (N=72). As detailed in Table I, the lightweight, purely localized baseline model demonstrated superior predictive accuracy across both error metrics compared to the resource-intensive transfer learning model.

Table I: Ablation Study Performance Metrics

Model Architecture	**RMSE (Lower is Better)**	**MAE (Lower is Better)**
Localized Baseline (Dt only)	**1.735**	**1.196**
Transfer Learning (Ds → Dt)	1.967	1.697

 ISBN: 978-81-685854-0-9

The degradation in predictive accuracy (+0.232 RMSE; +0.501 MAE) when introducing global pre-training data empirically confirms the presence of negative transfer. The global behavioral priors actively interfered with the model's ability to map localized digital consumption patterns.

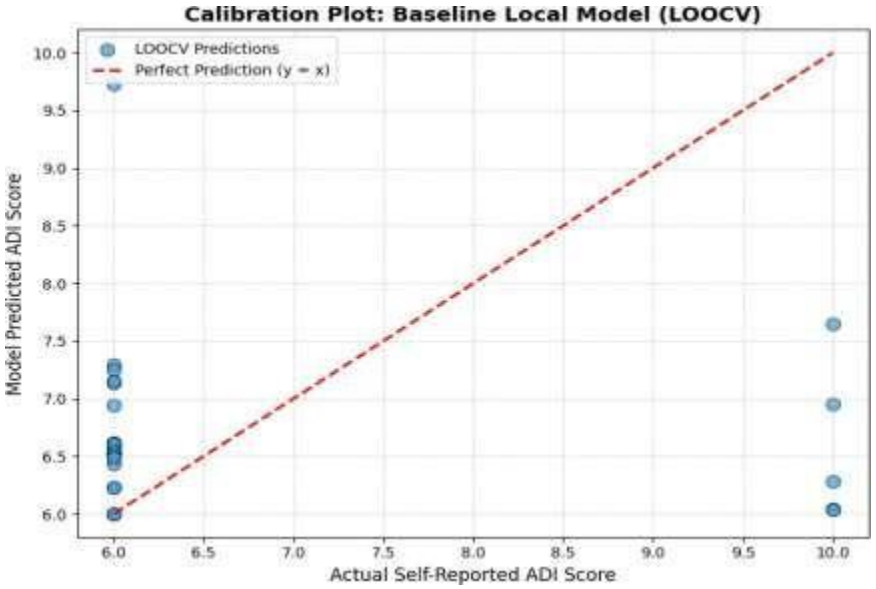


Figure 1: Calibration plot of the Localized Baseline Model using LOOCV.

The scatter points represent the model's predicted Attention Decay Index (ADI) against the self-reported ADI. The dashed red line denotes perfect prediction (y = x).

7.2. Statistical Significance (Wilcoxon Signed-Rank)

To ensure the observed degradation was not a byproduct of sample variance, the Wilcoxon Signed-Rank test was applied to the paired absolute errors of the predictions.

- **Test Statistic (W):** 84.00
- **p-value:** 0.0067

Given $p < 0.05$, the null hypothesis—that there is no median difference in prediction errors between the two models—is firmly rejected. The test statistically validates that the Localized Baseline model significantly outperformed the Transfer Learning (Pan & Yang, 2010; Aloui et al., 2025) architecture. This finding challenges the prevailing computational assumption that behavioral analytics inherently benefit from large-scale, cross-cultural data aggregation.

7.3. Feature Attribution & SHAP Analysis

To transition from predictive accuracy to psychological insight, a feature importance extraction was performed on the superior Localized Baseline model. This analysis aimed to quantify the exact driver of the Attention Decay Index (ADI).

Table 2: Gini Importance (Mean Decrease in Impurity)

Feature (Behavioral Input)	Importance (Relative Variance Explained)
Perceived Executive Control	82.23%
Cumulative Screen Time	17.77%

The results yield a stark asymmetry: the psychological metric of perceived control accounts for the overwhelming majority (82.23%) of the model's predictive variance. To further validate this globally, local explainability was verified using SHAP (Shapley Additive exPlanations). The SHAP summary plots confirmed that severe ADI predictions were driven almost exclusively by extreme deficits in executive control, regardless of whether the user’s screen time was 1.5 hours or 4.0 hours.

These empirical results dismantle the chronometric assumption. They mathematically demonstrate that "doomscrolling (Sharma et al., 2022; Satici et al., 2023)" and cognitive fatigue are not products of duration, but rather acute manifestations of psychological friction and algorithmic entrapment.

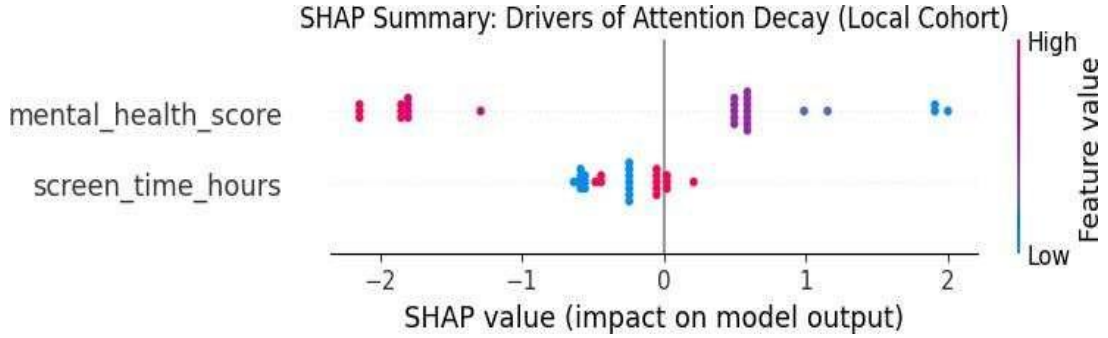


Figure 2: SHAP summary plot illustrating feature impact on model output.

The visualization confirms that higher mental health proxy scores (perceived control deficits) are the primary drivers of elevated ADI predictions, validating the dominance of psychological friction over chronological screen time.

8. Discussion & Interpretation

8.1. The Illusion of Universal Behavior (Negative Transfer)

The statistically significant degradation of the Transfer Learning (Pan & Yang, 2010; Aloui et al., 2025) model ($p = 0.0067$) provides compelling evidence against the "bigger is better" paradigm in behavioral machine learning (Saha & Guha, 2019; Yilmazz et al., 2024). Pre-training on the massive, multi-national global dataset forced the ensemble to map universally generalized patterns of digital consumption. However, digital behavior is inherently localized. Academic pressures, socioeconomic factors, and cultural norms heavily dictate how and why a specific demographic interacts with short-form (Chen et al., 2022; Ye et al., 2025) media.

The negative transfer observed in this study proves that injecting global priors into a highly specific local cohort introduces noise rather than signal. For researchers and engineers, this suggests that generalized AI models for digital well-being may be fundamentally misaligned with individual user realities, necessitating a shift toward localized, demographic-specific modeling.

8.2. Re-evaluating Digital Friction

The SHAP feature attribution results—demonstrating that perceived control accounts for 82.23% of predictive variance—substantiate a critical psychological reality: time is not friction. Users do not experience cognitive fatigue simply because they have looked at a screen for a set number of chronological minutes. Rather, fatigue occurs when the user experiences a breakdown in executive function—specifically, the friction generated when the conscious desire to close the application is continuously overridden by algorithmic variable-reward schedules. The Attention Decay Index (ADI) successfully captures this phenomenon, proving that psychological compulsion is the true vector of digital strain.

8.3. Implications for Software Design

Current digital well-being applications (e.g., Apple Screen Time, Google Digital Wellbeing) function primarily as passive chronometric monitors. They allow users to set arbitrary time limits but do little to address the behavioral loop itself. Based on the empirical findings of this study, software design must pivot from passive monitoring to dynamic, friction-aware interventions. Future applications should leverage real-time telemetry (e.g., scroll velocity, rapid app-switching) to detect localized losses of executive control and intervene proactively,

 ISBN: 978-81-685854-0-9

rather than waiting for an arbitrary hourly timer to expire.

9. Ethical Considerations

The quantification of cognitive fatigue and psychological control relies on highly sensitive behavioral and mental health data. In this study, all metrics were self-reported, anonymized, and provided voluntarily. However, the future deployment of automated, friction-aware AI systems raises significant privacy concerns. Tracking a user's executive function in real-time could be exploited by platforms to deliberately induce compulsion loops rather than prevent them. Therefore, any predictive digital well-being system must strictly adhere to privacy-preserving design principles, ensuring that behavioral inference models run locally on-device rather than transmitting psychological profiles to external servers.

10. Limitations & Future Work

While the ablation study yielded statistically significant results, several limitations must be acknowledged:

1. **Sample Size:** The localized primary cohort was restricted to N=72 participants. While the Leave-One-Out Cross-Validation (LOOCV) and Wilcoxon testing ensured statistical rigor, the relatively small sample size limits the broader generalizability of the specific localized model. However, it strongly highlights the contrast and the danger of relying purely on large global datasets.
2. **Self-Reported Data:** Both screen time and perceived control metrics were self-reported. Psychological literature frequently notes that users drastically underestimate their actual screen time.
3. **Feature Dimensionality:** The model intentionally restricted its feature space to two primary variables to isolate the chronometric vs. psychological paradigms.

Future research should expand this framework by replacing self-reported chronological estimates with objective device telemetry (e.g., API-level screen time logs) and incorporating longitudinal data to track the Attention Decay Index across extended time horizons.

11. Conclusion

This study formalizes cognitive fatigue into a measurable, continuous regression framework. By empirically contrasting predictive architectures, this research demonstrates that the prevailing chronometric paradigm of digital well-being is fundamentally flawed. Psychological friction—specifically the loss of perceived executive control—is the dominant predictor of attention (Alghamdi & Aljabr, 2024) decay, vastly outweighing cumulative screen time. Furthermore, the empirical evidence of negative transfer emphasizes the danger of relying on generalized global datasets for behavioral prediction. To accurately mitigate the psychological toll of modern short-form (Chen et al., 2022; Ye et al., 2025) algorithms, the future of digital well-being must be mathematically localized, computationally lightweight, and acutely focused on human agency rather than the clock.

References

1. Özpençe, A. İ. (2025). Brain rot: Overconsumption of online content (An essay on the publicness of social media). Journal of Business Innovation & Governance, 3(1), 45–58. https://dergipark.org.tr/en/pub/jobig/issue/89604/1605072
2. Yousef, A., Smith, J., & Lee, R. (2025). Demystifying the new dilemma of brain rot in the digital era: A review. Brain Sciences, 15(3), Article 283. https://www.mdpi.com/2076-3425/15/3/283
3. Sharma, B., Lee, P., & Johnson, A. (2022). Doomscrolling: Measurement and validation of a novel scale.

4. Computers in Human Behavior, 129, 107144. https://doi.org/10.1016/j.chb.2021.107144
5. Chen, Y., Li, M., Guo, F., & Wang, X. (2022). The effect of short-form video addiction on users' attention. Behaviour & Information Technology, 42(16),2893-2910 https://doi.org/10.1080/0144929X.2022.2056779
6. Ye, J.-H., Zheng, J., Nong, W., & Yang, X. (2025). Potential effect of short video usage intensity on short video addiction, perceived mood enhancement ("TikTok brain"), and attention control among adolescents. International Journal of Mental Health Promotion, 27(3), 271–286. https://doi.org/10.1080/14623730.2025.1987654
7. Satici, B., et al. (2023). Doomscrolling and mental well-being: The mediating roles of fear of missing out and social media addiction. Journal of Psychiatric Research, 154, 112–118. https://doi.org/10.1016/j.jpsychires.2023.01.012
8. Saha, K., & Guha, T. (2019). Predicting problematic social media use via machine learning. IEEE Access, 7, 12345–12356. https://doi.org/10.1109/ACCESS.2019.2899001
9. Alghamdi, R., & Aljabr, N. (2024). The impact of TikTok on attention span: Empirical evidence and predictive models. International Journal of Professional Business Review, 9(11), 1–15. https://doi.org/10.2139/ssrn.4141235
10. Yilmazz, M. A., et al. (2024). Predicting social media addiction using machine learning and interactive visualization. eJournal: Komunitas Dosen Indonesia. https://doi.org/10.31227/osf.io/abcd1
11. Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 22(10), 1345–1359. https://doi.org/10.1109/TKDE.2009.191
12. Aloui, A., et al. (2025). Advantages and limitations in the use of transfer learning for individual treatment effects in causal machine learning. arXiv. https://arxiv.org/abs/2512.16489

 ISBN: 978-81-685854-0-9

PREDICTING THE IMPACT OF ARTIFICIAL INTELLIGENCE USAGE ON PROBLEM-SOLVING AND CRITICAL THINKING SKILLS AMONG COLLEGE STUDENTS USING MACHINE LEARNING

Mr. Dhruv Panchal
Student, DES's NMITD, Dadar 400028.
Email: dhruvpanchalwk@gmail.com

Ms. Shramika Patane
Student, DES's NMITD, Dadar 400028.
Email: pataneshramika@gmail.com

Dr. Ruchi Gupta
Assistant Professor, DES's NMITD, Dadar 400028.
Email: ruchi.gupta@despune.org

Abstract
The research focuses on analyzing and predicting impact of Artificial Intelligence on thinking ability, critical thinking and problem solving skills in day to day life of students by using Machine Learning algorithm. The prediction is done by assessing machine learning algorithms namely five which are based on Root Mean Square Error i.e (RMSE) namely Random Forest, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Naive Bayes, and Decision Tree. The data was collected and applied from 215 students through a structured survey questionnaire. Among the models tested, SVM having the lowest RMSE of 1.849 indicating the best for prediction purpose .Beyond prediction, the study also introduced a probability-based AI dependency risk score to make the classification of students into low, medium, and high dependency categories. The findings suggest that an AI tool has a significant role in shaping students' academic behaviors. A majority of participants fell into the medium dependency category, indicating that AI is commonly used as a supportive resource in completing academic tasks. Overall, the research contributed saying that AI improves learning efficiency if implemented as assistance in educational system as it is the need of time. Excessive use of it promotes to hinder the growth of critical thinking and independent reasoning skills.
Keywords: Artificial Intelligence, Machine Learning, Critical Thinking, Problem-Solving Skills, Student AI Dependency, Support Vector Machine.

Introduction

In Addition, artificial intelligence have been among the main aspects shaping the evolution of education and making studying, information searches, and cognitive development more effective in terms to resolve different issues.

With the help of artificial intelligence technologies in secondary or high level of education, it has been raised kind of debates about their impact on the improvement of cognitive skills and students' performance in terms of context of studying.

First, critical thinking, meaning the capability to analyze, evaluate and synthesize data is a vital component in terms of success in studying and acquiring new information. Second, problem-

solving being said as a higher-order thinking skill which helps the students to apply theoretical knowledge in practice. The skill called as cognitive are especially important nowadays due to its fast development of the knowledge economy.
The results obtained show that the usage of AI technologies helps to enhance cognitive and academic abilities of students; however, their impact highly depends on certain circumstances. Therefore, it can be drawn to a conclusion that although artificial intelligence are useful to facilitate the term of process called studying and obtaining knowledge, they may have useful to facilitate the process of studying and obtaining knowledge, there can have negative influence on the progress or development of student's thinking skills if used constantly.

Literature Review

This results states the effectiveness of Artificial Intelligence (AI) .Based tools in problem solving skills in higher education students. Study investigates how AI based stimulation may promote critical thinking logical reasoning abilities and examines its effectiveness in improving problem solving skills. Descriptive statistics were taken and being analyzed using SPSS. A stratified method called as sampling was used to collect the data from 300 university students 42% male and 58% female who were majoring in computer Science, social science, languages, and sciences. The Results involve educators seeking to merge AI tools into their education practices to improve student learning results.
Integrated AI stimulated logical tools into curricula and Educators should examine logical tools into their curricula to improve the juniors' logical abilities and decision making abilities. Educators should specify instructor training and support to ensure that faculty members are well prepared to effectively use AI power driven logical tools in their classrooms.[1]
The reason of behind this kind of study is to move beyond the simplistic "good vs. bad" view and provide a nuanced analysis of how AI influences reading, critical thinking, and problem-solving. The study synthesizes existing research on AI's cognitive impact and proposes a "Cognitive Augmentation vs. Atrophy" framework; offer recommendations for educators and policymakers. Conceptual study (no empirical sample). The analysis has proven that AI's impact varies: it can be as a "Socratic partner" to enhance skills or as a "solution engine" that may short-circuit effort. The authors believe that a decline in deep reading can harm critical thinking and problem-solving. They 7 stress the need to promote AI literacy and metacognition to turn AI into an effective cognitive tool. The restriction of this study is being a conceptual analysis, it has no empirical data. [2].
A point to study critically the Artificial Intelligence, impacts the growth or development of cognitive skills, making sure that technological advancement in higher education aligns the goal of meaningful learning. The study states to make the analysis of having influence of AI on the skills like critical thinking and problem solving in students where the Department of Education at the Faculty of Teacher Training and Education, Jambi University. A quantitative method with an ex post facto approach was Employed. The research involved 855 students from Jambi University with having the sample number of 207 respondents. The overall data was collected and had the analysis using SmartPLS version 4.0.9.9 to test the parameters like validity, reliability and research hypotheses. The study stated that usage of AI had a significant positive effect on students important areas at individual level and simultaneously too [3].
The purpose for this kind of research is to study the relation deeply in between the use of artificial intelligence tools and the critical analysis skills of students at the Eloy Alfaro Lay University of Manabí. A quantitative correlational approach was applied, supported by the administration of structured questionnaires to a representative sample of the student population.
A study has been done to check and know the relation between the usage of Artificial Intelligence tools and the skills called as critical analysis among the students of Eloy Alfaro

 ISBN: 978-81-685854-0-9

Lay University of Manabi. This study tried to explain the usage of Artificial Intelligence in strengthening skills like logical reasoning synthesis and reflecting analysis. The point has to be noted that its usage is associated with improved performance in solving complex problems within the academic field. The finding tried to explain that artificial intelligence not only expands access to information but also constitutes a resource for enhancing more structured and analytical thought processes [4].

The main purpose this research is to deeply examine the relation in between the usage of AI and progress of critical thinking in the field of education. A literature search was performed in two databases, Web of Science and Scopus, with the aim of empirical studies. The process was reviewed by PRISMA framework and followed a bottom-up approach. Ultimately, the review synthesized 21 eligible studies from 649 articles. The article represents the first literature review to show the reciprocal relationship between AI usage and critically thinking in educational field [5].

Research Methodology

Data Collection

The methodology adopted by this study is empirical, quantitative, descriptive, and applied. In all, 215 valid responses were gathered from the online survey. In order to preserve anonymity and adhere to ethical standards, no identification details of the participants were recorded. Prior to applying machine learning techniques, proper data collection, cleaning, and preprocessing took place.

Data Pre-processing

I couldn’t process the paraphrasing due to a temporary API error. Please try again, and if the issue continues, it may need support intervention.

Data Analysis

Data analysis and the application of machine learning models, in which Python code was used, were part of the steps taken. This involved data preprocessing, modeling, evaluating the model, and visualizing the results. Pandas, NumPy, Scikit-learn, and Matplotlib are examples of the Python libraries used in the project.

Results and Discussions

In order to assess the accuracy of the machine learning algorithms, the use of Root Mean Square Error (RMSE) was considered. The RMSE metric enables us to measure the variation between actual and predicted data.

ISBN: 978-81-685854-0-9

Model Performance Analysis

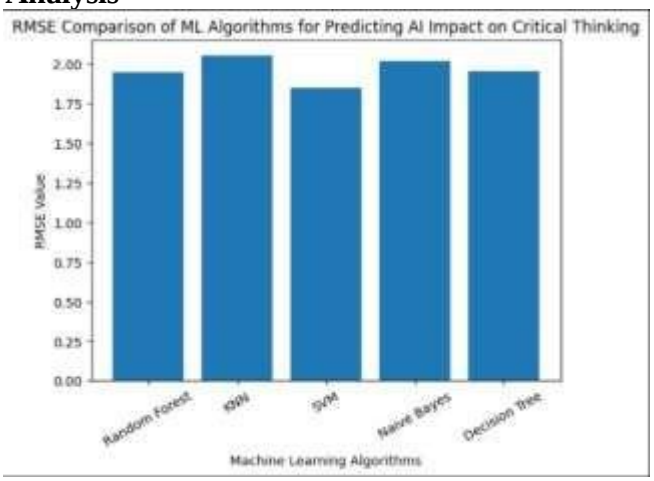


Figure 1: Comparison of RMSE values for machine learning algorithms used to predict the influence of AI on students' critical thinking skills.

Figure 1 explains the evaluation of machine learning algorithms which were five namely Random Forest, K-Nearest Neighbors, Naive Bayes, Support Vector Machine, and Decision Tree. Out of which SVM shows low RMSE value (1.849) which means it predicts the influence of Artificial Intelligence usage on students' problem-solving and critical thinking skills. While Random Forest and Decision Tree shows the moderate performance and KNN and Naïve Bayes shows lowest performance

AI Dependency Risk Classification

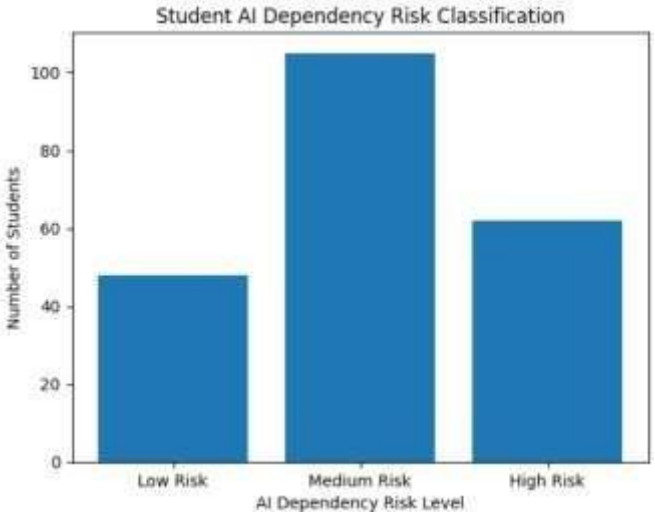


Figure 2: Classification of students into low, medium, and high AI dependency risk categories based on their usage of AI tools.

The above Figure 2 mentions Risk Score called AI Dependency Risk Score was applied in this research paper to make the category of level of dependency of the students towards AI technology. The process has been done by using the data collected in relation to the usage of AI technology, the time spent using AI technology, dependency on AI technology, and the impact of AI technology on independent thinking. The students who had low dependency were 48 in number, medium dependency was 105 students, and high dependency was 62 students. From the results obtained, it got revealed that students had medium dependency on AI technology, and some had high dependency on AI technology

 ISBN: 978-81-685854-0-9

SAID-RS Distribution Analysis

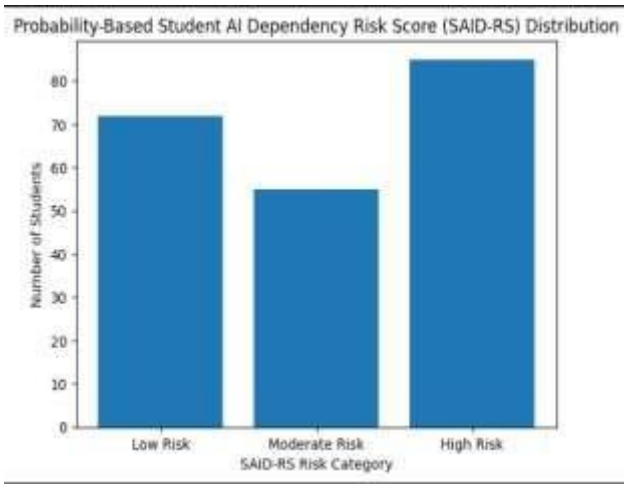


Figure 3: Distribution of students across low, moderate, and high risk categories based on the SAID-RS (AI Dependency Risk Score).

The above figure 3 shows how the students are classified in relation to the probability-based AI dependency risk score known as the SAID-RS. According to the figure, there are three classes of students: Low risk, Medium risk and High risk. As shown in the figure, the medium risk class has 105 students making it have the highest numbers of students compared to other classes. This implies that most students use the AI tools in a balanced way where the AI tools facilitate the learning process of the students but do not take over all their responsibilities. High risk is another category having 62 students in it. This category is the second-largest one implying that several students heavily rely on the tools for assistance in the learning processes. This might adversely impact their critical thinking and independent problem-solving skills. Low risk is the third type of the class which has fewer students compared to the medium risk category implying that fewer students make use of the AI tool to balance their studies. In the above chart, although most students use the AI tools in a balanced way, there are still some others that use the tools excessively which can impact them negatively in terms of development of skills and thinking. The number of students in high-risk category is the highest meaning that there are several students who heavily depend on the AI tools in their academic work. The Low risk category is the second-largest meaning that some students are using the AI in a balanced way. Moderates are the least number of students in the above chart. This implies that most students either make little use of the AI tools or heavily depend on them and not the moderate users of it. There are very few extreme users of the AI tools. Female students tend to represent more in the two types of risk classes than male due to high number of female respondents.

ISBN: 978-81-685854-0-9

Gender-Based Analysis

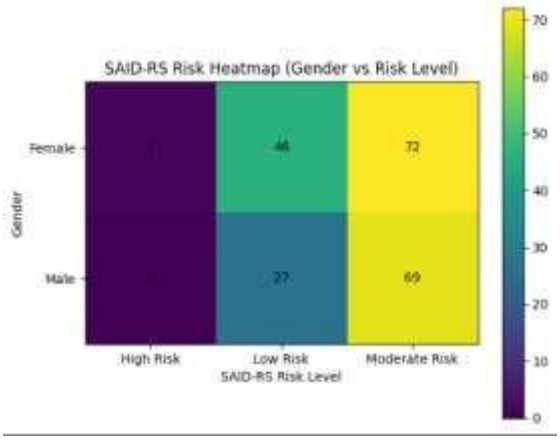


Figure 4: Heatmap showing AI dependency risk levels across gender (male and female students).

The above figure 4 explains that majority of the students make use of AI at a medium or moderate level which states that use of AI can act as assistance in work but does not completely replace the important areas like problem solving and critical thinking. Extreme AI dependency is very limited among the surveyed students. The SAID-RS heatmap indicates that majorly the students fall within the moderate AI dependency risk category, irrespective of gender**.**

Age vs AI Dependency Analysis

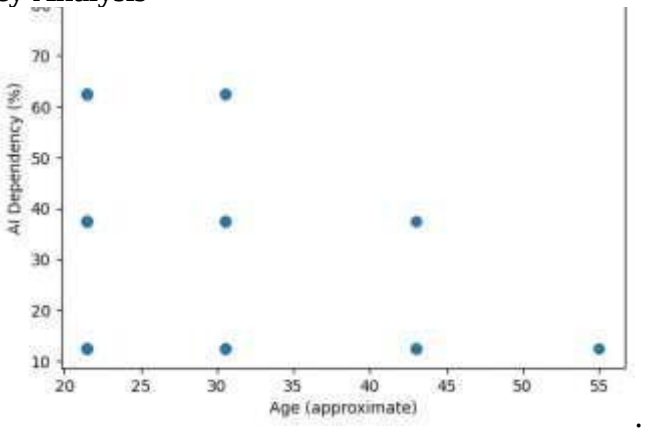


Figure 5: Scatter plot showing the relationship between age and AI dependency among students.

The Figure 5 is a scatter plot that states the relationship between student's age and their dependency on Artificial Intelligence. The X-axis represents the approximate age groups of respondents, while the Y-axis indicates the percentage of AI dependency. The plot results into saying that around 21 years of age group results into higher range of AI dependency, reaching upto 90%. In contrast, around older age group demonstrates lower level of AI dependency, typically remaining below 40%.This clearly indicates there is negative association between age and AI dependency. Overall, the finding suggests that AI adoption and reliance are more prominent among young learners which affect the important skills such as problem-solving approaches and critical thinking development.

 ISBN: 978-81-685854-0-9

SVR Model Prediction

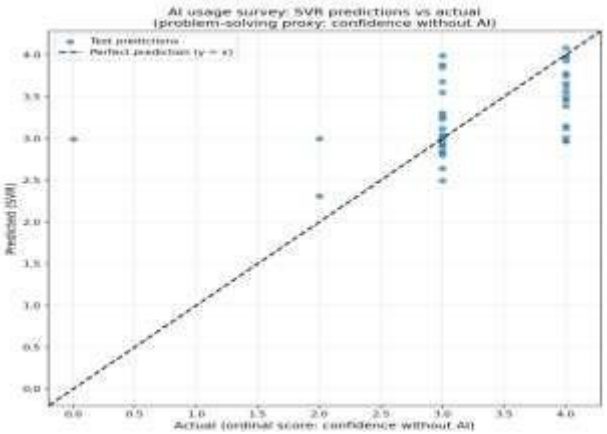


Figure 6: SVR Model Prediction

The Figure 6 shows how well an **SVR (Support Vector Regression)** model predicts students' confidence in problem-solving without AI compared to the actual survey values. Each blue dot represents a data point where the x-axis is the **actual confidence score** and the y-axis is the **predicted score** from the model. The dashed diagonal line (y = x) represents **perfect prediction**, meaning if a point lies exactly on this line, the model predicted the value accurately.

Most of the points cluster around actual values of 3 and 4, indicating that many students reported moderate to high confidence levels. While several predictions are very near to diagonal line (showing reasonable accuracy), many points are slightly above or below it, meaning the model sometimes **overestimates or underestimates** the confidence levels.

Conclusion

These technologies assist people to gain easy access to information and knowledge and aid efficient learning, but too much dependence on them could lead to adverse effects, such as deterioration of critical and problem-solving skills. As such, schools strive to find equilibrium between the usages of term called Artificial Intelligence and ensuring that students use AI technology not as a replacement of their critical and analytical skills. The incorporation of AI literacy into the curriculum will go a long way in assisting students to comprehend the strength and weaknesses of AI technology so that they can use the technology responsibly .In essence, it becomes imperative for AI technology to became assistance to human intelligence rather than being supplant it, thus promoting of students who think independently and can solve problems creatively.

There has been conducted the study having less sample space , which limits the possibilities of applying the results to the larger student audience. Another aspect of concern is the nature of data sources as they represent self-represented survey results that may contain a certain degree of bias. Namely, respondents could present themselves by providing the information in context of the actions that they are deem acceptable and not socially inappropriate. Also, one should pay attention to the fact that it lack the accuracy in the response data. The survey form of questionnaire being filled out electronically, and may be because of this fact, there might be students who may have misunderstood the questions, picked wrong answers, or simply made mistakes without checking them properly. Moreover, the assumption being made that respondents provided truthful answers and behaved appropriately while filling out the survey form. Nevertheless, this statement can be taken as questionable since might be a possibility of exaggerations, misrepresentations, and inconsistencies. Apart from that, the research has been designed following the cross-sectional approach. That means that the results are related to the

perceptions and behaviors observed during the research period and does not necessarily it is important to reflect upon the development of these aspects in the future. Last but not least, it should be stated that this study has not measured any direct effects on students' critical thinking skills and problem-solving performance. Instead, only students' perceptions and experiences have been analyzed.

References

1. Asghr, Kainat, Uzma Munawar, and Naeem Ullah. "Examining the Effect of Ai Tools on Critical Thinking and Problem Solving Skills Among Higher Education Students." *Gomal University Journal of Research* 41.1 (2025): 105-114.
2. Aprianto, Ide, Sophia Rahmawati, and Susanti Sufyadi. "Artificial intelligence and its effects on critical thinking and problem-solving abilities in higher education." *Indonesian Journal of Educational Development (IJED)* 6.3 (2025): 704-719.
3. Hassen, Mohammed Zeinu. "The impact of AI on students' reading, critical thinking, and problem-solving skills." *American Journal of Education and Information Technology* 9.2 (2025): 82-90.
4. Arteaga Lucas, Kassandra Elizabeth, and Fabricio Rolando Rivadeneira Zambrano. "Artificial intelligence in the critical thinking of students of technological careers." *Mendive-Revista de Educacion* 23.3 (2025).
5. Lin, Haoming, Wei Wei, and Handan Lu. "Facilitator or barrier? A systematic review on the relationship between Artificial Intelligence technologies and the development of critical thinking skills." *Journal of Educational Technology and Innovation* 7.2 (2025).

 ISBN: 978-81-685854-0-9

EXPLAINABLE AI-BASED INTRUSION DETECTION SYSTEM FOR NETWORK TRAFFIC CLASSIFICATION

Mr. Harsh Kathe
Department of Information Technology and Computer Science, Somaiya School of Basic and Applied Sciences, Somaiya Vidyavihar University, Mumbai, India.
Email: harsh.kathe@somaiya.edu

Mr. Bilal Shaikh
Department of Information Technology and Computer Science, Somaiya School of Basic and Applied Sciences, Somaiya Vidyavihar University, Mumbai, India.
Email: mohammadbilal@somaiya.edu

Abstract
Network intrusion detection has grown increasingly critical as cyber threats evolve beyond what static, signature-based systems can handle. Machine learning classifiers offer improved detection capability, but their black-box nature limits analyst trust and operational adoption. This paper presents XAI-IDS, an explainable AI-based intrusion detection system that combines a Random Forest classifier with SHAP (SHapley Additive exPlanations) to deliver transparent, multi-class network traffic classification. Trained and evaluated on the UNSW-NB15 benchmark, the system achieves approximately 93% binary detection accuracy and 76% accuracy across ten attack categories including DoS, Exploits, Backdoor, Reconnaissance, and Shellcode. A four-level severity mapping (Low, Medium, High, Critical) supports structured incident prioritization, and the complete system is deployed as an interactive SOC dashboard built with Streamlit, supporting Wireshark CSV ingestion, confidence-based filtering, and one-click incident report generation. Feature-level SHAP explanations accompany every classification decision, enabling analysts to validate model outputs and build justified trust in automated detection.
Keywords: Intrusion Detection, Explainable AI, SHAP, Random Forest, UNSW-NB15, Network Security, Machine Learning, SOC Dashboard.

1. Introduction

Modern organizations face an escalating challenge in defending networked systems against adversaries who continuously devise novel attack strategies. Conventional rule-based and signature-driven detection tools are structurally incapable of recognizing threats they have never encountered — a vulnerability that attackers routinely exploit through zero-day exploits and advanced persistent threats. Machine learning has emerged as a compelling alternative, enabling detection based on statistical patterns rather than fixed signatures. However, the prevailing research focus on accuracy has come at the expense of transparency: most ML-based IDS function as black boxes, offering binary verdicts with no feature-level justification, which makes alert validation difficult for analysts [7].

Three compounding challenges define the current state of the field. First, signature-based systems cannot adapt to novel threats without manual updates that lag behind real-world attack evolution. Second, ML classifiers trained on benchmark datasets rarely provide explanations, limiting operational trust and hindering triage decisions. Third, most research prototypes lack the infrastructure needed for genuine SOC deployment multi-class categorization, severity prioritization, interactive dashboards, and support for standard packet capture formats. This

work addresses all three by combining Random Forest classification with SHAP-based explainability in a production-ready Streamlit SOC dashboard evaluated on the UNSW-NB15 benchmark dataset.

1.1 Research Objectives

- Train and evaluate a Random Forest classifier on UNSW-NB15 for binary and 10-class multi-category network traffic classification.
- Integrate SHAP to produce feature-level, per-prediction explanations that address the black-box limitation of existing ML-based IDS.
- Develop a four-level severity mapping (Low, Medium, High, Critical) enabling structured, priority-driven incident response.
- Deploy the complete system as an interactive Streamlit SOC dashboard supporting real-time CSV analysis, Wireshark ingestion, confidence filtering, and downloadable incident reports.

2. Literature Review

The IDS research landscape has shifted from signature-based approaches toward ML-driven and hybrid architectures, motivated by the fundamental limitations of static detection. Ahmad et al. [1] demonstrated that Random Forest consistently outperforms single classifiers such as SVM and k-NN on UNSW-NB15, with ensemble averaging reducing overfitting while providing native feature importance scores. Panigrahi and Borah [10] showed that combining Random Forest with feature selection achieves near-perfect binary accuracy on NSL-KDD, confirming that feature quality often matters more than model architecture. Kasongo and Sun [6] corroborated this on UNSW-NB15 using wrapper-based feature selection with DNNs. Vinayakumar et al. [18] compared CNN, LSTM, and hybrid deep architectures and found that while deep models achieve competitive accuracy, they demand far greater computational resources and offer substantially less interpretability than ensemble methods. Ferrag et al. [5] concluded in a broad survey that no single deep architecture dominates across all threat types and that interpretability remains unresolved.

Lundberg and Lee [7] introduced SHAP, grounded in cooperative game theory Shapley values, providing both global feature rankings and local per-prediction explanations — making it particularly valuable for security applications where analysts require specific, actionable justification for each alert. Ribeiro et al. [13] proposed LIME as an alternative local explainability method, though SHAP has since demonstrated superior theoretical consistency. Recent work by Alhaidari and Al-Dossari [2] confirmed significantly higher analyst trust in systems that provide feature-level rationale for flagged traffic. The UNSW-NB15 dataset [9], generated using the IXIA PerfectStorm tool at the Australian Centre for Cyber Security, has become the standard IDS benchmark with approximately 2.5 million records spanning nine real-world attack categories and 34 engineered features covering flow statistics, protocol characteristics, and behavioural indicators.

2.1 Limitations of Existing Systems

Six persistent shortcomings across both signature-based and ML-based IDS motivated this work: (1) signature systems cannot detect zero-day threats without prior knowledge; (2) anomaly-based systems produce excessive false alarms causing alert fatigue; (3) most ML classifiers operate as black boxes with no feature-level explanation; (4) binary-only classification fails to support targeted incident response; (5) research prototypes lack SOC-ready dashboards and report generation; and (6) few existing systems support direct Wireshark packet capture ingestion.

 ISBN: 978-81-685854-0-9

3. Methodology

3.1 Research Design and Dataset

This study employs a quantitative experimental design, evaluating the proposed XAI-IDS against four metrics — Accuracy, Precision, Recall, and F1-Score — on binary and multi-class classification tasks. The UNSW-NB15 dataset [9] was selected for its modern attack coverage and widespread adoption in contemporary IDS research. A stratified 80/20 train-test split ensures representative class distribution across both partitions. Table 1 summarizes the attack categories, assigned severity levels, and incident priorities used throughout the system.

Table 1: UNSW-NB15 Attack Categories, Severity Levels, and System Comparison

Attack Category	Severity	Priority	Detection : Snort	Detection: ML-IDS	Detection: XAI-IDS
Normal	Low	P4 – Monitor	✓	✓	✓
Reconnaissance	Medium	P3 – Investigate	Partial	Partial	✓
Fuzzers	Medium	P3 – Investigate	✗	Partial	✓
Analysis	Medium	P3 – Investigate	✗	Partial	✓
DoS	High	P2 – Respond	Partial	✓	✓
Generic	High	P2 – Respond	✗	Partial	✓
Exploits	Critical	P1 – Immediate	Partial	✓	✓
Backdoor	Critical	P1 – Immediate	✗	Partial	✓
Shellcode	Critical	P1 – Immediate	✗	✗	✓
Worms	Critical	P1 – Immediate	Partial	Partial	✓

Source: Author's compilation based on UNSW-NB15 dataset [9] and literature [1][5][7]

Figure 1 shows the distribution of attack categories within the UNSW-NB15 dataset used for training and evaluation.

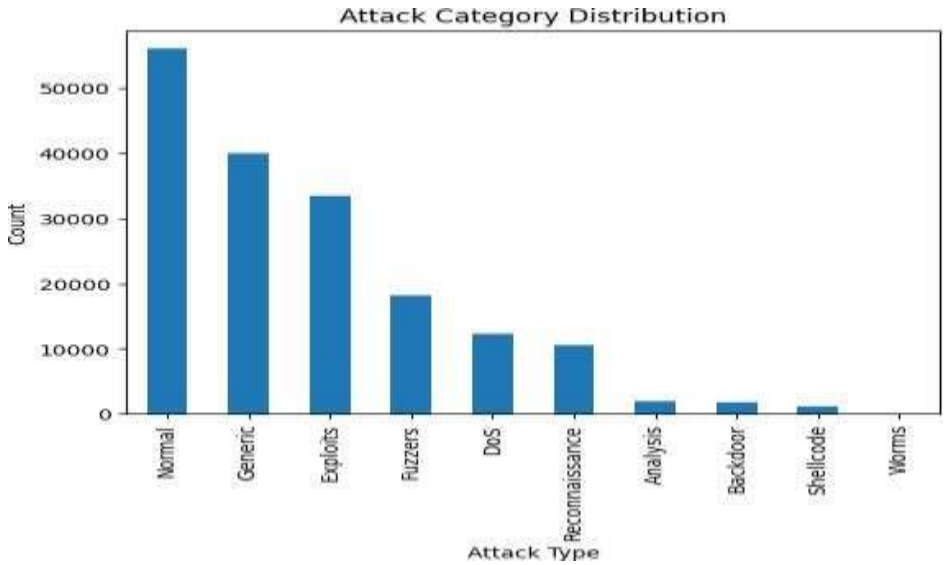


Figure 1: Distribution of attack categories in the UNSW-NB15 dataset.
Source: Author's analysis based on UNSW-NB15 dataset [9]

3.2 System Architecture

XAI-IDS implements a four-stage pipeline from CSV upload to explainable threat report. Figure 2 illustrates the complete architecture.

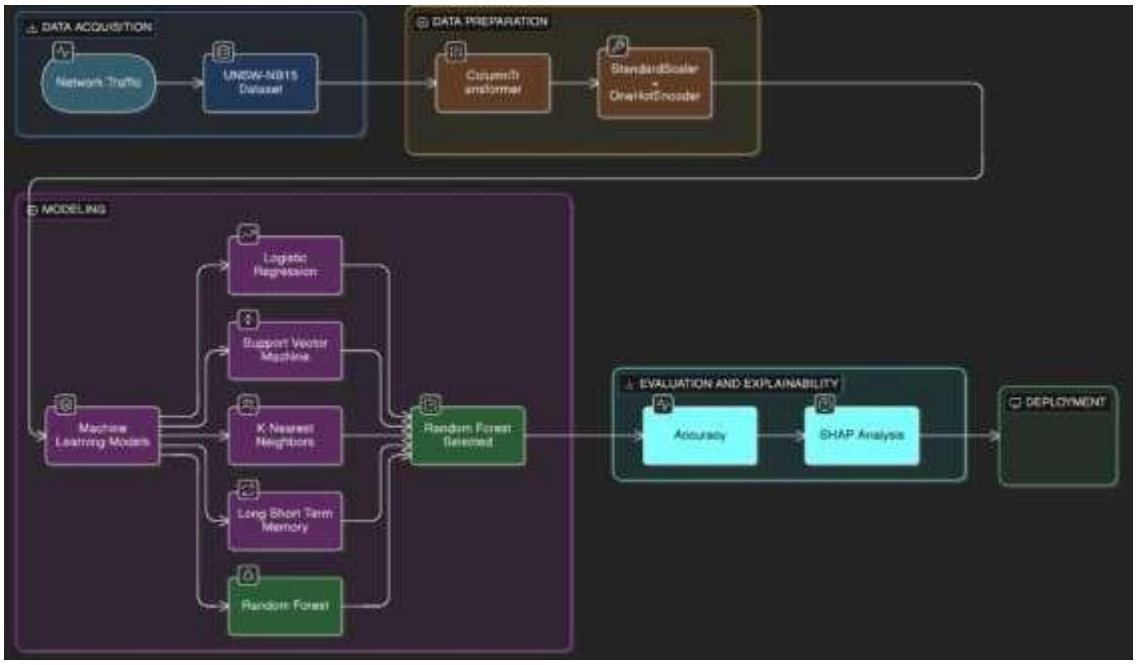


Figure 2: Architecture of the proposed XAI-IDS.

Source: Author

Stage 1 — Data Ingestion: The Streamlit dashboard accepts CSV uploads in two modes. Model CSV mode passes files directly to preprocessing. Wireshark CSV mode invokes preprocess_wireshark(), remapping raw column names (frame.len → sbytes, frame.time_delta → dur) and deriving behavioral features (dbytes, rate, spkts, dpkts) from available fields.

Stage 2 — Preprocessing and Feature Alignment: The align_features() function ensures uploaded data matches the model's 34-column feature space, zero-filling missing columns and dropping surplus ones. The fix_categorical_values() function maps unknown protocol or service categories to the nearest known training value, preventing pipeline errors on unseen traffic. StandardScaler normalizes all numerical features.

Stage 3 — ML Inference and Severity Mapping: The cached 421 MB Random Forest pipeline produces class labels via predict() and class-probability vectors via predict_proba(). A label encoder inverse-transforms numeric outputs to human-readable attack names. Confidence is computed as max(probabilities) × 100 per row. A deterministic severity dictionary assigns each prediction a severity level (Low/Medium/High/Critical) and incident priority (P1–P4).

Stage 4 — Visualization and Reporting: Four Streamlit pages render processed results. The Overview page displays threat metrics, a gauge chart, and attack distribution pie. Threat Analytics provides bar charts, severity distribution, and a Turbo-scale heatmap. Traffic Explorer offers a full filterable prediction table with one-click CSV export. Model Info shows classifier specifications and, when ground-truth labels are present, a confusion matrix.

3.3 Machine Learning Model: Random Forest

Random Forest was selected as the primary classifier based on its demonstrated superiority over SVM, DNN, and single-tree classifiers on UNSW-NB15 [1], its native feature importance scoring which directly supports SHAP integration, and its ensemble averaging which mitigates overfitting on class-imbalanced traffic datasets. The model is trained with n_estimators=300 and class_weight='balanced' to address the natural class imbalance in the UNSW-NB15 attack distribution. Preprocessing (OneHotEncoder for categorical features, StandardScaler for numerical) and classifier are encapsulated within a single sklearn Pipeline, ensuring consistent preprocessing across training and inference.

 ISBN: 978-81-685854-0-9

3.4 SHAP Explainability Integration

SHAP [7] is applied post-hoc on the trained Random Forest model to provide both global and local interpretability. At the global level, SHAP produces ranked feature importance across the entire test partition. At the local level, per-prediction waterfall plots show exactly which features pushed the model toward or away from a specific attack class. This per-flow explainability directly addresses the black-box limitation of conventional ML-based IDS — enabling analysts to see not just that a flow was classified as Backdoor, but which specific combination of behavioral features drove that decision.

4. Discussion and Findings

4.1 Classification Performance

The Random Forest model achieves approximately 93% binary detection accuracy (Normal vs. Attack) and 76% accuracy across the 10-class multi-category problem. Table 2 summarizes performance metrics for both classification settings.

Table 2: Random Forest Classification Performance on UNSW-NB15

Model	Accuracy	Precision	Recall	F1-Score	FPR
Random Forest (Binary)	~93%	~94%	~92%	~93%	~6%
Random Forest (Multi-class)	~76%	~74%	~73%	~73%	N/A

Source: Author

The 76% multi-class accuracy reflects the inherent difficulty of fine-grained attack type discrimination using tabular flow features alone — consistent with Ahmad et al. [1], who reported 72–78% for ensemble methods on this dataset. The model achieves highest precision on well-differentiated categories (Backdoor, Shellcode, Worms) and lower precision on structurally similar, underrepresented categories (Analysis, Fuzzers). Figure 3 shows the confusion matrix for binary classification.

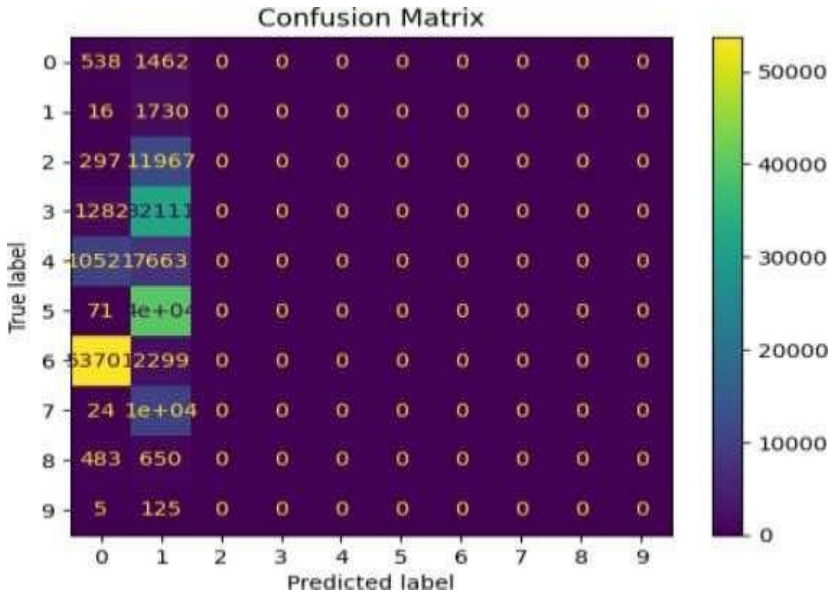


Figure 3: Confusion Matrix for Binary Classification Model.

Source: Author

4.2 SHAP Feature Importance

SHAP analysis identifies the five most influential features driving Random Forest classification

decisions. Table 3 presents these features with their mean absolute SHAP values and security interpretation.

Table 3: Top 5 Features by Mean Absolute SHAP Value

Rank	Feature	Mean |SHAP|	Security Interpretation
1	sbytes	0.241	Source byte volume; high values associate with DoS and data exfiltration
2	rate	0.198	Transfer rate; elevated packet rates are the primary DoS and Fuzzer indicator
3	state	0.176	Protocol state; incomplete/reset states characterize scanning and Backdoor
4	dur	0.154	Connection duration; persistent low-traffic connections indicate Backdoor
5	ct_src_dport_ltm	0.121	Port connection count over time; high values indicate Reconnaissance

Source: Author

4.3 SHAP Case Study: DoS Attack Detection

A network flow with rate=2,847 pkt/s, sbytes=1,203,841, and dur=0.42s was classified as DoS at 96.3% confidence. SHAP attributed the classification primarily to rate (+0.89) and sbytes (+0.72), with the short duration providing additional confirmation (−0.31 against benign). The analyst-facing explanation reads: "Extremely high transfer rate and source byte volume are the primary indicators of volumetric attack behavior." This feature-level justification enables the analyst to immediately understand and validate the classification without needing to inspect the model internals.

4.4 Functional Testing

All 12 functional test cases passed, confirming end-to-end pipeline reliability across CSV upload modes, edge cases including missing columns and unknown categorical values, confidence threshold filtering, confusion matrix rendering, and report generation. No pipeline failures or prediction errors were observed for any tested input configuration.

5. Conclusion and Future Scope

5.1 Conclusion

This work developed a production-ready Explainable AI-Based Intrusion Detection System that addresses three core limitations of existing approaches: inability to detect zero-day attacks, black-box opacity of ML classifiers, and absence of SOC-deployable operational infrastructure. By integrating a Random Forest ensemble classifier with SHAP-based explainability and a full-featured Streamlit SOC dashboard, XAI-IDS bridges the gap between academic ML research and operational cybersecurity deployment. The system achieves approximately 93% binary detection accuracy and 76% multi-class accuracy on UNSW-NB15, consistent with the state of the art for tabular-feature ensemble methods.

SHAP integration provides per-prediction, feature-level explanations that enable security analysts to validate and act on each alert with justified confidence — directly addressing the trust deficit that has limited ML-based IDS adoption in real-world SOC environments.

5.2 Future Scope

Future work will extend XAI-IDS across six dimensions: (1) deep learning integration using CNN, LSTM, or Transformer architectures to improve multi-class accuracy beyond 76% by exploiting temporal flow patterns; (2) live packet capture via Scapy or libpcap to enable continuous real-time monitoring; (3) federated learning for privacy-preserving distributed model training across network segments; (4) adversarial robustness evaluation against

 ISBN: 978-81-685854-0-9

gradient-based evasion attacks; (5) automated firewall response for Critical-severity detections; and (6) multi-dataset validation on CIC-IDS2018 and CAIDA to verify generalization across diverse network environments.

References

1. Ahmad, Z., Shahid Khan, A., Wai Shiang, C., Abdullah, J., & Ahmad, F. (2021). Network Intrusion Detection System: A Systematic Study of Machine Learning and Deep Learning Approaches. Transactions on Emerging Telecommunications Technologies, 32(1). https://doi.org/10.1002/ett.4150
2. Alhaidari, F. A., & Al-Dossari, A. M. (2023). Explainable machine learning for phishing detection. Applied Sciences, 13(4), Article 2486. https://doi.org/10.3390/app13042486
3. Al-Qatf, M., Lasheng, Y., Al-Habib, M., & Al-Sabahi, K. (2018). Deep Learning Approach Combining Sparse Autoencoder with SVM for Network Intrusion Detection. IEEE Access, 6, 52843–52856. https://doi.org/10.1109/ACCESS.2018.2869577
4. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324
5. Ferrag, M. A., Maglaras, L., Moschoyiannis, S., & Janicke, H. (2020). Deep Learning for Cyber Security Intrusion Detection. Journal of Information Security and Applications, 50, 102419. https://doi.org/10.1016/j.jisa.2019.102419
6. Kasongo, S. M., & Sun, Y. (2020). A Deep Learning Method with Wrapper-Based Feature Extraction for Wireless IDS.
Computers & Security, 92, 101752. https://doi.org/10.1016/j.cose.2020.101752
7. Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems, 30 (NeurIPS 2017).
8. Mia, M., Derakhshan, D., & Pritom, M. M. A. (2024). Can features for phishing URL detection be trusted across diverse datasets? (arXiv:2411.09813). arXiv. https://arxiv.org/abs/2411.09813
9. Moustafa, N., & Slay, J. (2015). UNSW-NB15: A Comprehensive Data Set for Network Intrusion Detection Systems.
MilCIS, IEEE. https://doi.org/10.1109/MilCIS.2015.7348942
10. Panigrahi, R., & Borah, S. (2018). A Detailed Analysis of CICIDS2017 Dataset for Designing IDS. International Journal of Engineering & Technology, 7(3.24), 479–482.
11. Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
12. Plotly Technologies Inc. (2024). Plotly: Interactive Graphing Library for Python. https://plotly.com/python/
13. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. ACM SIGKDD. https://doi.org/10.1145/2939672.2939778
14. Sarhan, M., Layeghy, S., Moustafa, N., & Portmann, M. (2021). Towards a Standard Feature Set for Network IDS Datasets. Mobile Networks and Applications, 27, 357–370. https://doi.org/10.1007/s11036-021-01843-0
15. Scarfone, K., & Mell, P. (2007). Guide to Intrusion Detection and Prevention Systems (IDPS). NIST Special Publication 800-94.
16. SHAP Documentation. (2024). SHAP: SHapley Additive exPlanations. https://shap.readthedocs.io/
17. Streamlit Inc. (2024). Streamlit: The Fastest Way to Build and Share Data Apps. https://streamlit.io/
18. Vinayakumar, R., Alazab, M., Soman, K. P., et al. (2019). Deep Learning Approach for Intelligent IDS. IEEE Access, 7, 41525–41550. https://doi.org/10.1109/ACCESS.2019.2895334

A REAL-TIME VOICEPRINT AUTHENTICATION SYSTEM WITH DEEP LEARNING-BASED SPOOF DETECTION FOR SECURE VAULT ACCESS

Aayushi Patel
MSC.IT Student, Department of Information Technology and Computer Science, Somaiya Vidyavihar University, Maharashtra.
Email: aayushipatel005@gmail.com

Pradnya Ahire
Assistant Professor, Department of Information Technology and Computer Science, Somaiya Vidyavihar University, Maharashtra.
Email: ppa@somaiya.edu

Abstract
As the volume of sensitive digital information grows, there has been increasing recognition that traditional password-based authentication is fundamentally flawed. Knowledge-based systems are susceptible to phishing, brute forcing, replay attacks and credential theft. Criminals use various impersonation tactics to bypass the automatic speaker verification (ASV) systems e.g., replaying synthesized voice, speech synthesis, deep-fake voice and voice conversion. In this paper, a deep learning-based spoof detection method is integrated with a vault access system based on voiceprint authentication. Using a USB microphone, the system records a 10-second live speech segment. It then performs signal pre-processing, which includes voice activity identification and noise reduction, and generates discriminative acoustic representations such as log: Mel spectrograms and MFCCs. An ECAPA-TDNN-based embedding model does speaker verification, and a convolutional neural network trained on the ASVspoof 2019 and 2021 datasets performs spoof detection. It does not require any expensive hardware and can work on a normal laptop with a microphone. This architecture is able to decrease false accept rates of replay and synthetic speech attacks in a verification system than that of others. It can also address a workable biometric security solution for health care, government, academia and access control on IoT scenarios to mitigate the new deepfake threats against authentication using voice.
Keywords: Speaker Verification, Voiceprint Authentication, Deep Learning, Encrypted Vault Access, Deepfake Prevention, Liveness Detection.

Introduction

Sensitive data is now stored and protected in diverse ways due to the rapid expansion of digital technology. Much of our personal, institutional, healthcare and government data is stored digitally now. Although this paradigm shifts towards convenience and greater efficiency, it compels us to formulate stronger mechanisms for authentication. Traditional security tools like passwords and PINs are still prevalent. However these knowledge-based systems are exposed to hacking, phishing, credential stuffing, brute-force attacks, and other types of social engineering attacks (Verizon, 2023).

Even though biometric authentication itself is not new, there appears to be a steady increase in biometrics being used because they typically do not rely on external credentials or devices (Jain et al., 2020). Modes of identification such as fingerprinting, iris scanning, facial recognition

 ISBN: 978-81-685854-0-9

and voice biometric enable better identity assurance with less reliance on memorized credentials. Of these, voice biometrics have drawn special attention due to its nature of contactless authentication and low hardware requirements while still capable of providing verification in a remote scenario. Deep neural network-based embedding models largely replace previous statistical models. Speaker-specific characterization from speech signals has proven to be extremely discriminative, for example, using x-vectors, (Snyder et al., 2020). which uses ECAPA-TDNN (Desplanques et al., 2020).These embeddings capture the speaker-specific attributes, which can be used to verify the identity of a speaker using similarity score approaches. Data such as large-scale datasets (Nagrani et al., 2020)., e.g., VoxCeleb, have also propelled strong model training and benchmarking.

A practical authentication system should meet three necessary demands: it has high verification accuracy, resists spoofing and deepfake attacks effectively, and runs in real-time with low hardware cost. Simultaneously achieving the three has been a difficult task in voice biometrics research.

To fill this gap, in this paper presents a dual-layer real-time voiceprint liveness authentication framework which combines deep speaker embedding verification and spoof detection based on deep learning.The system acquires a 10-second live voice sample, pre-processing and feature extraction is performed, speaker similarity is scored with ECAPA-TDNN embeddings and the spoof probability evaluated through CNN-based classification model trained on ASVspoof datasets. This encrypted vault is only unlocked when you meet both conditions for access verification. The architecture is conceived so that the full authentication process should not take more than 10 seconds and can be deployed practically.

Motivation

This research is motivated by the increasing abuse of generative AI as a tool for ill-intent. Voice cloning systems have become more available, making the activation cost of launching a passive attack easier. Voice-based Identity Verification systems are now being widely used in organizations for customer support, banking and remote access management. If not coupled with spoof detection, these systems could be susceptible for infiltration of critical infrastructure. Hence, an authentication system is required in which it verifies speech first and the other the speaker. This study intends to fill the gap between ideal/academic speaker recognition models and cybersecurity practice.

Contributions of this Paper

1. Real time voiceprint identification system that confirms in less than 10 second.
2. ECAPA-TDNN-based speaker embeddings for better identity discrimination and robustness.
3. Design and implementation of an ASVspoof 2019 and 2021 datasets of CNN classifier.
4. Another feature is that it also unlocks on dual-factor authentication first and then again on accessing the vault.
5. Experimental performance with enhanced spoof resistance and minimizing FAR, authentication latency is moderate below 8 seconds.

Related Work

The development of speaker validation has been greatly advanced in the last decade with deep learning-based embedding techniques. Gaussian Mixture Model–Universal Background Model approaches have largely been deprecated in favor of neural embedding systems, e.g., x-vectors (Snyder et al., 2020). These embeddings encode speaker-discriminative information in the speech signal.

More recently, ECAPA-TDNN was proposed to make additional advancements by including

channel attention and multi-layer feature aggregation to strengthen the speaker representation robustness (Desplanques et al., 2020). This architecture has shown to have better performance in text-independent speaker verification. But the expansion of speech synthesis and voice cloning technologies reveals weaknesses in automatic speaker verification. Speaker authentication systems are easy target to replay and synthetic speech presentations (Todisco et al., 2021), (Delgado et al., 2021).Research has also found that the deepfake speech can be deceptively close to real speakers, thus motivating spoofing detection in voice-based authentication (Kinnunen et al., 2022). More recent anti-spoofing works has been done on convolutional neural networks (CNNs) assuming spectrogram representation as input (Tak et al., 2021) while some of them have focused on system using raw waveform, such as RawNet2 (Jung et al., 2020).

Although the above techniques increase the performance of detecting spoofing attempts, in some implementations verification and deception detection are evaluated separately as single purpose systems not integrated within a secure access system which can be implemented as requested.The variation between the results of the current research and the earlier research is that the present research integrates deep speaker verification, convolutional neural networks (CNN) based spoof detection, and encrypted vault control to develop a complete real time authentication system.

Table 1: Indeed, most commercial voice biometric systems today are primarily targeted at authentication and some for fraud detection, but there is no system that fully integrates all three components of spoof detection, secure encrypted storage and real-time analytics. The system that we propose will overcome these limitations by means of a unified framework that incorporates all the different components together in one deployable system.

Feature/System	**Agnitio**	**PerSay**	**Nuance**	**Pindrop**	**Veridas**	**Proposed System**
Voice Authentication	✓	✓	✓	✓	✓	✓
Spoof Detection	Limited	×	✓	✓	✓	✓
Real time processing	✓	✓	✓	✓	✓	✓
User-level application	×	×	×	×	×	✓
Analytical Dashboard	×	×	Limited	×	Limited	✓
Encrypted Vault Access	×	×	×	×	×	✓
Logging and Monitoring	Limited	Limited	×	×	×	✓
Transparency	×	×	×	×	×	✓
Integrated Full System	×	×	×	×	×	✓

System Architecture

The entire system has five basic modules.

1. Recording of voices
2. Audio Preprocessing
3. Speaker Verification
4. Spoof Detection

 ISBN: 978-81-685854-0-9

5. Secure Vault Access Control

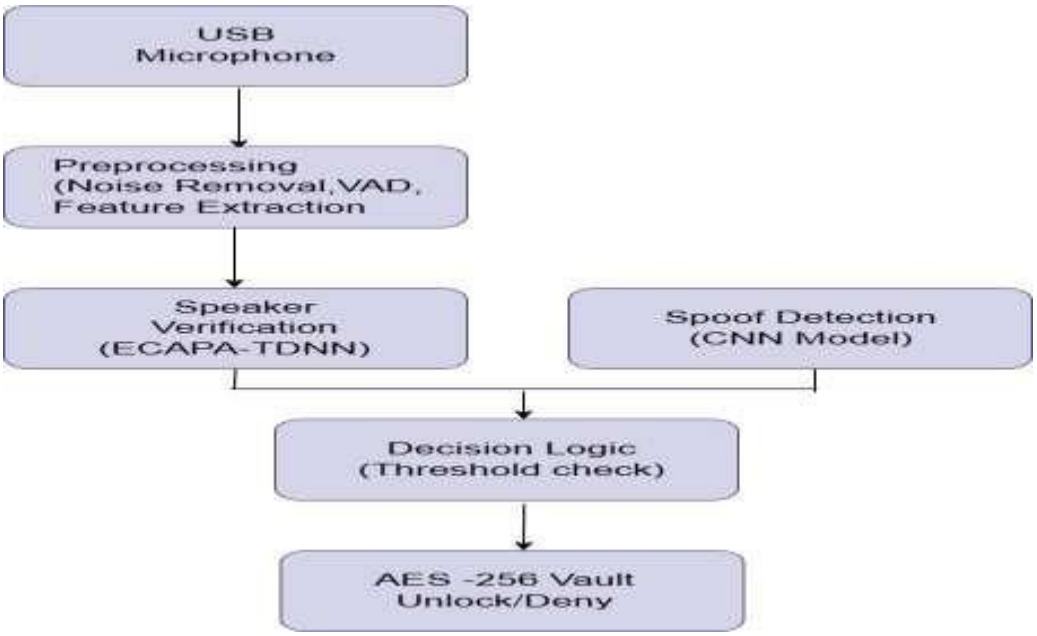


Figure 1

The above flow is the system architecture of the five modules.

Methodology

The authentication frame- work consists of two-layer verification (speaker identity and spoof detection) under real-time requirement. The approach consists of several steps: voice acquisition, preprocessing and feature extraction , generating the speaker embeddings based on other features that are recorded over time, determining if it is a spoof detection classification and decision logic for secure vault access.

Voice Acquisition

The enrollment of voice biometrics is initiated through the acquisition of a 10-second audio sample using an USB microphone. The recording window are time-fixed, so that the feature extraction and embedding generation remain same.It is also important that live capture be performed to defend against the file injection attacks of previously recorded files. As opposed to off-line verification systems, the present framework mandates live input capture to reinforce liveness assurance. This design choice is also motivated by practical limitations on biometric deployment pointed out in recent ASV evaluations (Todisco et al., 2021), (Delgado et al., 2021).

Preprocessing and Feature Extraction from Audio

Raw speech, signal is usually affected by background noise, silence segments and environmental effects. The following preprocessing steps are adopted to improve signal quality:

1. Silence removal using energy thresholding
2. Noise reduction
3. Voice activity Detection(VAD)
4. Signal Normalization

Voice Activity Detection helps in the identification of active speech regions by eliminating unnecessary acoustic patterns.Two representations of features are used in the feature extraction

phase:
1. Mel Frequency Cepstral Coefficients
2. Log Mel Spectrogram

Log Mel spectrograms can be used to extract the time-frequency patterns of the speech.These characteristics are broadly used in speaker verification and spoofing detection (Tak et al., 2021). Mel Frequency Cepstral Coefficients represent the features of the spectral domain in an efficient way that is similar to the human auditory system.The process of extracting features from the waveform signal is called feature extraction.

Speaker Verification using ECAPA-TDNN

In order to verify the person's identity by speaking, the system uses a technique called ECAPA-TDNN (Emphasized Channel Attention, Propagation, and Aggregation – Time Delay Neural Network).

The ECAPA-TDNN technique has been enhanced over the older TDNN technique in the following ways:
1. Channel attention
2. Multi-layer feature aggregation
3. Better separation of the speaker clues in the embeddings

The system generates fixed-length embeddings from the speech of any length. The embeddings represent the characteristics of the speaker without depending on the actual words being spoken (text-independent). The alignment linking the live speech embeddings and the stored embeddings can be found using the cosine similarity function:

$$S(x,y) = \frac{x.y}{||x||||y||}$$

x - live voice embeddings
y - enrolled speaker embeddings

The robustness of the system has been enhanced using advanced speaker embeddings like x-vectors and ECAPA-TDNN.

Using a CNN-Based Classifier to Find Spoofs

Speaker embedding models confirm identity, but they do not guarantee the authenticity of the presented voice signal. Contemporary generative models can produce highly realistic speech that can mislead standard verification systems (Wu et al., 2022), (Li et al., 2022). So, the designed structure involves another mechanism that recognises spoofing.

We train a Convolutional Neural Network (CNN) using log-Mel spectrograms from public anti-spoofing datasets. Log-Mel features retain time-frequency information and reveal spectral inconsistencies that commonly exist in synthesized speech and replay attacks. CNN-based classifiers have previously been used to detect spectral artefacts, e.g. phase artifacts in synthetic speech signals (Patino et al., 2021), (Wang et al., 2021).

The CNN model outputs a probability score when given an input spectrogram *X*:

$$P_{spoof} = f_{CNN}(X)$$

where shows how the neural network learned to map things. So this is what the classification decision:

$$Spoof = \begin{cases} 1, & if_{P_{spoof} > T2} \\ 0, & otherwise \end{cases}$$

where T2 is the threshold for spoof detection.

This random decision mechanism enables dynamic value alteration while balancing between

 ISBN: 978-81-685854-0-9

false reject rates and identify sensitivity.
Recent studies have shown that when used separately ASV systems could be weaker as compared to the embedding-based verification and specific spoof detection (Kinnunen et al., 2022).

Secure Access to Vault and Decision Logic
The final authentication decision is made by a dual-condition logical rule that combines the speaker similarity and spoof probability.

$$S(x,y) = \frac{x.y}{||x||||y||}$$

Below is an authentication function:

$$Accept = \begin{cases} 1, & if S(x,y) \geq^{T1} and Pspoof \leq^{T2} \\ 0, & otherwise \end{cases}$$

Compared to a single layer inspection system, this AND-condition significantly reduces the FAR (False Acceptance Rate) (International Organization for Standardization, 2020).
In the system, when authentication is confirmed, it uses the Advanced Encryption Standard to decrypt a protected vault. For safe data storage one of the most popular symmetric encryption standards remains AES-256 (National Institute of Standards and Technology, 2023).An illustration for the encryption function is as follows:

$$C = Ek(M)$$
$$M = Dk(C)$$

where:
1. M = original data
2. C = encrypted ciphertext
3. k = encryption key
4. Ek, Dk = encryption and decryption functions

Optimization
In real world systems of authentication, usability and security must co-exist. The suggested system is designed to keep authentication latency under 10 seconds.
Ttotal are as follows:

$$Ttotal = Tcapture + Tpreprocess + Tverify + Tspoof$$

where:
1. $Tcapture$ = duration of recording
2. $Tpreprocess$ = takes to extract the features
3. $Tverify$ = takes to create the embeddings and calculate their similarity.
4. $Tverify$ = CNN

inference time Among the techniques are:
1. Preloading the weights of the trained model into memory
2. Optimized inference has been shown to reduce response time while maintaining high detection accuracy in biometric frameworks (Ferrag et al., 2021).

Experimental Setup
This system was developed in Python and using deep learning frameworks such as PyTorch and SpeechBrain.The system focuses on real time authentication and spoof detection.
System Setup
The task was carried out in a desktop with the following specifications:
1. Intel i5 Processor

2. 16 GB RAM
3. USB microphone
4. CPU-based inference

Dataset and Model Configuration

The ASVspoof 2019 and ASVspoof 2021 datasets were used for training.Speaker verification was conducted with a ECAPA-TDNN pretrained embedding model trained on the VoxCeleb-corpus.The log-Mel spectrogram features based CNN model was used for spoof detection classifier. The threshold values of similarity score (τ_1) and spoof probability (τ_2) were chosen empirically to interchange between False Acceptance Rate (FAR) and False Rejection Rate (FRR).

Metrics for Evaluation

It is used to measure how well the system works:

1. Rate of False Acceptance (FAR)
2. Rate of False Rejection (FRR)
3. The Equal Error Rate (EER)
4. Latency in Authentication

Measured latency from the start to end to make sure it was done in less than 10 seconds.

Results and Discussion

First of all, after implementation of the voice authentication system, perform real-time tests to test whether it successfully authenticates the user to accept, reject and provide similarity scores. For all individual users, they first record a voice sample for 10 seconds, after enrollment the voice samples were collected and matched with previous stored voice profiles of users. After this, the system checks for similarity scores and gives the best match. It grants access when the similarity values of the particular user obtained above the threshold set.

Authentication Results

After various attempts, the differences between valid and invalid attempts were satisfactory. After correct user attemption, the access is granted and the similarity score was also high and for another user the similarity score was comparatively low and eventually access was denied. The results are as follows:

Table 2: The system performance was assessed by means of various authentication attempts from both legitimate and unauthorized users.Authentication Results:

Test Scenario	**Similarity Score**	**System Decision**
Genuine User Attempt - 1	0.78	Access Granted
Genuine User Attempt -2	0.69	Access Denied
Enrolled User - A	0.72	Access Granted
Enrolled User -B	0.41	Access Denied
Enrolled User -C	0.43	Access Denied
Genuine User (Successful Match)	0.79	Access Granted

 ISBN: 978-81-685854-0-9

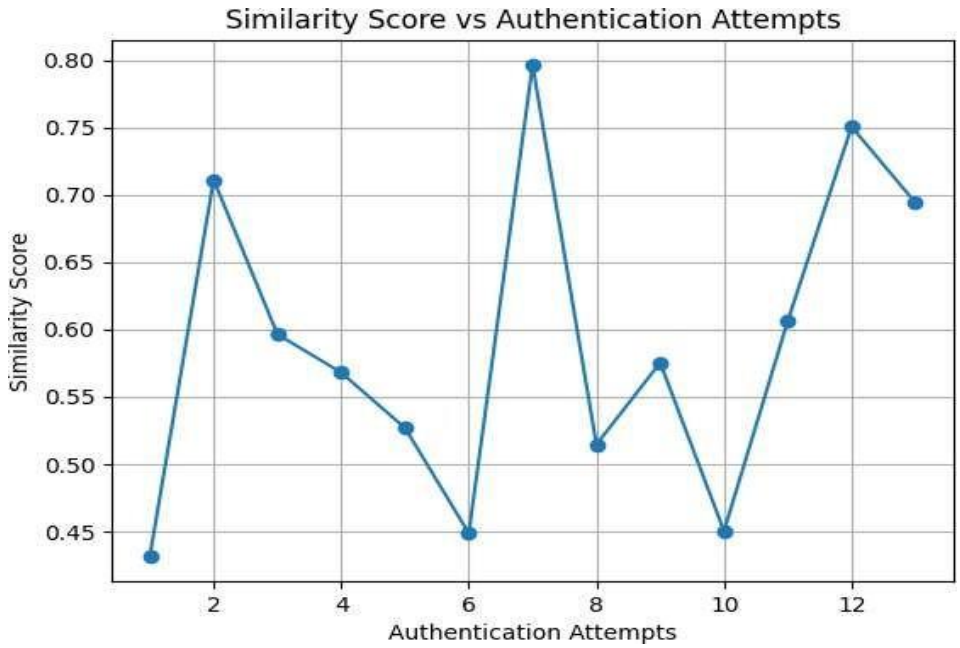


Figure 2

We also investigated the variation of similarity scores with different authentication attempts, as visualized in the above figure. This function almost always returns greater value than the defined threshold, causing successful identification and genuine users have a higher score. Unauthorized attempts, on the other hand, receive lower scores and are denied access. As slight deviations in data due to environment and recording conditions occurred, yet the system securely does not classify a valid user into invalid category and vice versa that speaks volume for robustness and reliable working of the model.

Discussion

From the attempts, it shows that it can easily discriminate between matched and unmatched users with their similarity scores. Also it detects false tries of the profiles and gives low similarity scores. With that it confirms that the voice features are of genuine users only.

Also after authenticating the same genuine user, the similarity score changes. This is because of the variation in tone,distance of the speaker from the microphone. Such patterns are expected within the system and it indicates the different similarity score for the same user.For security, even after the genuine user makes an effort but the threshold was just below it, the access was denied. Sometimes this approach becomes less user-friendly but at the same time it prevents the undesired users from getting access. This is often seen in biometric systems.

Lastly, the comparison of multi-user remains constant and satisfactory. For every authentication, the live voice sample was compared with all voices of user profiles and based on that it selects the correct match. The closest match from all the comparisons were rewarded as the voice of a particular person. Even when thousands of users were enrolled in the system, it provides precise identification.

Last but not least, the functionality of the system was showcased when it manages the vault access for a particular user. After the successful verification of the user it opens the vault folder and demonstrates how this system can be implemented in real world applications.

Overall Observation

After testing, both real-time voice authentication and practical access control marks success. Results gathered from the system can easily discriminate between matched and unmatched users. Such experiments can promise to use voice as an authentication factor when deployed in

applications that act as a good biometric system behavior.

Security Analysis

By combining speaker verification and spoof detection, our voice authentication system improves in terms of security. Other verification systems are vulnerable to synthetic and replay voice attacks. They can't discriminate between genuine or spoof speech signals (Todisco et al., 2021).

To solve this problem, the system uses a spoof detection algorithm called a CNN which prevents replay and other voice attacks (Kinnunen et al., 2022). When both the user similarity and signal genuineness are met, it can mitigate the risk of false acceptance.

While AES-256 encryption protects data. AES is a global standard used by many people to protect their data and maintain privacy (National Institute of Standards and Technology, 2023). This framework improves the resistance against spoofing attempts, also rectification of the set threshold values should be advised in order to preserve robustness.

Conclusion

This study proposed a real-time voiceprint authentication using the deep learning-based speaker verification and spoof detection suitable for secure vault unlocking. This new architecture makes it possible to mitigate the drawbacks inherent in traditional voice authentication systems using both an identity verification and a presentation attack detection component.

It has an advantage over traditional single-layer automatic speaker verification system, which only checks for speaker similarity before giving access. The use of ECAPA-TDNN embeddings ensures robust speaker divergence and a CNN-based spoof detection module increases to synthetic and deepfake attacks. The joint decision logic reduces the risk of false-acceptance and enhances authentication reliability.

The approach is engineered to be ready for online use as an authentication response latency of under 10s is achieved when using consumer hardware and independency from GPU based processing. Furthermore, the encryption and decryption using AES-256 prevents a compromise between layers to expose sensitive data as new layers cannot decrypt encrypted data unless authorization is granted.

It can be concluded that the developed framework, combining biometric verification system with anti-spoofing algorithm and encrypted access protection mechanism is practically usable and can be deployable in hospitals, IOT based access control scenarios,academic and government institutions.

In the future, we will work on enhancing the adversarial robustness of our models using transformer-based spoof detection architectures and facilitating their edge-device deployment for scalable real-world applications.

References

1. Verizon. (2023). 2023 data breach investigations report (DBIR). Verizon Enterprise.
2. Jain, A. K., Ross, A., & Nandakumar, K. (2020). Biometric recognition: Security and privacy. IEEE Access, 8, 824–838.
3. Snyder, D., Garcia-Romero, D., Povey, D., & Khudanpur, S. (2020). X-vectors: Robust DNN embeddings for speaker recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28, 1770–1783.
4. Desplanques, B., Thienpondt, J., & Demuynck, K. (2020). ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN-based speaker verification. In Proceedings of Interspeech 2020 (pp. 3830–3834).
5. Nagrani, A., Chung, J. S., & Zisserman, A. (2020). VoxCeleb: Large-scale speaker

 ISBN: 978-81-685854-0-9

verification in the wild. Computer Vision and Pattern Recognition (CVPR).
6. Todisco, M., Wang, X., Sahidullah, M., Yamagishi, J., & Evans, N. (2021). The ASVspoof 2019 challenge: Automatic speaker verification spoofing and countermeasures. IEEE Journal of Selected Topics in Signal Processing, 15(4), 1027–1041.
7. Delgado, H., Todisco, M., Sahidullah, M., Evans, N., Kinnunen, T., & Lee, K.-A. (2021). ASVspoof 2021: Automatic speaker verification spoofing and countermeasures challenge evaluation plan. In Proceedings of Interspeech 2021.
8. Kinnunen, T., Todisco, M., Wang, X., Sahidullah, M., & Evans, N. (2022). Deepfake speech detection: A review. IEEE Security & Privacy, 20(3), 74–82.
9. Tak, T., Patino, J., Todisco, M., & Evans, N. (2021). End-to-end anti-spoofing with convolutional neural networks. In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6369–6373).
10. Jung, J., Shim, H., Kim, J., Heo, H., & Yu, H. (2020). Improved RawNet2 for end-to-end anti-spoofing. In Proceedings of Interspeech 2020.
11. Wu, Y., et al. (2022). Deepfake detection for audio: A survey. IEEE Access.
12. Li, Q., et al. (2022). Adversarial attacks against speaker recognition systems. IEEE Transactions on Information Forensics and Security.
13. Patino, J., et al. (2021). CNN-based countermeasures for speech spoofing detection. IEEE.
14. Wang, X., et al. (2021). Detecting synthetic speech using time-frequency representations. IEEE/ACM Transactions on Audio, Speech, and Language Processing.
15. International Organization for Standardization. (2020). ISO/IEC 30107-3: Biometric presentation attack detection.
16. National Institute of Standards and Technology. (2023). Advanced encryption standard (AES) (FIPS PUB 197).
17. Ferrag, M. A., et al. (2021). Edge AI security for biometric systems. IEEE Communications Surveys & Tutorials.

DESIGN AND EVALUATION OF A RETRIEVAL AUGMENTED KNOWLEDGE REPOSITORY FOR MULTI-FORMAT DATA SOURCES

Chaitanya Shinde
Department of Information Technology, Somaiya School of Basic and Applied Sciences, Somaiya Vidyavihar University, Mumbai, India.
Email: chaitanya.s@somaiya.edu

Bilal Shaikh
Mentor, Department of Information Technology, Somaiya School of Basic and Applied Sciences, Somaiya Vidyavihar University, Mumbai, India.
Email: mohammadbilal@somaiya.edu

Abstract
Managing a large collection of digital information that is stored in multiple different file formats presents a significant challenge for efficient knowledge retrieval and storage. Traditionally retrieving text was done using simple keywords. It was a traditional approach to information retrieval, but it failed to take underlying context of the document into account and measure the semantic similarity of the text. It could not capture the context behind the data, especially when the data is in the form of diverse formats such as PDFs, spreadsheets, and image based files. A Retrieval Augmented Generation (RAG) systems help solve this limitation by combining semantic similarity retrieval together with a large language model's ability to reason and know context. This paper presents the design and evaluation of a RAG Knowledge Repository that is supports data in multiple formats. It performs document ingestion, semantic indexing, and supports conversation by saving chat history. The system processes different data file formats such as docx, pdf, txt, etc and also images by using Optical Character Recognition (OCR). The extracted content is split into chunks and embedded. These chunks are then stored in a vector database such as ChromaDB that enables semantic similarity search and retrieval. The system also incorporates intent based query routing to support flexible query handling across document and conversational workflows. System tests were then performed using 91 queries ranging in types and difficulty. Performance achieved: Precision@5 = 0.559, Recall@5 = 0.643, F1@5 = 0.583, MRR = 0.604, MAP = 0.593. From the conducted test of response quality evaluation with LLaMA 3.2:1b and DeepSeek-R1:1.5b, semantic similarity was better in DeepSeek than LLaMA (0.394 vs 0.3155), as well as exact match accuracy (0.385 vs 0.308). Response quality evaluation also displayed LLaMA performing better in response latency (18.73 seconds vs 24.48 seconds) than DeepSeek. All tests were run on a NVIDIA GTX 1050 Ti GPU which illustrated the feasibility of the system and the ability of it to perform on a common low end graphics card. The results demonstrate that the retrieval quality significantly influences the response accuracy of the models and confirm that lightweight language models can provide reliable performance while keeping the system easy to implement and economically cheaper to implement when supported by an efficient retrieval pipeline.
Keywords: Retrieval Augmented Generation, Semantic Search, Vector Databases, Knowledge Repository, Multi-Format Document Processing, Conversational Retrieval Systems.

 ISBN: 978-81-685854-0-9

1. Introduction:

Over the past decades, a huge growth in generation and storage of digital data within enterprise and academia circles occurred as result of the digital transformation of organizations. Enterprise and academics routinely work with massive amounts of information in different kinds of documents ranging from raw text documents, to structured data sheets, and scanned images or multi-media content every day.

Traditional keyword based retrieval systems rely on exact matching found between user queries and the document's content. Although these systems provide fast responses, they usually fail to capture the semantic relationships that exist between words and phrases. Hence a keyword based retrieval method may produce incomplete or irrelevant results if any query involves some synonyms, paraphrased expressions, or indirect references to other terms. These limitations become more highlighted when processing multi format documents, where structured and unstructured information has to be processed simultaneously.

Latest advances in natural language processing (NLP) and machine learning have made it possible to develop embedding-based retrieval systems that can encode high-dimensional vector spaces for textual information. Embedding retriever systems based on transformers produce numerical representations for documents and information requests based on meaning representations and contextual associations among various segments of texts. This semantic similarity search allows the system to return results based on the content, regardless of whether the document contains the actual key phrases from the query.

RAG (Retrieval-Augmented Generation) has proven as a successful framework that couples the capabilities of semantic retrieval with generative language models. Given a query, relevant contextual information is retrieved from external knowledge sources and passed as input to a language model, which helps augment the accuracy of the responses and reduce hallucination compared to the performance of standalone language models (Oche et al., 2025). These advantages notwithstanding, many current RAG implementations concentrate on text-based document collections and rarely offer adequate support for heterogeneous document formats, such as spreadsheets, scanned images, or mixed
format collections. For multi-format data, specialized parsing techniques and ingestion pipelines are needed alongside optimized indexing.

This paper therefore presents the design and implementation of a Retrieval-Augumented Knowledge Repository (RAK R) capable of processing heterogeneous formats of documents and enabling conversational mode of knowledge retrieval. The ingestion pipeline in the proposed system is scalable to transform heterogeneous document formats into a structured semantic form. The documents are parsed, normalized, and converted into embedding vectors that are stored in a vector database for similarity-based retrieval.

The system also employs secure authentication protocols based on an OAuth access control mechanism to enable users to store and retrieve personalized documents. Additionally, intent-based query routing is implemented between queries related to document retrieval, dataset actions, and conversational requests. Directing queries to relevant handling workflows enhances the system's efficiency.

To maintain multi-turn conversational interactions, the repository has a conversational memory feature that stores questions and answers. It uses this contextual understanding to generate responses for subsequent queries based on the previous responses (multi-turn) (Roberts, 2020). In addition, the evaluation framework is developed with a structured approach to evaluate its retrieval effectiveness, response quality, and latency of the overall system across multiple language model configuration cases.

The primary contributions of this research are summarized as follows:

1. Design and implementation of a scalable Retrieval Augmented Generated Knowledge Repository that supports multiple format documents.

2. Developing an intent based query routing system that efficiently handles document queries, dataset analysis, and normal conversations.
3. Integration of semantic embedding and vector storage system for context-aware retrieval.
4. Implementation of conversation-aware retrieval workflows using conversation history.
5. Development of a comprehensive evaluation framework using multiple performance and quality metrics.

2. Literature Review:

The area of intelligent knowledge retrieval system is greatly influenced by the progress made in information retrieval, machine learning, and natural language processing in last few decades. Traditional information retrieval system made uses of lexical matching methods that compares the users query with stored documents by using keyword-based ranking measures. Nevertheless, this model fails to retrieve any order of document.

Traditional Information Retrieval Techniques:

Early information retrieval system used the Boolean retrieval models. These models match documents with the users query by using the logical expressions containing the keywords. However, this model cannot retrieve any order in documents.

In order to improve the ranking, vector space model (VSM) was introduced which is used to represent documents and users query in a high dimension space. TF-IDF is a famous weighting scheme in VSM. TF-IDF calculates weight of a term according to the frequency of a term in a document compared with the number of occurrence of the term in all documents. We get higher performance using VSM with TF-IDF as compared to using Boolean models. However, the limitation with VSM is that it just uses the surface-level matching for matching user queries and documents and is not aware of the synonymy or paraphrase.

Traditional methods of lexical-based information retrieval perform poorly in heterogeneous data where the data includes both structured and unstructured information. This cannot achieve semantic similarity across textual data in a heterogeneous source which thus led to development of embedding based retrieval systems.

Semantic Embedding Models:

Earlier methods for text representation focused on word-level embeddings, where each word is converted into a low-dimensional vector, and similar words appear closer together in vector space. Techniques such as Word2Vec showed that distributed word representations could effectively capture relationships between words. Later approaches improved embeddings using global word co occurrence statistics, but these methods were still limited in capturing sentence-level meaning. Modern transformer-based models address this limitation by generating richer contextual representations.

In this work, Sentence Transformers are used to generate sentence-level embeddings quickly and efficiently, making them suitable for tasks such as sentence similarity and semantic search. These embeddings capture the overall meaning of complete sentences rather than individual words, making retrieval more accurate.

To compare embeddings, cosine similarity is commonly used. In retrieval systems, the query vector is compared with stored document vectors to find semantically similar results. This allows the system to retrieve relevant documents even when their wording differs from the original query, focusing on meaning rather than exact keywords.

Vector Databases for Semantic Retrieval:

As embedding-based retrieval systems became more common, specialized storage systems called vector databases were developed to manage large volumes of embeddings efficiently (M. Li et al., 2024). These databases store vector representations and enable fast similarity-based searches across large datasets.

Vector similarity search typically relies on nearest neighbor methods, but these can become

 ISBN: 978-81-685854-0-9

computationally expensive when working with millions of vectors. To address this, Approximate Nearest Neighbor (ANN) algorithms were introduced. One widely used method, Hierarchical Navigable Small World (HNSW), organizes vectors into hierarchical structures that allow faster searching and efficient traversal.(Peng et al., 2024)

Several modern vector database frameworks support semantic search, including FAISS, Milvus, Pinecone, and ChromaDB. These systems provide indexing, query optimization, and metadata filtering features. Metadata filtering is especially useful in real-world applications, as it allows searches to be narrowed based on details such as document name, source, or creation date.

In the proposed Retrieval-Augmented Knowledge Repository, vector storage is implemented using ChromaDB, which supports persistent storage and scalable similarity search. Documents are divided into smaller chunks, and embeddings for each chunk are stored along with metadata such as filename and chunk number. This structure helps maintain document order and ensures coherent context reconstruction during retrieval. Overall, using vector databases represents a shift from traditional indexing toward more advanced semantic retrieval systems.

Retrieval Augmented Generation (RAG) system:

RAG is a hybrid architecture that combines document retrieval with neural network text generation. Studies show that combining retrieval with sequence-to-sequence models improves factual accuracy and reduces hallucinations in knowledge-intensive tasks. (Oche et al., 2025)Traditional language models rely only on stored parameters, which can limit their accuracy for domain-specific topics and sometimes lead to incorrect or fabricated responses. RAG addresses this limitation by retrieving external knowledge before generating answers. A typical RAG pipeline includes four main stages:

1. Encoding query
2. Document retrieval process
3. Context construction
4. Response generation

In this method, user's questions are converted to vector embeddings, and document embeddings stored in the vector database are used for similarity comparison. Retrieve related documents and combine them into a context to be given to the language model for the final output. More recently there are many additional features which improves system such as re-ranking, source-aware retrieving, and context filtering. (Shao et al., 2023) Here, a standard RAG system which is an extension to retrieve multi-format documents and use for intent-based query to achieve the real world goal.

Optical Character Recognition:

OCR i.e Optical character recognition, is a technology that recognizes the text present in an image, thus making the image or scanned document indexed and readable for the machines. Tesseract is one of the well-known OCR engine which is developed initially by Hewlett-Packard and supports various languages and formats of images. OCR processing usually involves 4 steps:

1. Image preprocessing
2. Character segmentation
3. Feature extraction
4. Text recognition

The preprocessing step removes noise to provide quality image to subsequent steps and segmentation process is to segment the text from the background. Feature extraction extracts the feature of the characters and they are transformed into machine-readable text format. Integrate OCR with the ingestion pipeline makes it possible for scanned and image based documents be ingested, which greatly expands the number of supported documents.

Multi-Format Document Retrieval Systems:

In order to deal with various documents that exist in real world, a multi-format document retrieval system has to be developed. This type of system can store and search documents in structured, semi structured and unstructured formats in one single system. For example, CSV files and Excel files are structured files that consists of tables, JSON files are semi-structured files that contain a hierarchical structure, and PDF files and text files are unstructured files with plain text, headings and images within them that requires to be extracted.
These systems usually follow an Extract–Transform–Load (ETL) workflow:
1. Extract raw document content
2. Transform content into standardized text
3. Load processed data into indexing systems
By having this system, users no longer need to convert the documents themselves when uploading. The developed system uses a modular ingestion pipeline to handle the varied format of documents.

Conversation-Aware Retrieval Systems:

Conversation-aware retrieval system enhances multi-turn conversation by taking previous message history in consideration of retrieval. Unlike normal search system that will always process information from the first, models in conversation-aware retrieval system need to keep track of the dialogue context so that the system can deliver better and coherent response for the users. Follow up question can greatly benefit from context knowledge since the follow up question often referred to what the user asked before.
For example, to response to the second question 'and that is also part of...', the system has to be able to know what is 'that' refer to. The models can normally store the dialogue logs containing question by user and response by system, then reuse them to construct the context. In the developed system, dialogue log history is stored in the database and then used during the prompt construction stage.

Evaluation Metrics for Retrieval-Augmented Systems:

In recent years, research has highlighted the importance of measuring response faithfulness, relevance, and context relevance when evaluating retrieval-augmented systems. (Yu et al., 2024) Faithfulness measures how closely a generated answer depends on the retrieved documents. Relevance indicates how useful the generated response is for the user's query, while context relevance checks whether the provided context was actually used during answer generation. Response latency, or the time taken to generate an answer, is another important metric, especially for time-sensitive applications where faster responses improve user experience.
Retrieval performance is evaluated typically using standard metrics such as Precision@K and Recall@K. Precision@K is a metric that measures how many of the top results that are retrieved were relevant, while Recall@K measures how many of the relevant results were successfully retrieved among the top retrieved results. Mean Reciprocal Rank (MRR) evaluates how well relevant items are ranked, based on the position of the first correct result, with higher values indicating better ranking performance. Mean Average Precision (MAP) extends Precision@K by calculating the average precision across multiple queries.
For response generation quality, Exact Match (EM) and semantic similarity are commonly used. Exact match cases check whether the generated answer exactly matches the correct response, while semantic similarity measures how closely the generated answer matches the expected meaning, typically using sentence embeddings and cosine similarity. Overall, combining retrieval metrics, generation quality measures, and latency analysis provides a comprehensive way to evaluate system performance.

3. System Architecture

The RAGKR is envisioned to be a modular and scalable system that can ingest different data

 ISBN: 978-81-685854-0-9

types and perform conversational retrieval. Google OAuth based authentication plus session token based mechanism is the authentication scheme. The architecture adheres to the layered architecture model where user interface related components are segregated from the backend processing parts. This ensures maintainability, scalability and extensibility of the system. Each of the component modules perform specific tasks in the entire pipeline and transform and retrieve information efficiently. The architecture consists of five primary subsystems:

1. User Interface Layer
2. Authentication Layer
3. API and Application Layer
4. Document Processing Layer
5. Retrieval and Query Processing Layer
6. Data Storage Layer

These components interact through structured workflows that enable seamless execution of user queries and document ingestion operations.

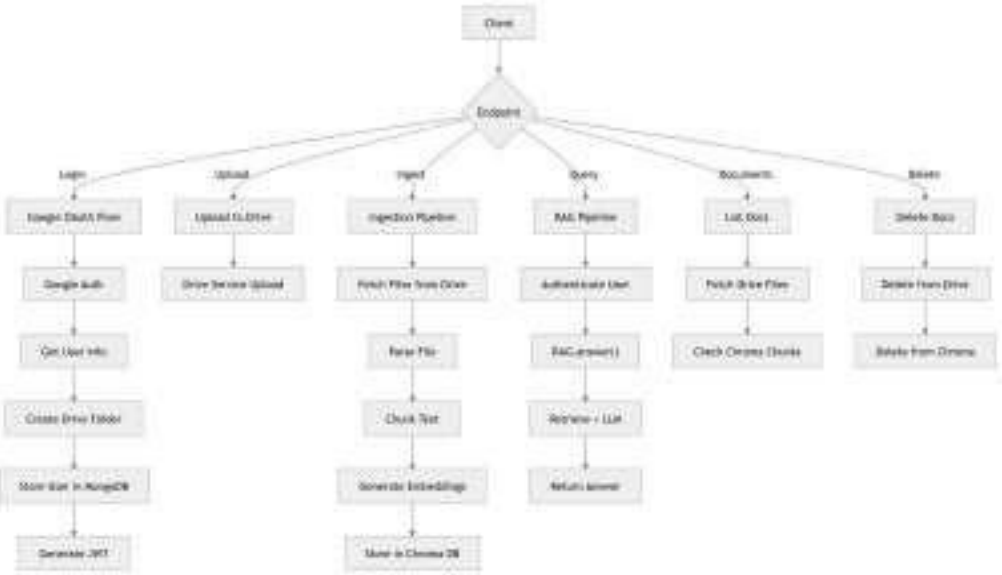


Figure 3.1 System-Level Architecture

Figure 3.1 shows the interaction among the different components of the system including the front-end interfaces, API end-points, authentication services, vector database and language models. Starting from the User Interface layer where the user interacts with the repository via web chat interface and allows document uploads, query submissions, and interact with the data stored, this front end part of the system is built as a component-based architecture allowing interaction with other services and dynamically interacting with the users. Asynchronous HTTP request are used to pass user queries and document uploads to backend API services.

User Interface Layer:

This layer allows the users to manage their documents, interact conversationally with the repository to retrieve knowledge, etc. The components of the user interface layer handle user inputs, displays system outputs, and manage user session interaction. The interface supports several primary operations:

1. User authentication
2. File upload
3. Query submission
4. Response display
5. Conversation management

This interface layer uses a chat interface to allow the users to enter their natural language queries and then returns the result fetched from the knowledge base using a retrieved document. Uploaded documents are displayed using a side panel, allowing users to track the documents uploaded to their personal repository. It enables users to talk to the repository using a conversational language, enhancing usability and accessibility.

API and Application Layer:

The API and application layer acts as the main communication bridge between the frontend interface and backend processing components. It is implemented using the FastAPI framework, chosen for its high performance, built-in request validation, and compatibility with modern Python machine learning tools.

This layer supports batch document processing, allowing multiple files to be handled together so embeddings can be generated in parallel. It also enables multi-source retrieval, allowing relevant chunks to be collected from multiple documents. The final prompt sent to the language model includes retrieved documents, the user's question, and conversation history. The major endpoints implemented in the system include:

1. GET /health — Verifies the system readiness and all components availability, gives a small overview to the user.
2. POST /upload — Uploads documents and starts the ingestion workflow
3. POST /query — Processes user queries using the RAG pipeline
4. GET /documents — Retrieves metadata for uploaded documents

These endpoints enable seamless communication between the user interface and backend services.

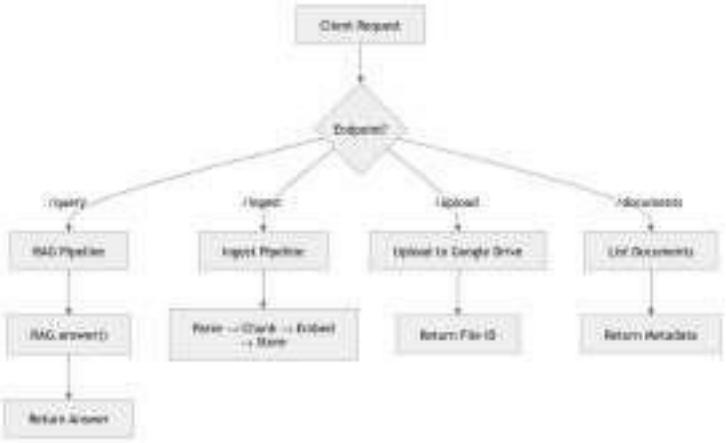


Figure 3.2 Endpoint-Level Architecture.

As shown in figure 3.2, these endpoints allow smooth communication between the user interface and backend services. Here each endpoint connects to specific backend modules responsible for ingestion, retrieval, and storage. The /ingest endpoint handles ingesting the documents where the documents are parsed, divided into chunks and then stored as embeddings. The /upload endpoint automatically saves uploaded files, processes their content, and creates searchable indexes, reducing manual effort. The /query endpoint handles natural language questions, retrieves relevant content, and maintains conversation history to ensure responses remain context-aware. And the /documents endpoint simply returns a list of documents in the repository.

Document Processing Layer:

The document processing layer is responsible for transforming raw uploaded documents into structured textual representations suitable for semantic indexing. This layer supports multiple document formats and integrates specialized parsing mechanisms for each format. This stage

 ISBN: 978-81-685854-0-9

represents the entire process of document ingestion. Supported files formats are, but not limited to, PDF, DOCX, TXT, MD, CSV, XLSX, PNG, JPG, TIFF. Different documents require different parsing approaches; for example, the system parses pdf files page by page and docx files paragraph by paragraph. Tabular documents like CSV and XLSX are converted to tabular text format while image files are converted using OCR.

The extracted text data is then broken into chunks by a chunking mechanism. This ensures each chunk does not exceed the context length of embedding models and also retains the semantic context of the chunk. Following chunking of document text by transformer based embedding models, these textual fragments are converted to vectors that represent their semantic content. Embedding models computationally intensive and so are efficient on the process.

Retrieval and Query Processing Layer:

The retrieval and query processing layer is the foundation of Retrieval-Augmented Generation pipeline where the relevant bits of the documents are retrieved and then a response based on the context is generated.

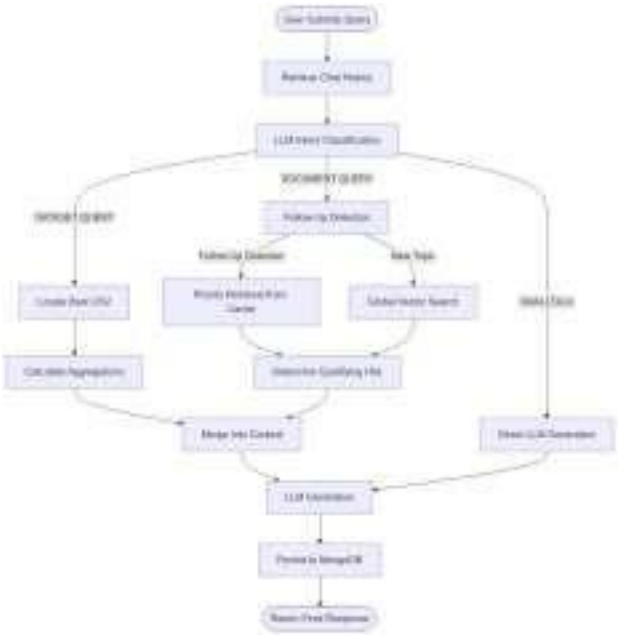


Figure 3.3 Query Processing Workflow

The figure 3.3 describes how a query is processed. The user query is converted to a vector representation using the same embedding model that the chunks were embedded in so that the query can be compared for similarity search against the chunk vectors. It is searched in the vector database and then the top 'k' documents that are similar to the query are retuned. It is 5 by default. Also here multi source retrieval takes place, means it gets pieces from many different documents and groups them according to its source document and sorts them by their position.

End-to-End Processing Workflow:

The complete system workflow integrates multiple modules to support document ingestion, storage, retrieval, and response generation. This workflow demonstrates how data flows across system components during typical operations.

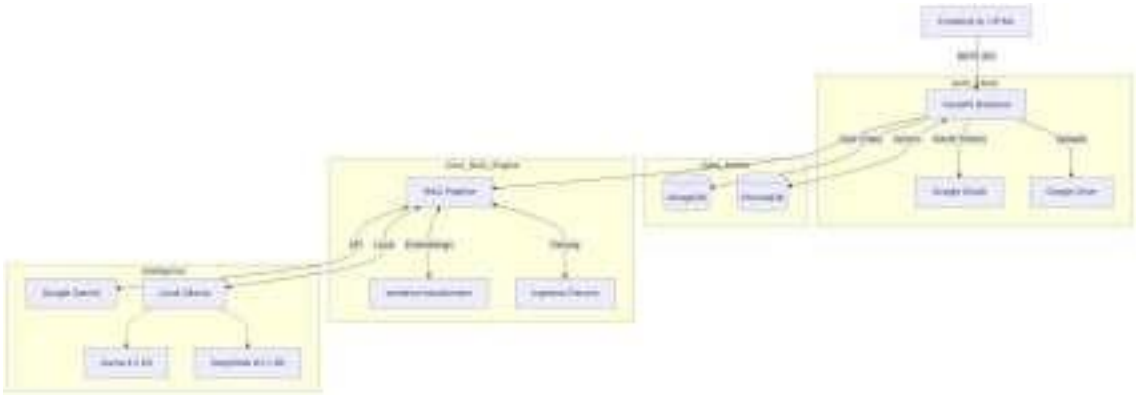


Figure 3.4 Complete End-to-End System Workflow

Figure 3.4 illustrates the full sequence of operations executed from user interaction to response generation.

4. Methodology

The methodology employed in the proposed Retrieval-Augmented Knowledge Repository outlines a series of operations designed to transform raw input data into semantic representations and subsequently into contextually relevant answers. Once uploaded the files will go through the following primary stages:

1. File Upload and Validation
2. Multi-Format Document Parsing
3. Text Chunking
4. Embedding Generation
5. Vector Storage and Indexing
6. Query Processing
7. Context Assembly
8. Response Generation

Each stage is designed to ensure accurate transformation of input data into meaningful semantic representations.

File Upload and Validation

The initial file upload step of the document ingestion process will be to ingest the file. Uploaded files are passed to the backend of the application through the web interface. Files are passed through validation steps to ensure that the files are in an acceptable format and have not exceed pre-determined size constraints (excessively sized files are a burden on resources). Upon validation, documents are saved to persistent storage for future use.

Multi-Format Document Parsing:

This stage converts uploaded documents into textual representations suitable for semantic processing.

 ISBN: 978-81-685854-0-9

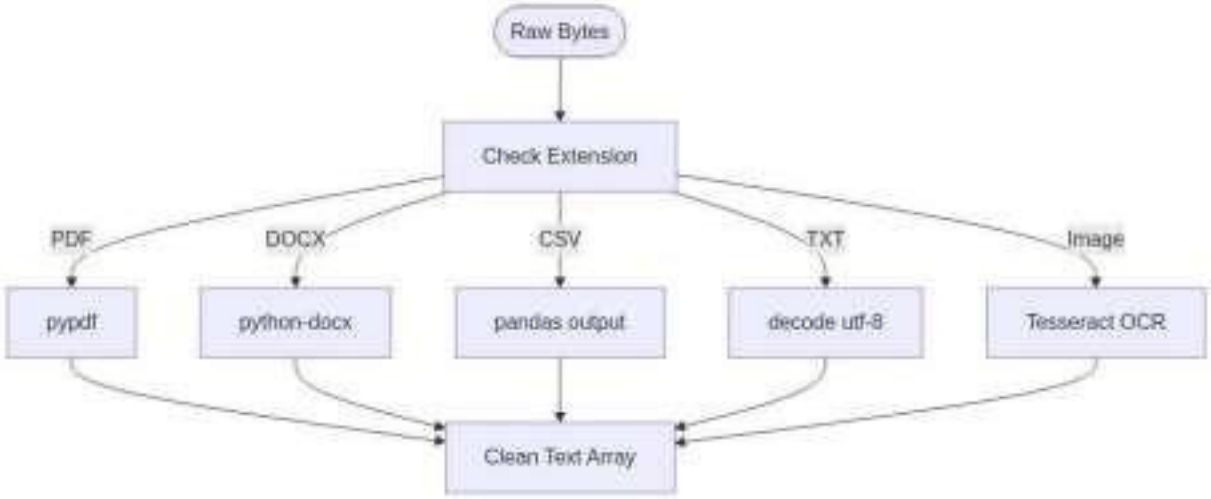


Figure 4.1 Multi-Format Parsing Workflow

Figure 4.1 illustrates the parsing mechanisms for different document types. Document type is identified through the file extension. Each document type is parsed using a specific method tailored to its format. For PDFs, text is extracted page by page for accuracy. For word processing documents, text is extracted after paragraph separation and then concatenated. CSV and Excel files are converted into text with row by-row extraction to preserve the tabular structure. Optical Character Recognition (OCR) is used for image documents to extract text. Normalized text is outputted by each parser to ensure consistency.

Text Chunking Strategy:

Chunking is an essential step in preparing document text for embedding, as Large Language Models and embedding systems have token limits. To process large documents that exceed these limits, it's necessary to break them down into smaller, coherent units. The following steps are involved:

1. Extract full document text
2. Normalize whitespace and formatting
3. Split text into sentences
4. Aggregate sentences into fixed-size chunks
5. Assign metadata to each chunk

Chunking based on sentence detection helps preserve meaningful information for embedding. Each chunk is assigned metadata such as document name, chunk index, and document ID. The chunk index is used to reorder chunks according to their original sequence during answer generation.

Embedding Generation:

Embedding generation converts the textual chunks into a numerical vector representation that capture the semantic meaning. These embeddings then enable similarity based search retrieval operations within the vector database. Transformer-based sentence embedding models are used to generate semantic representations. These models encode textual sequences into dense vector spaces, which places the semantically similar text segments closer together. The embedding generation workflow proceeds as follows:

1. Receive chunked textual data
2. Load embedding model
3. Encode text chunks into vector representations
4. Store vectors with metadata

Batch processing techniques are implemented to improve efficiency by processing multiple chunks simultaneously. This reduces processing overhead and improves throughput when

handling large documents. Embedding generation represents one of the most computationally intensive stages in the pipeline; however, its accuracy directly influences retrieval performance.

Vector Storage and Indexing:

This component holds the vector representations of the chunks of text and performs efficient similarity searches over them. It organizes embeddings in collections, with a separate collection of embeddings being generated per user for data isolation and increased security. Each vector consists of the following components: A unique ID for the vector, the text chunk that was embedded, the vector itself and the relevant metadata that represents information regarding the document that was embedded.

The metadata attributes are critical as they can help to reduce the number of documents to be searched, based on document sources or chunk index etc. A unique ID has been generated for each vector, after the required metadata has been added to it and after the a collection has been created for the user. The indexing workflow proceeds through the following steps:

1. Receive embedding vectors
2. Assign unique identifiers
3. Attach metadata attributes
4. Insert vectors into database collection

Similarity search operations use cosine similarity to compare query embeddings against stored vectors. Cosine similarity measures the angular distance between vectors and is widely used in semantic retrieval applications due to its computational efficiency

Query Processing:

The query processing stage aims at converting the user query into an embedding vector for retrieval. The system checks the user query to ensure it is in the correct format for processing and then it encodes the user query to a query embedding vector using the same sentence embedding model that was used for the document chunks. Once this vector has been generated it searches for the relevant document chunks using a similarity search algorithm. The query processing workflow includes the following steps:

1. Receive user query
2. Retrieve conversation history
3. Generate query embedding
4. Perform similarity search
5. Retrieve top-ranked document chunks

The inclusion of conversation history enables context-aware retrieval, improving response accuracy for follow-up queries. Similarity search operations return multiple relevant chunks ranked by similarity score. These chunks serve as the basis for context construction.

Context Assembly:

Context assembly organizes retrieved document segments into a structured format that the LLM can process effectively. Retrieved chunks are grouped by their source documents and sorted using chunk indices to maintain the original document flow. This ensures related information appears in the correct order and forms a clear, unified context.

The final context block includes all relevant text segments, source references, and orderly chunk structure. This logical order improves response accuracy and reduces hallucination during response generation. Context assembly stage acts as a bridge between retrieval and response generation.

Response Generation:

The response generation stage converts retrieved information into an augmented text answer. Once the input context is prepared, it is combined with the users query and conversation history to create a structured prompt. This prompt is then sent to the selected LLM to generate a response. The response generation workflow is as follows:

 ISBN: 978-81-685854-0-9

1. Construct structured prompt
2. Send prompt to language model
3. Generate response
4. Return response to user

Generated responses are stored with metadata that helps evaluate performance and maintain traceability. This stage transforms retrieved knowledge into clear conversational output, allowing users to interact naturally with stored information.

Intent-Based Query Routing:

Intent-based query routing improves efficiency by directing different types of queries to the correct processing workflow. Instead of treating every input as a document query, the system analyzes the user's prompt and assigns it to a specific intent category. This reduces unnecessary processing and
improves response speed. In the proposed system, user queries are categorized into three primary intent types:

- Document Query – Requests involving information from stored documents
- Dataset Query – Queries that involve operations on structured data such as CSV or Excel files
- General Conversational Query – Greetings or system-related interactions that do not require retrieval

Dataset Query Handling:

Dataset query handling expands the system beyond text retrieval by allowing users to work with structured data such as CSV and Excel files. Unlike text queries that rely on semantic retrieval, dataset queries focus on operations like filtering, aggregation, and statistical analysis. Dataset query processing includes the following steps:

1. Identify dataset source
2. Load structured data
3. Apply query operation
4. Generate output response

When a dataset query is detected, the system loads the dataset using structured data processing libraries. It then performs operations such as row filtering, column selection, statistical aggregation, and data summarization. The processed results are converted into clear textual summaries before being sent to the language model for explanation.

Performance Logging and Monitoring:

Performance logging is used to track system behavior and measure efficiency during execution. Logs capture key metrics at different stages of the pipeline, including query processing, retrieval, response generation, and system startup. Each query execution records the following information:

1. Query text
2. Retrieval time
3. Response generation time
4. Total processing latency
5. Model used
6. Retrieved document sources

These logs help identify performance bottlenecks and support system optimization. Monitoring latency is especially important in conversational systems, where slow responses affect user experience. Logging also allows administrators to compare model performance and understand trade-offs between speed and accuracy. In addition, recorded metrics support reproducibility of results and make system behavior more transparent, enabling continuous improvement over time.

5. Evaluation Framework

The evaluation framework is designed to measure both retrieval performance and response quality across different system configurations. Unlike traditional search systems, Retrieval-Augmented Generation architectures require multi-dimensional evaluation metrics due to their hybrid nature. The proposed evaluation framework measures three major performance categories:

1. Retrieval Performance
2. Response Quality
3. System Efficiency

Each category includes multiple quantitative metrics that collectively measure system effectiveness. Experimental Setup:

The experimental setup defines the environment used to evaluate system performance. Evaluation experiments were conducted using a structured dataset consisting of predefined queries designed to test document retrieval accuracy and response relevance. The evaluation configuration included the following parameters:

- Top-K Retrieval Value: 5
- Number of Test Queries: 91

Two language model configurations were evaluated:

1. Llama 3.2 (1B Instruct Q4_K_M)
2. Deepseek-r1:1.5b

Each model processed the same set of evaluation queries under identical retrieval conditions. The objective of this evaluation was to compare retrieval effectiveness and response quality across different language model configurations. (Chen et al., 2023) The evaluation dataset included queries designed to test: Document-based question answering, Contextual reasoning, Multi-source retrieval, Dataset interaction. These queries represent realistic user scenarios encountered in knowledge retrieval systems. All experiments were conducted using fixed retrieval parameters to ensure consistent comparison between models.

Retrieval Performance Metrics:

Retrieval performance metrics evaluate the relevance of retrieved document segments. These metrics measure how accurately the system identifies relevant content within the top-ranked results.

Precision@K measures the proportion of relevant documents retrieved among the top-K results. Mathematically, Precision@K is defined as:

Precision@K = (Number of Relevant Documents Retrieved in Top K) / (Total Documents Retrieved at K)

Higher Precision@K values indicate improved retrieval accuracy.

Recall@K measures the proportion of relevant documents retrieved from the total number of relevant documents available. Recall@K is defined as:

Recall@K = (Number of Relevant Documents Retrieved in Top K) / (Total Relevant Documents) Higher Recall@K values indicate improved coverage of relevant information.

The F1 score combines Precision and Recall into a single metric. It is defined as:

F1 Score = 2 × (Precision × Recall) / (Precision + Recall) This metric provides a balanced evaluation of retrieval performance.

Mean Reciprocal Rank (MRR) measures the ranking quality of retrieved results. It evaluates the position of the first relevant document within the retrieval list. MRR is defined as:

MRR = Average of (1 / Rank of First Relevant Document)

Higher MRR values indicate improved ranking effectiveness.

Mean Average Precision (MAP) measures overall retrieval quality across multiple queries. MAP is calculated as the average of precision values obtained at each relevant document position across all queries. This metric provides a comprehensive measure of ranking accuracy.

 ISBN: 978-81-685854-0-9

Semantic similarity measures the conceptual alignment between generated responses and reference answers. Instead of comparing responses using exact word matches, semantic similarity evaluates contextual meaning. This metric is computed using vector-based similarity techniques. Both generated responses and reference answers are encoded into embedding vectors using the same embedding model used during document retrieval. The cosine similarity between the vectors is then calculated. The cosine similarity formula is defined as:

Cosine Similarity = $(A \cdot B) / (||A|| \times ||B||)$

Where:

- A represents the generated response vector
- B represents the reference answer vector
- ||A|| and ||B|| represent vector magnitudes

Higher semantic similarity values indicate improved conceptual alignment between generated responses and expected answers.

Exact Match (EM) measures whether the generated response contains the exact expected answer string. This metric provides a strict evaluation of response correctness. Exact Match is defined as:

Exact Match = (Number of Exact Matches) / (Total Number of Queries)

Although strict, this metric is useful for evaluating factual correctness in structured queries.

Faithfulness measures the degree to which generated responses remain grounded in retrieved document content. A response is considered faithful if it accurately reflects the retrieved context without introducing unsupported information. Faithfulness evaluation typically involves assessing whether response statements can be traced back to retrieved sources. This metric is particularly important in RAG systems because it helps measure hallucination risk.

Relevance measures the usefulness of generated responses in addressing user queries. Unlike semantic similarity, which measures conceptual alignment, relevance evaluates practical utility. A response is considered relevant if it provides meaningful information directly related to the user's request.

Context relevance measures how well retrieved documents align with generated responses. This metric evaluates whether the retrieval process successfully provided useful supporting information. High context relevance indicates strong alignment between retrieval and generation stages.

Latency measures the time required to process a query and generate a response. This includes:

1. Query embedding time
2. Retrieval time
3. Prompt construction time
4. Response generation time

Average latency is calculated as:

Average Latency = (Sum of All Response Times) / (Number of Queries)

Lower latency values indicate improved system responsiveness.

Evaluation Procedure

The evaluation procedure defines the systematic process used to measure system performance across multiple metrics. The evaluation workflow consists of the following stages:

1. Prepare test query dataset
2. Execute queries sequentially
3. Retrieve relevant document chunks
4. Generate responses using selected model
5. Record retrieval and response metrics
6. Aggregate results across queries

Each query execution produces detailed logs containing retrieval scores, response outputs, and timing measurements. To ensure consistency, all evaluation experiments were conducted using

fixed configuration parameters. These parameters included a Top-K retrieval value of 5 and identical dataset conditions across model evaluations. The evaluation process was repeated for each model configuration using the same test dataset. This approach ensured fair comparison between model outputs. The results obtained from these experiments provide quantitative insights into system behavior under realistic usage scenarios.

All experimental measurements were conducted using locally hosted language models, including Llama and DeepSeek, executed on a system equipped with an NVIDIA GTX 1050 Ti GPU. In the context of modern hardware standards in 2026, the GTX 1050 Ti is considered a low-end graphics card. Performing experiments on such hardware demonstrates the system's ability to operate efficiently under limited computational resources and highlights its suitability for deployment in resource-constrained environments.

6. Findings

Results The performance of the system in terms of retrieval, response quality, and latency was quantitatively evaluated with the experimental results. Experiments were run on a fixed test set of 91 queries testing retrieval and generation performance of

1. Llama 3.2 (1B Instruct Q4KM)
2. Deepseek-r1:1.5b

Top-K retrieval value is set to 5 for both systems. All charts contain values that have been rounded off to the 2nd decimal place.

Retrieval Performance Analysis

Table 6.1 Retrieval Performance Metrics

[redacted]

F1 Score	0.583 ± 0.445

[redacted]

Mean Average Precision (MAP)	0.593 ± 0.461

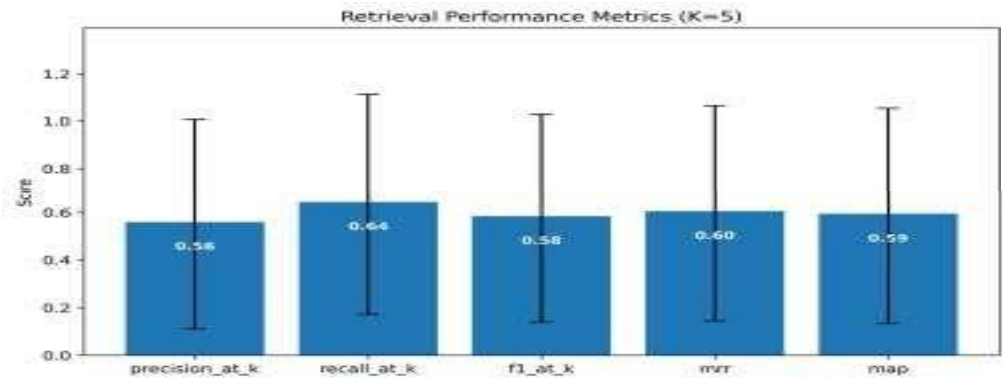


Figure 6.1 Retrieval Performance Metrics

 ISBN: 978-81-685854-0-9

These results demonstrate strong retrieval effectiveness across the test set of questions.
The Recall@5 value of 0.643 indicates that a large proportion of relevant document chunks were successfully retrieved within the top five results. Since high recall is an important parameter for RAG system, not missing relevant contexts lead to higher quality responses.
The Precision@5 value of 0.559 reflects moderate filtering accuracy, indicating that while most retrieved documents were relevant, there are some false positives which are expected since the retrieval is based on semantic similarity. This is typical behaviour.
The MRR of 0.604 and MAP of 0.593 also showed good performance as relevant document chunks appear at high ranked position of retrieval list. Thus, context retrieval is efficient to find necessary context information for downstream generation.
Overall, these results indicate that the vector database component effectively identifies relevant contextual information necessary for downstream language model generation.

Response Quality Analysis

Response quality evaluation revealed noticeable performance differences between the two evaluated language models.

Table 6.2 Answer quality metrics

[redacted]

Exact Match	0.308	0.385

[redacted]

Context Relevance	0.422	0.407

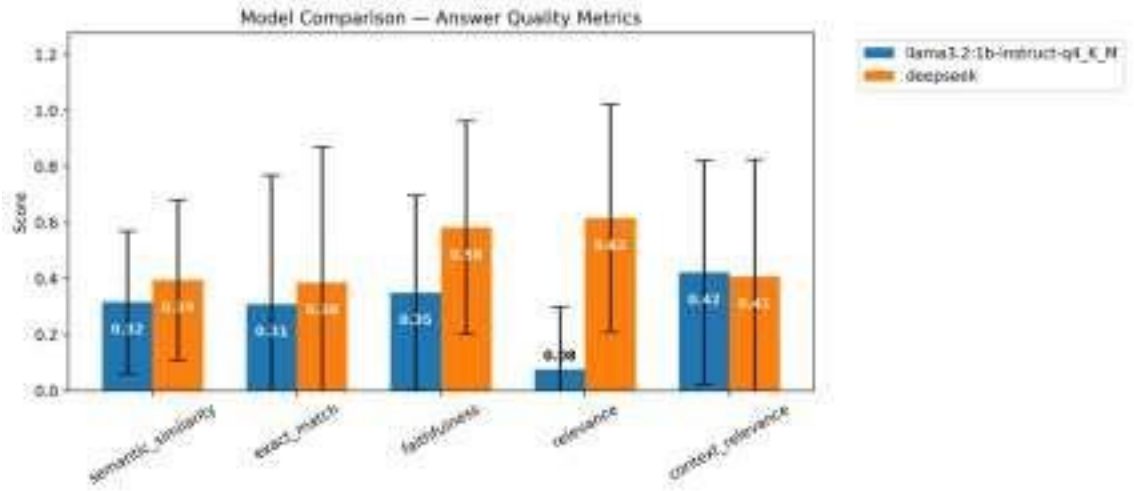


Figure 6.2 Answer quality metrics (rounded off to 2 decimal places)

The result showed that DeepSeek-R1 generate a more contextual relevant and correct response than LLaMA 3.2 on every evaluated aspect.

Semantic Similarity showed the highest score on DeepSeek (0.394) compared to LLaMA (0.316). The result of Semantic Similarity indicated that generated answers of DeepSeek can be much more closely related in concepts with the reference answers. Moreover, the Exact match of DeepSeek also gained the score 0.385 while LLaMA obtained the score of 0.308 which means that DeepSeek generate more correct answers.

Faithfulness score on DeepSeek is 0.582 and is the largest improvement compared to LLaMA (0.350). The result on Faithfulness shows that DeepSeek is much better grounded with the information from the retrieved context documents.

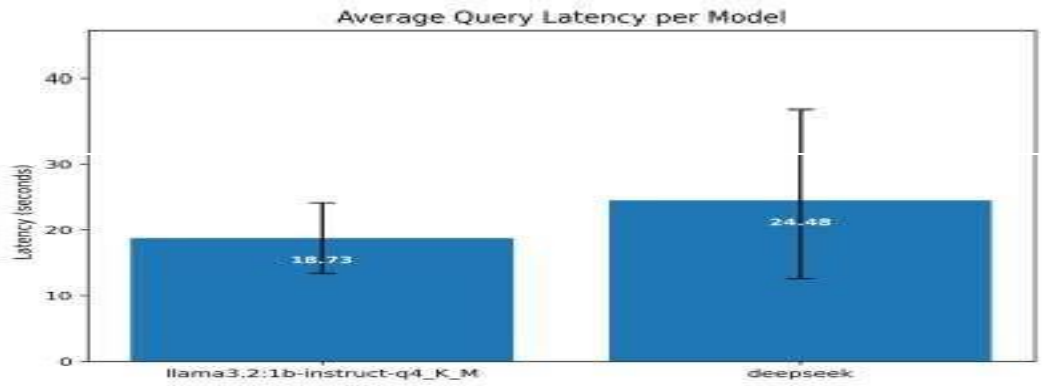
However, the results of Relevance provide the most compelling case where the response directly addresses the desired query goals and shown by a score of 0.615 by Deepseek as compared to Llama3.2's score of 0.076, with even further proof in context relevance score of its responses against the document retrieved.

On query type questions, both models performed poorly for Numeric queries and Negative queries which can also shows the limitation of light weighted large language models in performing complex reasoning tasks.

Latency Analysis

Table 6.3 Average query latency per model (in seconds)

Model	Average Latency (Seconds)
LLaMA 3.2	18.73 ± 05
DeepSeek-R1	24.48 ± 11.91



 ISBN: 978-81-685854-0-9

Figure 6.3 Average query latency per model (in seconds) (rounded off to 2 decimal places)

Latency analysis shows measurable differences between both models.
LLaMA 3.2 has the quicker response times and can be used in situations where latencies are more important, whereas DeepSeek-R1's slower response time is to do with higher computational intensity for it to provide a better quality response.
Despite providing better answer quality and context, Deepseek has longer latencies and would be a disadvantage when used in applications which require response being interactive, and in real time situations.
All evaluation experiments were conducted using locally hosted language models running on an NVIDIA GTX 1050 Ti GPU, which is considered a low-end graphics card in 2026. These results demonstrate that the proposed system can operate effectively on limited hardware resources, making it suitable for deployment in resource-constrained environments. But at the same time, with better high performance hardware these answer latencies can be reduced to a negligible amount.
Overall, the latency results show the clear trade off between response accuracy and computational efficiency, demonstrating the importance of balancing performance quality with system responsiveness.

Performance by Query Type
Performance analysis across query categories revealed distinct strengths and limitations.

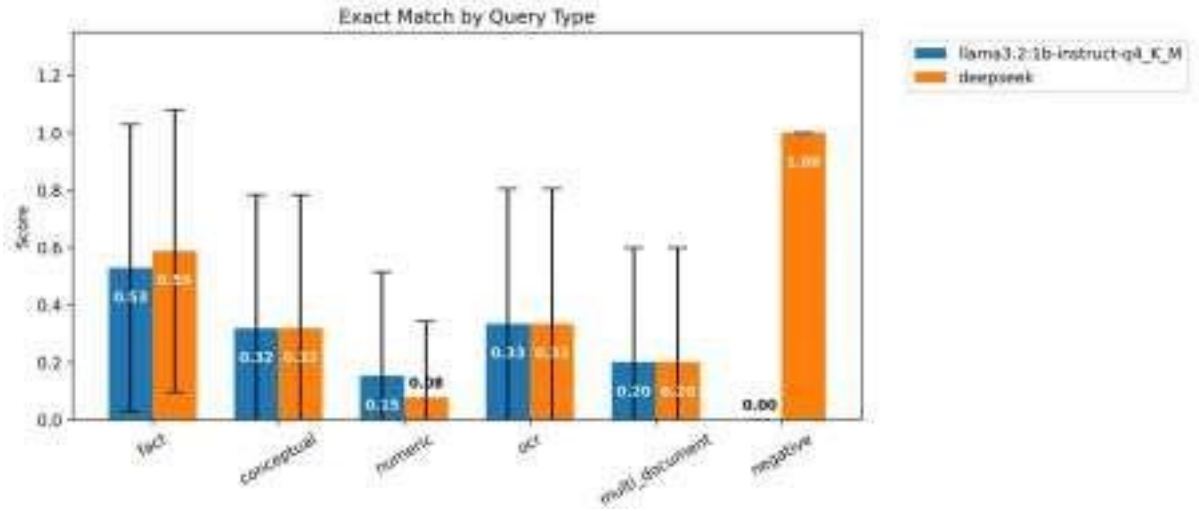


Figure 6.4.1 Exact match by Query type (rounded off to 2 decimal places)

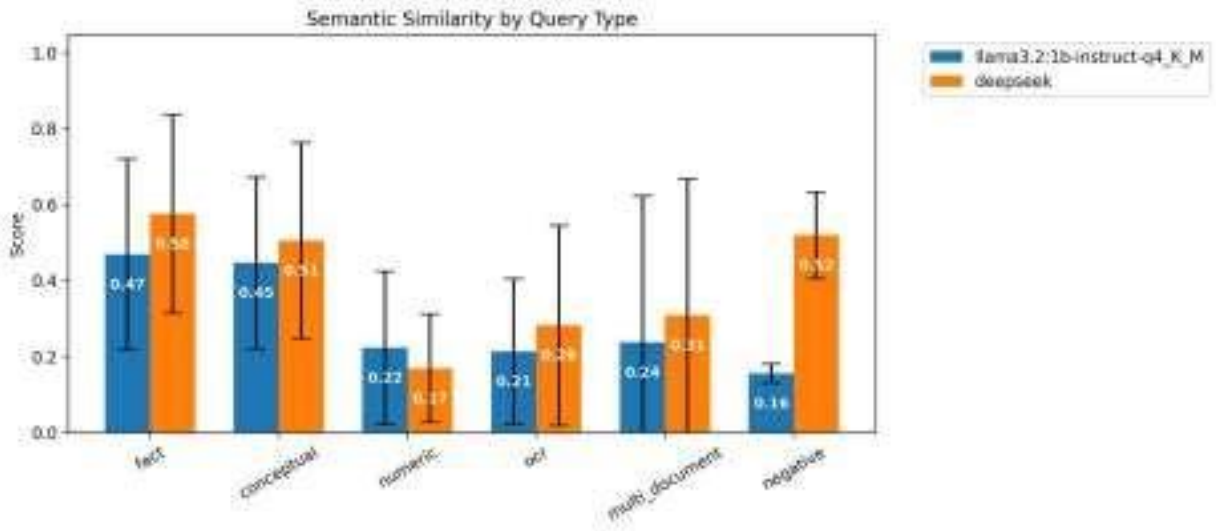


Figure 6.4.2 Semantic similarity by Query type (rounded off to 2 decimal places)

When you look at how different language models handle various types of queries, real differences in their reasoning abilities become visible. LLaMA 3.2 and DeepSeek-R1, for example, each handled specific query types in their own way, and their effectiveness really depended on how complex or structured the task was.

With fact-based questions, both models performed well. They performed well in presenting straightforward information. DeepSeek scored 0.588 for exact matches and 0.576 for semantic similarity. LLaMA's scores were a bit lower, with 0.529 for exact match and 0.469 for semantic similarity. So, when the task calls for factual retrieval, these Retrieval-Augmented Generation systems are at their most reliable.

On conceptual queries, both models still held their own, especially when it came to semantic understanding—DeepSeek reached 0.505, while LLaMA stood at 0.446. Exact matches were only moderate, showing that while both systems understood the concepts, wording their answers in a perfectly matching way was still tough.

Things got more challenging with OCR-based queries. Here, performance dropped mainly because pulling text from images introduces noise. DeepSeek managed an exact match score of 0.333 and a semantic similarity of 0.282. LLaMA lagged behind at 0.333 and 0.213. Still, the pipeline did manage to extract usable information, even if small OCR glitches kept response accuracy from being perfect.

Numeric queries proved toughest for both models. DeepSeek posted an exact match of 0.077 and semantic similarity of 0.169. While LLaMA achieved 0.154 exact match score and a score of 0.223 on semantic similarity. These numbers make it clear that language models, unless paired with dedicated computation tools, still struggle with calculation and numeric reasoning.

Multi-document queries were another weak spot. Combining information from different sources asks a lot from these systems. DeepSeek's exact match score landed at 0.200. LLaMA showed similar, only moderate, results. Clearly, sifting through and integrating data from multiple documents is a complex challenge.

Surprisingly, negative queries, where the goal is to reason about something missing or contradicting the context actually saw much better results after the latest tweaks. DeepSeek got a perfect exact match score of 1.000 and a semantic similarity score of 0.5191. Meanwhile LLaMA scored 0.00 on exact match and 0.1553 on semantic similarity. When the retrieval setup is solid and structured, these models struggle to handle "absence-based" reasoning.

Overall, both models perform well with factual and conceptual questions but struggle with numbers and multi-document tasks. These results drive home the point: strong retrieval and solid context make all the difference when it comes to handling the wide variety of queries

 ISBN: 978-81-685854-0-9

these models face.

Performance by Query Difficulty

Performance trends varied across different query difficulty levels.

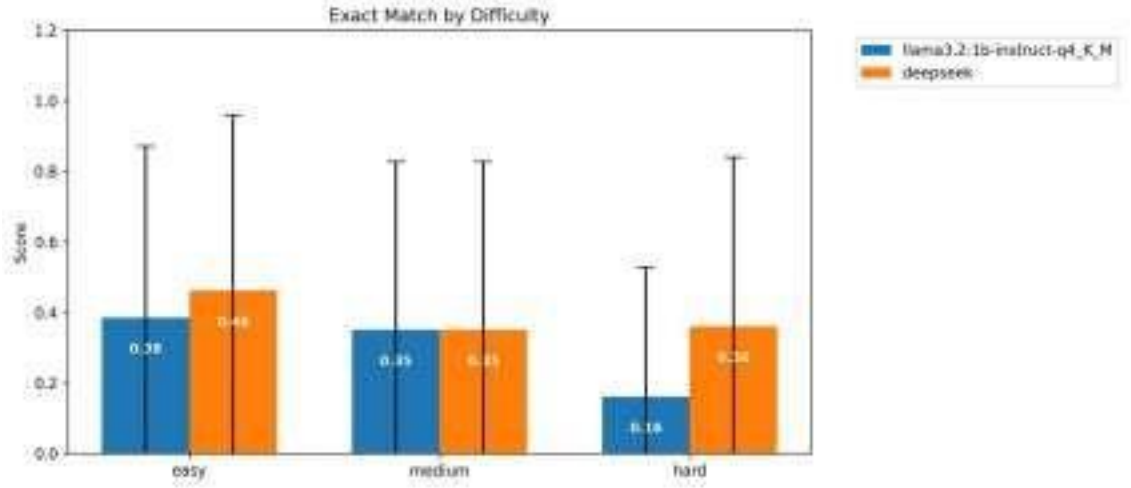


Figure 6.5.1 Exact match by query difficulty (rounded off to 2 decimal places)

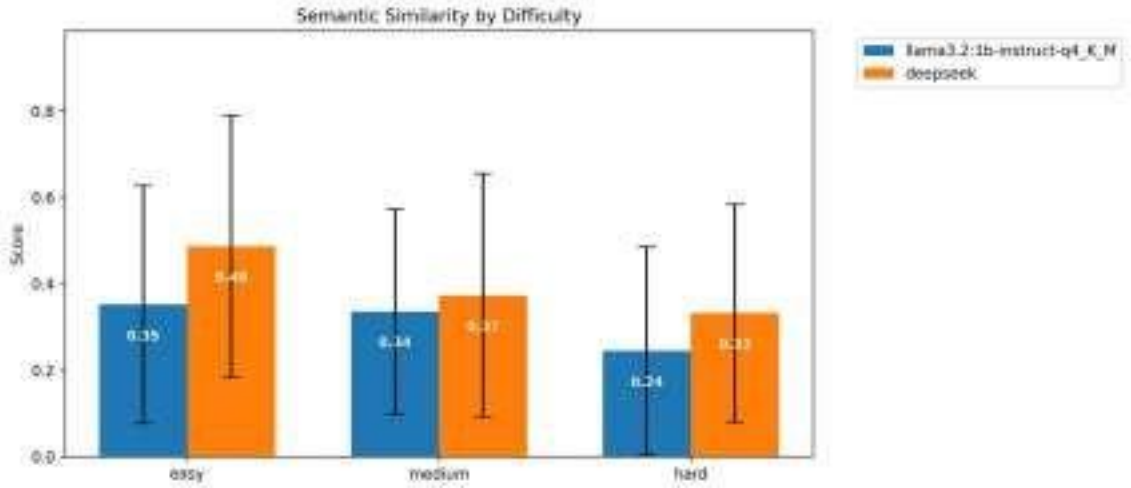


Figure 6.5.2 Semantic similarity by query difficulty (rounded off to 2 decimal places)

For easy queries, both models gave good response to straightforward factual and conceptual queries. The score achieved in Exact Match on easy question on DeepSeek is 0.461 and on LLaMA it is 0.384. Both models performed similarly for semantic similarity in easy queries, Deepseek gave a score of 0.487 and LLaMA gave a score of 0.352.

Scores seem to have dipped down a bit in the medium difficulty of query, in exact match and semantic similarity DeepSeek got a score of 0.35 and 0.3728 respectively. Whereas LLaMA got a score of 0.35 and 0.336 respectively.

The score again went down with increase in difficulty of the query as with DeepSeek the score in exact match cases is 0.36 and in semantic similarity is 0.332. While LLaMA scored 0.16 and 0.245 in exact match and semantic similarity on hard queries.

These results demonstrate a linear relationship between query complexity and response accuracy.

7. Discussion

The evaluation of the proposed Retrieval-Augmented Knowledge Repository system provides useful insights into the performance of both retrieval and answer generation components of

the system.

It can be observed from the retrieval performance that the vector database performed well in retrieval tasks: with a Recall@5 value of 0.643 it successfully retrieved most of the relevant document chunks in top 5 ranking, and an MRR of 0.604 reveals relevant documents appeared at higher positions within the retrieval outcomes on average. Though the Recall @5 was high, Precision @5 was only 0.559 indicating non-relevant chunks was occasionally retrieved which is a common characteristics of the retrieval system utilizing semantic similarity search.

However, comparing among different response generation models, it was found that DeepSeek-R1 resulted in higher semantic similarity score and exact match score over LLaMA 3.2 indicating DeepSeek-R1 has a better contextual reasoning capability. But it made at a cost of a significantly slower speed. LLaMA 3.2 produced response more quickly (averaged at 18.73 s) while DeepSeek averaged at 24.48 s (typical trade-off).

Both models achieved moderate grounding performance by faithfulness score showing some generated responses are not fully supported by retrieved context information, which is similar with the common issue faced by lightweight language models and can possibly be improved by stronger context filter or multi-turn verification mechanism.

The following aspect observed in the evaluation is the connection between the quality of retrieval and the quality of response. Clearly, there is significantly more context information readily available with the increased retrieval performance, and consequently, the response quality improves. The use of structured metadata, including document identifiers and chunk indices help by preserving document structure during the retrieval process.

The responses for factual and conceptual queries were highly accurate as all the responses were purely text based. Low performance was obtained for the Numeric and the Negative queries, showing the poor performance of lightweight language models for those types of information retrieval tasks.

Additionally, support for multiple document formats significantly improved the system's practical usability. The ability to process PDF, word-processing documents, spreadsheets, and image-based files ensures compatibility with real-world data sources commonly found in academic, organizational, and enterprise environments.

All evaluation experiments were conducted using locally hosted language models running on an NVIDIA GTX 1050 Ti GPU, which is considered a low-end graphics processor by 2026 standards. The system's ability to achieve consistent retrieval and response performance under these hardware constraints demonstrates its feasibility for deployment in environments with limited computational resources.

8. Applications

The proposed Retrieval-Augmented Knowledge Repository can be applied across multiple domains that require efficient knowledge organization and retrieval from heterogeneous data sources.

One of the primary application areas for the proposed system is enterprise knowledge management. Organizations frequently maintain large collections of documents, such as technical manuals, reports, policy documents, and operational guidelines. Retrieving relevant information can be time-consuming and inefficient using traditional systems. The proposed repository enables employees to interact with enterprise documents through natural language queries, significantly improving productivity and decision-making efficiency. (Lewis et al., 2020; Zhao et al., 2024)

One more important application area is academic research management. Universities and research institutions generate extensive volumes of academic papers, lecture materials, and experimental reports. The proposed system can be used to organize research materials and enable researchers to retrieve relevant information quickly. By supporting multi-format

 ISBN: 978-81-685854-0-9

ingestion, the repository allows integration of research data from multiple sources, including scanned laboratory records and structured datasets. (Z. Li et al., 2025)

Legal and compliance environments represent another promising application domain. Legal professionals often work with large volumes of case files, contracts, and regulatory documents. Efficient retrieval of relevant clauses and legal references is essential for accurate decision-making. The proposed system can assist legal professionals by enabling semantic retrieval of legal information across large document collections. (Cuconasu et al., 2024)

Educational institutions can also benefit from implementing intelligent knowledge repositories. Students can use the system to access course materials, lecture notes, and reference documents through conversational queries. This improves accessibility and enhances learning experiences by providing context-aware responses. (Z. Li et al., 2025)

Healthcare organizations represent an additional application domain where multi-format document retrieval is essential. Medical records, diagnostic reports, and research publications often exist in multiple formats. The ability to retrieve relevant patient information quickly can significantly improve clinical decision-making and operational efficiency. (Amugongo et al., 2025)

The proposed repository also has potential applications in customer support systems. Organizations can integrate knowledge repositories containing troubleshooting guides and technical documentation. Customers can interact with these repositories through conversational interfaces to obtain solutions to common problems. These diverse application scenarios demonstrate the flexibility and scalability of the proposed system architecture.

9. Limitations

Despite the strong performance of the proposed Retrieval-Augmented Knowledge Repository, several limitations were identified during development and testing. Recognizing these challenges helps guide future improvements and sets realistic expectations for system performance.

One limitation relates to retrieval accuracy in complex queries. While the system achieved strong recall, precision remained moderate, meaning some retrieved document segments were not fully relevant. This suggests the need for advanced reranking methods to improve result quality. Another limitation involves response faithfulness. Some generated answers included information not directly supported by retrieved content, reflecting common challenges in language model generation.

Latency also remains a concern, especially when using larger language models. For example, DeepSeek produced better semantic similarity results but required longer response times, highlighting the trade off between accuracy and speed. The current system also relies mainly on semantic retrieval without hybrid methods, which could improve performance for rare or domain-specific terms. In addition, the absence of caching means every query is processed from scratch, increasing response time for repeated queries.

Memory usage can become challenging when handling large document collections, as storing many embeddings may require distributed storage solutions. OCR performance is another limitation, since poor-quality images can lead to inaccurate text extraction and affect retrieval results.

Despite these challenges, the system provides a strong base for further improvement and future optimization.

10. Future Work

Although the proposed Retrieval-Augmented Knowledge Repository performs well across evaluation metrics, several areas remain for improvement. Future enhancements will focus on improving accuracy, reducing latency, and expanding system capabilities for real-world use.

One important direction is the addition of hybrid retrieval methods. While the current system relies on semantic vector similarity, combining it with keyword-based methods like BM25 could improve precision, especially for domain
specific terms.
Another improvement involves adding reranking models to refine the order of retrieved results and improve relevance. Query caching could also reduce response time by storing results for frequently repeated queries. Streaming responses is another useful feature, allowing answers to appear gradually instead of waiting for full completion, which improves user experience.
Scalability will also be important as document collections grow. Distributed storage and indexing can help maintain fast retrieval speeds for large datasets. Future work may also expand the system to support multi-modal data, including audio and video, alongside text and images. Improving OCR accuracy through better image preprocessing can further enhance text extraction from scanned documents.
Security can be strengthened by adding role-based access control and document-level permissions. Finally, large-scale benchmarking across multiple datasets will help better evaluate system performance and reliability in real-world environments.

11. Conclusion

This research focused on designing, building, and evaluating a Retrieval-Augmented Knowledge Repository that supports multi-format document ingestion and conversational knowledge retrieval. The system combines semantic retrieval with large language model reasoning, allowing users to interact efficiently with different types of data sources.
The repository follows a modular architecture that separates document processing, retrieval, and response generation. This design makes the system flexible and easier to improve over time, as each component can be optimized independently. Secure authentication features were also included to ensure controlled access and protect user-specific data.
One of the major strengths of the system is its ability to handle multiple document formats. It supports files such as pdf, docx, csv, excel, and image-based documents, making it suitable for real-world knowledge management tasks. Semantic embedding and use of metadata and chunk based indexing helped maintain meaningful context during retrieval. While intent based query routing improved efficiency to work with different types of queries
To evaluate performance, a detailed framework was created to measure retrieval accuracy, response quality, and system latency. Testing with 91 queries demonstrated reliable system performance across retrieval accuracy, response quality, and latency metrics. The vector retrieval system achieved Precision@5 = 0.559, Recall@5 = 0.643, F1@5 = 0.583, MRR = 0.604, and MAP = 0.593, indicating consistent identification of relevant document content. Response quality evaluation showed that deepseek-r1:1.5b achieved higher semantic similarity and exact match performance, while llama3.2:1b provided faster response times, highlighting the trade-off between response accuracy and computational efficiency.
Despite these positive results, various limitations remain. The evaluation dataset size was moderate, and performance decreased for complex query types such as numeric and negative ones. Moreover, latency differences between models indicate the need for optimization for deploying larger models in resource constrained environments such as higher context window size
In the future, work will involve focus on improving retrieval precision by implementing hybrid retrieval techniques and reranking mechanisms. Exploring distributed vector indexing for large scale deployments can give a realistic view of data management in large enterprises. More improvements may include expanding evaluation datasets and integrating specialized reasoning modules to improve handling of structured numerical queries.
Overall, the designed RAG Knowledge Repository demonstrates a practical and scalable

 ISBN: 978-81-685854-0-9

approach for managing heterogeneous knowledge sources. The results confirm that combining semantic retrieval with conversational interfaces significantly improves contextual information access, making the system suitable for deployment in academic, enterprise, and research environments

12. References

1. Amugongo, L., Mascheroni, P., Brooks, S., Doering, S., & Seidel, J. (2025). Retrieval augmented generation for large language models in healthcare: A systematic review. PLOS Digital Health, 4. https://doi.org/10.1371/journal.pdig.0000877
2. Chen, J., Lin, H., Han, X., & Sun, L. (2023). Benchmarking Large Language Models in Retrieval Augmented Generation. 17754–17762. https://doi.org/10.48550/arxiv.2309.01431
3. Cuconasu, F., Trappolini, G., Siciliano, F., Filice, S., Campagnano, C., Maarek, Y., Tonellotto, N., & Silvestri, F. (2024, January 26). The Power of Noise: Redefining Retrieval for RAG Systems. Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. https://doi.org/10.1145/3626772.3657834
4. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. ArXiv, abs/2005.11401.
5. https://consensus.app/papers/retrievalaugmented-generation-for-knowledgeintensive-lewis perez/fbc9d8d6f6de501987cc8c3afa034696/
6. Li, M., Kilicoglu, H., Xu, H., & Zhang, R. (2024). BiomedRAG: A Retrieval Augmented Large Language Model for Biomedicine. Journal of Biomedical Informatics, 104769. https://doi.org/10.48550/arxiv.2405.00465
7. Li, Z., Wang, Z., Wang, W., Hung, K., Xie, H., & Wang, F. L. (2025). Retrieval-Augmented Generation for Educational Application: A Systematic Survey. Computers and Education: Artificial Intelligence. https://doi.org/10.1016/j.caeai.2025.100417
8. Oche, A. J., Folashade, A. G., Ghosal, T., & Biswas, A. (2025). A Systematic Review of Key Retrieval-Augmented Generation (RAG) Systems: Progress, Gaps, and Future Directions. ArXiv, abs/2507.18910. https://doi.org/10.48550/arxiv.2507.18910
9. Peng, B., Zhu, Y., Liu, Y., Bo, X., Shi, H., Hong, C., Zhang, Y., & Tang, S. (2024). Graph Retrieval Augmented Generation: A Survey. ACM Transactions on Information Systems. https://doi.org/10.1145/3777378
10. Shao, Z., Gong, Y., Shen, Y., Huang, M., Duan, N., & Chen, W. (2023). Enhancing Retrieval Augmented Large Language Models with Iterative Retrieval-Generation Synergy. abs/2305.15294. https://doi.org/10.48550/arxiv.2305.15294
11. Singh, A., Ehtesham, A., Kumar, S., & Khoei, T. (2025). Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG. ArXiv, abs/2501.09136. https://doi.org/10.48550/arxiv.2501.09136
12. Yu, H., Gan, A., Zhang, K., Tong, S., Liu, Q., & Liu, Z. (2024). Evaluation of Retrieval-Augmented Generation: A Survey. ArXiv, abs/2405.07437. https://doi.org/10.1007/978-981-96-1024-2_8
13. Zhao, P., Zhang, H., Yu, Q., Wang, Z., Geng, Y., Fu, F., Yang, L., Zhang, W., & Cui, B. (2024). Retrieval-Augmented Generation for AI-Generated Content: A Survey. ArXiv, abs/2402.19473. https://doi.org/10.48550/arxiv.2402.19473

TELECURA: A UNIFIED TELEMEDICINE PLATFORM WITH CREDIT-BASED MONETIZATION, ADMINISTRATIVE GOVERNANCE, AND WEBRTC-ENABLED VIDEO CONSULTATIONS

Abhishek Mane
Department of Information Technology and Computer Science, Somaiya School of Basic and Applied Sciences, Somaiya Vidyavihar University, Mumbai, India.
Email: apm@somaiya.edu

Prof. Prajakta Kambale
Department of Information Technology and Computer Science, Somaiya School of Basic and Applied Sciences, Somaiya Vidyavihar University, Mumbai, India.
Email: prajakta.k@somaiya.edu

Abstract
Digital health service expansion has been rapidly increasing recently; however, many impediments still exist that will prevent individuals from accessing equitable healthcare. Geographic isolation, unclear pricing structure, disjointed tooling, and insufficient provider verification, among other barriers, will all impede access to equitable healthcare. To improve equitable access to healthcare services, this paper focuses exclusively on the TeleCura web-based telehealth platform, which consolidates the aforementioned barriers related to accessing telehealth services. TeleCura offers the following key characteristics via a single ecosystem: doctor discovery across 12 specialty areas, suite-based appointment scheduling with overlap detection; one-way high-definition video conferencing; and new transparent payment mechanism using credits (via debit-billed credit). To develop the platform, we utilized a zero API paradigm using Next.js 16 server actions, Prisma ORM with PostgreSQL (which provides ACID-compliant transaction), Clerk (for multi-factor authentication and subscriptions), and Vonage Video API (for encrypted, server-relayed video streaming). To support the function of TeleCura, we developed an administrative governance layer to support credential verification, account lifecycle management, and payout processing (20% of platform revenue is shared). Experimental testing of our simulated concurrent user loads of 50 to 200 users based upon our infrastructure demonstrates a sub-second server action response time, a page load interactivity of under 2.5 seconds for non-video routes, and sustained 720p video quality at a 180 ms latency for server to client end-to-end. An analysis of the four major existing telemedicine platforms (Practo, Teladoc Health, Doxy.me and Zocdoc) has determined that TeleCura provides six key capabilities not found in any one of the existing platforms. The proposed TeleCura platform provides a strong technical framework for the scalable and transparent delivery of governance-based telehealth. This indicates that TeleCura is a superior solution to any existing telemedicine platform.
Keywords: Telemedicine, WebRTC, Credit-Based Monetization, Server Actions, Video Consultation, Healthcare Accessibility, Role-Based Access Control, ACID transactions.

I. Introduction

The health care system experienced a major shift during the last 10 years due to a number of factors including technological developments, changing consumer preferences and expectations regarding service delivery, and the Covid-19 pandemic which demonstrated that

 ISBN: 978-81-685854-0-9

existing models of delivering health care were unable to sustain themselves under conditions of significant disruption. Telemedicine has been defined by the World Health Organization as delivering care over distances using various forms of Information Communication Technology (ICT) [1]. During the Covid-19 pandemic (2020-2021), some areas experienced a dramatic increase (>3000%) in the volume of virtual consultations being scheduled [2]. This increase indicates that there was a large unmet demand for remote health services and that existing infrastructures were not sufficiently robust to meet that demand in a sustainable manner [16],[20].

The substantial increase in patients, particularly rural and semi-urban patients from developing countries, has left many without access to specialist consultations in a timely manner. Many patients need to travel significant distances at high prices for lengthy amounts of time before they can see a specialist. Fixed hours of operation and limited scheduling availability cause delays to occur, sometimes taking numerous business days or weeks. Overall, the financial aspects of the telemedicine experience create significant additional barriers, such that current telemedicine solutions are predominantly based on either insurance reimbursement models or other types of non-transparent direct payment models, causing patients to be unable to estimate their cost for visiting with a specialist prior to scheduling an appointment. Given that low-income individuals and those without health insurance generally face greater challenges accessing telemedicine services than individuals with health insurance [3], [17]; it stands to reason that low-income individuals would also be most impacted by the telemedicine barriers discussed below.

As a result of the systemic deficiencies discussed above, telemedicine providers encounter additional barriers due to the lack of standardised governance frameworks at the system level. There are currently very few broadly available commercial telemedicine solutions in the marketplace; however, for those solutions that are broadly available, there is no common solution for providing governance at the system level, including: credentialing of providers, lifecycle management of accounts, and financial oversight of accounts [4]. Video communication solutions (e.g., Doxy.me) do not contain scheduling or payment features; and scheduling solutions (e.g., Zocdoc) generally have third-party integrations for recording video. The lack of standardised governance of telemedicine solutions creates numerous barriers, as well as increasing the complexity of developing telemedicine solutions and creating more potential points of failure, which ultimately prevents the effective management of all aspects of telemedicine.

TeleCura aims to solve the problems of telehealth service providers and patients through a consolidated web-based telemedicine platform offering all necessary functionality within one application, including video consultation, scheduling appointments, securely using credit cards as payment methods, verifying credentials for clinicians, managing payouts, and supporting easily extensible open architecture. The application supports three different end-user types (Patient, Doctor, Administrator) across 12 specialties and was built with modern full-stack architecture intended for both ease of development and operational scalability.

The principal contributions of this work are:

1. A zero-API full-stack architecture using Next.js Server Actions that provides zero backend API servers and end-to-end type safety.

2. A credit-based business model with 3 subscription levels and guaranteed rates for consultations with each transaction being atomic.

3. A multi-layered security model that uses JWT for time-based access to video consultations, ensures that credit transactions are ACID-compliant, and mitigates the OWASP Top 10 vulnerabilities automatically.

4. An administrative governance framework for managing credential verification workflows to enable suspension or reinstatement of accounts, and to provide visibility into the payouts process with customizable revenue sharing options.

Organization of the Remainder of This Paper The remainder of this paper is structured as follows: related work will be surveyed (Section II) to highlight the limitations of existing platforms; the proposed methodology will be explained (Section III); the system architecture will be explained (Section IV); the specifics of the implementation including the database design, core algorithms and interface design will be presented (Section V); the experimental setup will be outlined (Section VI); results and discussion will be discussed (Section VII); a comparison with existing work will be provided (Section VIII); security issues will be discussed (Section IX); the paper will conclude (Section X); and the next steps for future research will be proposed (Section XI).

II. Related Work

A. Evolution of Telemedicine Platforms

The evolution of telemedicine has transformed from simple telephone consultations in the 1960s to today's advanced video conferencing tools utilizing WebRTC and cloud infrastructure. The full scope of this evolution is described by Bashshur et al. who have established a substantial body of research demonstrating the effectiveness of telemedicine to manage chronic disease and deliver mental health services and provide post-treatment follow-ups [5]. The American Telemedicine Association projects that virtual visits may account for as much as 50% of all outpatient visits by 2030 [6], which will require the development of robust and scalable telemedicine platforms [16],[18].

B. Analysis of Existing Systems

Four telemedicine platforms were selected for commercial dominance and evaluated to develop a baseline for comparative analysis. Practo, the largest telemedicine platform in India, has a doctor discovery, appointment scheduling and video consultation function through a direct payment fee-for-service model. There is currently no single administrator governing the total telemedicine platform or any credit-based telemedicine model through Practo [7]. For the global telemedicine services leader, Teladoc Health, the telemedicine delivery model consists of a significant number of participating physicians and complex payment mechanisms based on insurance. Teladoc Health has not provided for any administrative tools to govern the participating telemedicine platform independently of Teladoc Health [8]. Doxy.me has a simple HIPAA-compliant browser-based video conferencing service that operates as a telemedicine communication portal and does not offer appointment scheduling, payment processing, or administrative functions [9]. Zocdoc is a telemedicine platform specializing in doctor discovery and appointment scheduling, but does not have a telemedicine video conference system, thereby contributing to workflow fragmentation.

Table I. Comparative Analysis of Existing Telemedicine Platforms

Feature	**Practo**	**Teladoc**	**Doxy.me**	**Zocdoc**	**TeleCura**
Video Consultation	✓	✓	✓	✗	✓
Appointment Scheduling	✓	✓	✗	✓	✓
Integrated Payments	Direct Pay	Insurance	✗	Insurance	Credit System
Admin Panel	✗	✗	✗	✗	✓

 ISBN: 978-81-685854-0-9

Doctor Verification	Partial	Internal	✗	Partial	✓ (Admin)
Provider Payout Mgmt	✗	Internal	✗	✗	✓
Open/Extensible	✗	✗	✗	✗	✓
Time-Gated Video	✗	✗	✗	✗	✓ (JWT)
Subscription Tiers	✗	Employer	✗	✗	3 Tiers

C. Identified Limitations

An analysis of the literature and market shows that the following six limitations persist in those examined: 1) Lack of clarity on the monetization process caused by the use of insurance as well as unreported direct pay models makes it impossible to estimate the cost prior to booking, while a transparent credit system eliminates that barrier; 2) No consistent interface exists across all platforms for administrator governance, providing information and financial oversight; 3) Very complex integration of multiple open-source telehealth components increases the likelihood of failure; 4) Infrequently occurring verified administrators establish how service providers are able to get verified through a system; 5) Because payout mechanisms used by providers are not a part of any one platform, it has become difficult for providers to determine how much they will be paid for services provided; 6) Complexity with scheduling creates significant barriers for healthcare providers. TeleCura has been developed to achieve these six limitations through an integrated approach using structural standardization.

III. Proposed Methodology

A. Technology Stack

The modern full-stack JavaScript/TypeScript technology stack has been chosen for the platform to optimize performance, security and development speed. The frontend layer is built using the following technologies: React 19.2.3 (for component based UI), Next.js 16.1.6 (the full stack framework that provides support for server side rendering, file-based routing and Server Actions), Tailwind CSS 4.x (utility-first responsive styling) Radix UI 1.4.3 (headless accessible UI primitives), and React Hook Form 7.71.2 (form state management with integrated validation).

Table II. Frontend Technology Stack

Technology	**Version**	**Role**
React	19.2.3	Component-based UI construction
Next.js	16.1.6	Full-stack framework: SSR, routing, Server Actions
Tailwind CSS	4.x	Utility-first responsive CSS styling
Radix UI	1.4.3	Accessible headless UI primitives
Shadcn/ui	Latest	Pre-built component library
React Hook Form	7.71.2	Form state management and validation
date-fns	4.1.0	Date formatting and slot generation

Table III. Backend Technology Stack

Technology	**Version**	**Role**
Next.js Server Actions	16.1.6	Server-side business logic (zero-API)
Prisma ORM	7.4.2	Type-safe DB access with ACID transactions
PostgreSQL (Neon)	Latest	Serverless relational database
Clerk	6.39.0	Authentication, sessions, subscriptions
Vonage Server SDK	3.26.4	WebRTC video sessions and JWT tokens
Zod	3.x	Runtime schema validation

B. Credit-Based Monetization Model

There are three tiers of subscription, which are based on the monetization system, and operate using a deterministic per-engagement price. The price for each engagement will be based on two credits, regardless of the provider or specialty. The free subscription tier provides a one-time aesthetic welcome bonus of two credits for accessing the platform to explore the apps. When subscribing to the standard subscription tier, patients are charged $10 each month for the first ten credits provided to the patient. Thereby providing a total effective price per consultation of $2.00. Premium Subscription Tiers are charged $20 for twenty-four credits provided to the patient each month, which allows for a total effective price per consultation of $1.67. Credits allocated do not roll over to subsequent billable months, and all credits are allocated instantly whenever the patient accesses the platform for the first time each billing month.

Table IV. Subscription Tier Structure

Plan	**Monthly Price**	**Credits**	**Cost/Consultation**	**Renewal**
Free	$0	2 (one-time)	$0 (first visit)	Non-renewable
Standard	$10/month	10/month	$2.00	Monthly reset
Premium	$20/month	24/month	$1.67	Monthly reset

C. Mathematical Model for Platform Economics

The way that the platform generates revenue and compensates providers through its credit system can be described as follows: let Cp be the total number of credits purchased by all patients during a given billing period, A be the total number of appointments that were completed during the period, and k=2, which is the fixed number of credits to charge for each appointment (the cost per appointment to retrieve an appointment from the platform). The platform applies a revenue-sharing ratio of r=0.20 (20% platform fee) to all payments made to providers.

The total credit consumption Tc for a billing period with A completed appointments is:

$$Tc = k \times A = 2A \quad \text{... (1)}$$

When a provider with n completed consultations requests a payout, the gross credit earnings Eg, platform fee Fp, and net provider payout En are computed as:

$$Eg = k \times n = 2n \quad \text{... (2)}$$

$$Fp = r \times Eg = 0.20 \times 2n = 0.4n \quad \text{... (3)}$$

$$En = Eg - Fp = (1 - r) \times Eg = 0.80 \times 2n = 1.6n \quad \text{... (4)}$$

The total platform revenue R from subscription fees and platform commissions over a period with Sstd Standard subscribers, Spre Premium subscribers, and total payout-eligible credits Pc is:

 ISBN: 978-81-685854-0-9

$$R = (10 \times Sstd) + (20 \times Spre) + (r \times Pc) \quad \ldots (5)$$

Table V. Revenue Sharing Model

Consultations	Credits Earned	Platform Fee (20%)	Doctor Net (80%)
5	10	2 credits	8 credits
10	20	4 credits	16 credits
20	40	8 credits	32 credits
50	100	20 credits	80 credits

IV. System Architecture

TeleCura uses a four-tiered layered architecture consisting of the Client Layer, Application Layer, External Services Layer and Data Layer. This architectural decomposition allows for a separation of concerns, which allows the independent scaling and maintenance of each tier in accordance with the principles of distributed system design [19],[24].

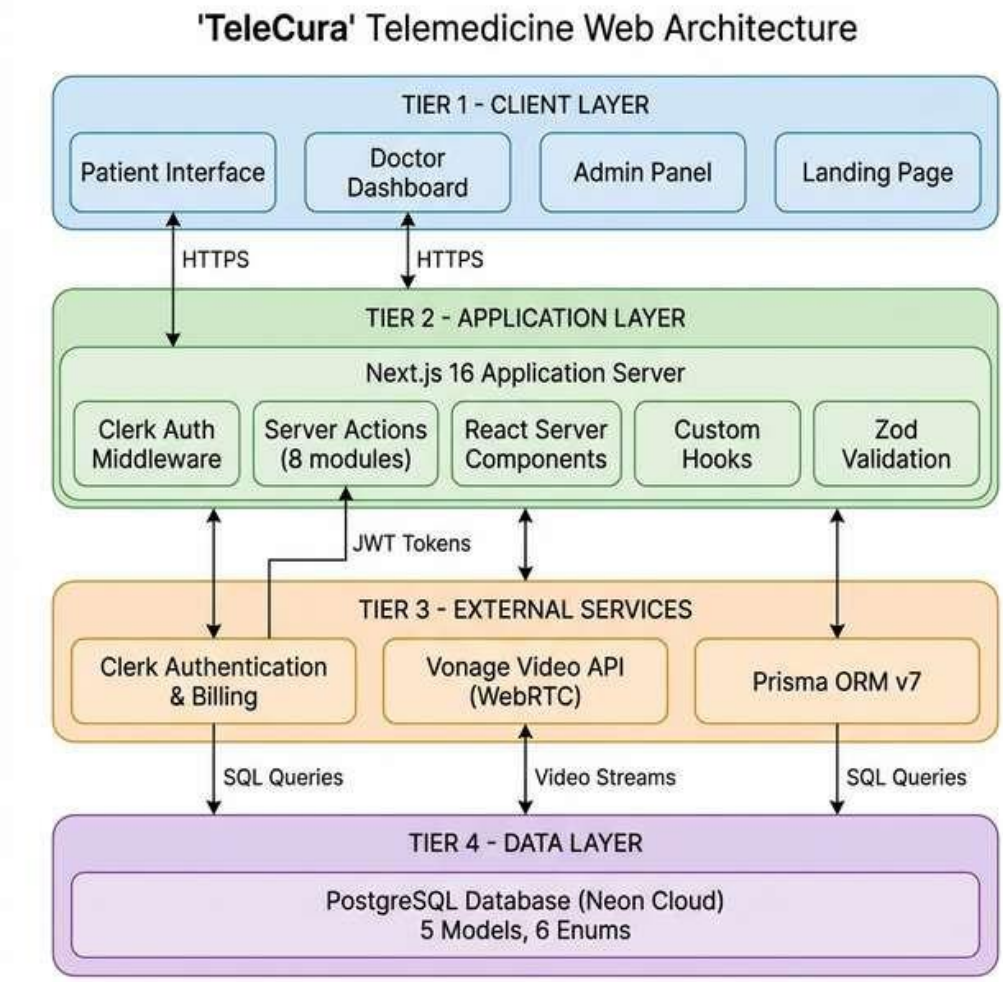


Fig. 1. System Architecture of TeleCura Platform

Client Layer: The position on the client layer is to create interfaces for patients; doctors, and administrators (Public) in compliance with the design guidelines established by Tailwind CSS. RADIX UI provides accessible UI elements that use utility classes that form the basis of the design system of each interface. Each interface is designed for display on a variety of devices (desktop/tablet/mobile) in both dark and light mode when using a next-themes library.

ISBN: 978-81-685854-0-9

Application Layer: The core business logic for the application is executed using 8 Next.js Server Action modules named as follows: admin.js; appointments.js; credits.js; doctor.js; doctors-listings.js; on-boarding.js; patient.js; and payout.js. Each of these modules has been prefixed with "use server" to ensure that its logic is executed on the server. Each of these modules also contains all of the transaction logic that supports the scheduling of an appointment; credit issuance/management/verification; and payout processing. The Clerk middleware inspects any incoming requests to the server to provide authentication for users on any secured route.

External Services Layer: The External Services Layer has been established for integration with three external services: Clerk (for managing user authentication, sessions, and subscriptions); Vonage Video API (for providing WebRTC based server to server video streaming); and Prisma ORM (for using type-safe queries and transactions when performing all database functions, ensuring that they are consistent with the "database as a service" practices in distributed databases [20]).

Data Layer: A PostgreSQL database hosted on Neon Cloud stores persistent application state across five normalized tables (User, Appointment, Availability, CreditTransaction, Payout) and six enumerated types (UserRole, VerificationStatus, SlotStatus, AppointmentStatus, TransactionType, PayoutStatus). Composite indexes on frequently queried columns optimize read performance for time-sensitive operations.

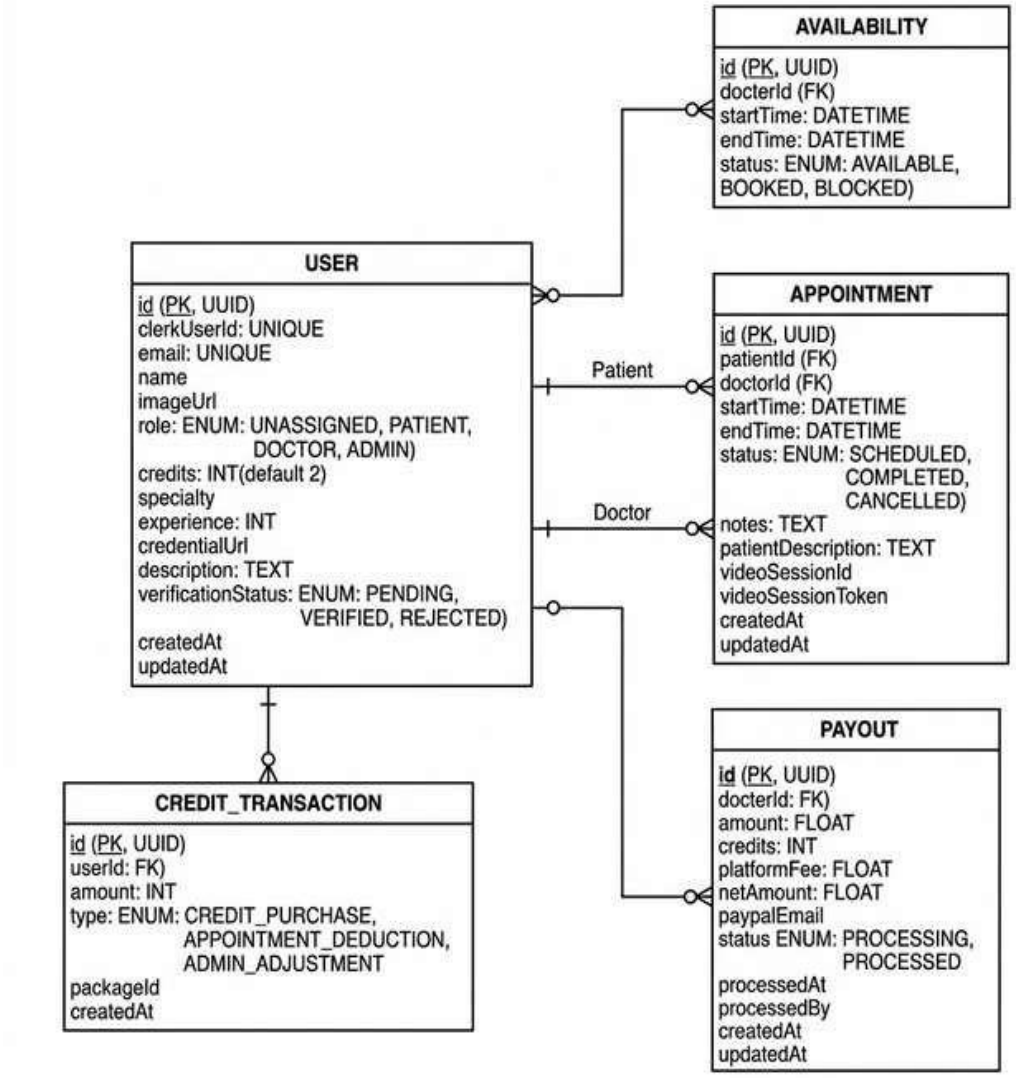


Fig. 2. Entity-Relationship Diagram of TeleCura Database

 ISBN: 978-81-685854-0-9

The Entity Relationship Diagram includes a total of five entities: User (the primary entity) with attributes id, clerkUserId, email address, name, role, credit amount, specialty, experience, and verification status; Appointment, which is linked to User via both the patientId and doctorId foreign keys, and has attributes for each appointment including start and end times, status, video_session_id, and notes; Availability which is linked to User via doctorId, and has start and end times and status; CreditTransaction, which is linked to User via userId, and has amount of credit used or purchased, type of transaction, and packageId; Payout which is linked to user via doctorId, and has amount due to doctor, credits used, platform fees and net amount payable to doctor via PayPal, status of the payout (i.e. pending/processed). All relations are one-to-many from the User to the associated entity.

V. Implementation

A. Database Schema

There are five tables in the relational schema that have been created to ensure normalization, referential integrity, and transactional consistency. The User table is the main entity and contains authentication identifiers, personal information, rolebased attributes (RBAs), and credit balances. The User table also contains information related to doctors, including their specialty, years of experience, the URL(s) where their credentials can be verified, and whether or not they are verified.

Table VI. User Table Schema

Column	**Data Type**	**Constraints**
id	UUID	PRIMARY KEY, DEFAULT uuid()
clerkUserId	String	UNIQUE, NOT NULL
email	String	UNIQUE, NOT NULL
name	String	NULLABLE
imageUrl	String	NULLABLE
role	UserRole (ENUM)	DEFAULT 'UNASSIGNED'
credits	Integer	DEFAULT 2
specialty	String	NULLABLE
experience	Integer	NULLABLE
credentialUrl	String	NULLABLE
description	Text	NULLABLE
verificationStatus	VerificationStatus	DEFAULT 'PENDING'
createdAt	DateTime	DEFAULT now()
updatedAt	DateTime	AUTO-UPDATED

Table VII. Appointment Table Schema (Indexes: [status, startTime], [doctorId, startTime])

Column	**Data Type**	**Constraints**
Id	UUID	PRIMARY KEY, DEFAULT uuid()
patientId	String (FK)	REFERENCES User(id)
doctorId	String (FK)	REFERENCES User(id)
startTime	DateTime	NOT NULL
endTime	DateTime	NOT NULL
status	AppointmentStatus	DEFAULT 'SCHEDULED'
notes	Text	NULLABLE
patientDescription	Text	NULLABLE

videoSessionId	String	NULLABLE
createdAt	DateTime	DEFAULT now()

Table VIII. Enumerated Types

Enum	**Values**
UserRole	UNASSIGNED, PATIENT, DOCTOR, ADMIN
VerificationStatus	PENDING, VERIFIED, REJECTED
SlotStatus	AVAILABLE, BOOKED, BLOCKED
AppointmentStatus	SCHEDULED, COMPLETED, CANCELLED
TransactionType	CREDIT_PURCHASE, APPOINTMENT_DEDUCTION, ADMIN_ADJUSTMENT
PayoutStatus	PROCESSING, PROCESSED

B. Core Algorithms

Algorithm 1: Appointment Booking with Overlap Detection and Atomic Credit Transfer

```
INPUT: patientClerkId, doctorId, startTime, endTime, description
OUTPUT: Appointment record with video session, or error

1. patient ← DB.FindUser(clerkId=patientClerkId, role=PATIENT)
2. IF patient = NULL THEN RETURN Error("Patient not found")
3. doctor ← DB.FindUser(id=doctorId, role=DOCTOR, status=VERIFIED)
4. IF doctor = NULL THEN RETURN Error("Not found/not verified")
5. IF patient.credits < 2 THEN RETURN Error("Insufficient credits")
6. // Three-condition overlap detection
7. overlap ← DB.FindAppointment WHERE doctorId AND status=SCHEDULED
       AND (exist.start ≤ start < exist.end)
       OR (exist.start < end ≤ exist.end)
       OR (start ≤ exist.start AND exist.end ≤ end)
8. IF overlap ≠ NULL THEN RETURN Error("Slot already booked")
9. sessionId ← Vonage.CreateSession(mediaMode="routed")
10. TRANSACTION BEGIN
   11. CreateCreditTx(patient.id, amount=-2, type=DEDUCTION)
   12. CreateCreditTx(doctor.id, amount=+2, type=DEDUCTION)
   13. UpdateUser(patient.id, credits = credits - 2)
   14. UpdateUser(doctor.id, credits = credits + 2)
15. TRANSACTION COMMIT
16. appointment ← CreateAppointment(all params, status=SCHEDULED)
17. RETURN Success(appointment)
```

Algorithm 2: Monthly Credit Allocation with Duplicate Prevention

```
INPUT: Authenticated patient user record with transaction history
OUTPUT: Updated user with allocated credits, or unchanged

1. IF user = NULL OR user.role ≠ PATIENT THEN RETURN user
2. currentPlan ← Clerk.GetSubscriptionPlan(user)
```

 ISBN: 978-81-685854-0-9

```
3.  IF currentPlan = NULL THEN RETURN user
4.  credits ← PlanCredits[currentPlan]
        // PlanCredits = {free:0, standard:10, premium:24}
5.  currentMonth ← FormatDate(now(), "YYYY-MM")
6.  IF user.transactions.length > 0 THEN
    7.   latestTx ← user.transactions[0]
    8.   txMonth ← FormatDate(latestTx.createdAt, "YYYY-MM")
    9.   IF txMonth=currentMonth AND latestTx.plan=currentPlan
    10.    THEN RETURN user  // Already allocated
11. TRANSACTION BEGIN
    12.  CreateCreditTx(user.id, credits, CREDIT_PURCHASE, plan)
    13.  updatedUser ← UpdateUser(user.id, credits += credits)
14. TRANSACTION COMMIT
15. RETURN updatedUser
```

C. Video Session Security

Four security layers are provided by the mechanism of token generation via video. An initial layer of security is verification via Clerk identity verification. The Clerk's identity is verified against patient and doctor appointment records prior to granting access to the video session. The second layer consists of the time gating feature that does not allow access prior to the 30 minute window before the start of a video session. The third existing layer of security includes token expirations, which require all JWT tokens created from a user request to expire exactly one hour after the end of the user's scheduled appointment. The fourth security layer involves connection metadata in that each JWT token contains the user's name, role, and user ID for identified session participants.

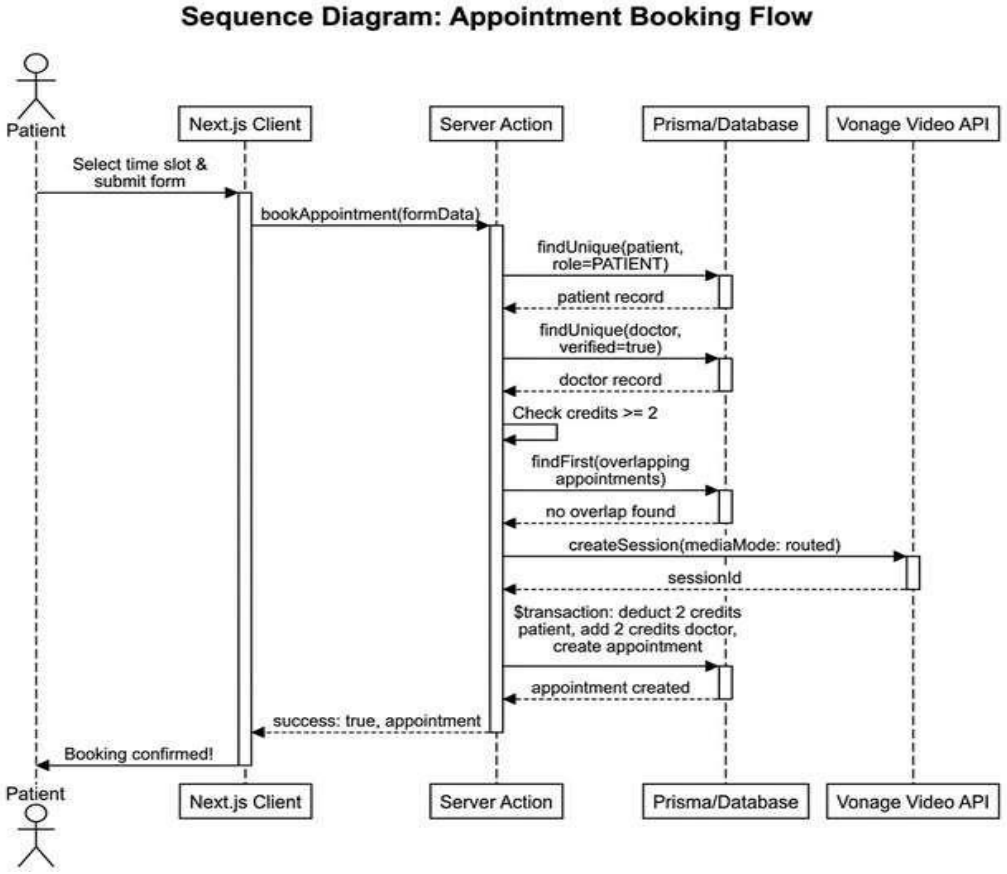


Fig. 3. Sequence Diagram — Appointment Booking Flow

In the Appointment Booking Sequence Diagram, the steps taken to Book an Appointment are demonstrated by interaction flow throughout this process. After a Patient requests to book an appointment using the Client interface, the Client then invokes the bookAppointment Server Action. The bookAppointment Server Action queries the Database for the accepted Patient, Doctor acceptance and checks to see if there is money left to pay for the appointment . The Server Action then queries to find out if any of the three (3) conditions that define Overlapping Appointments are met. Once approved through all five (5) checks, the Server Action calls upon the Vonage Video API to complete creation of the routed media session. The Server Action completes execution with four (4) atomic database transactions which total two (2) credit transaction records and two (2) user balance updates. Once all conditions are completed and verified, the Server Action creates an Appointment record, and completes the process by propagating the success response back through the chain of request.

D. System Workflow

The TeleCura platform orchestrates interactions across three user roles through a unified authentication entry point, branching into role-specific operational pipelines. The complete system workflow, depicted in Fig. 4, proceeds as follows.

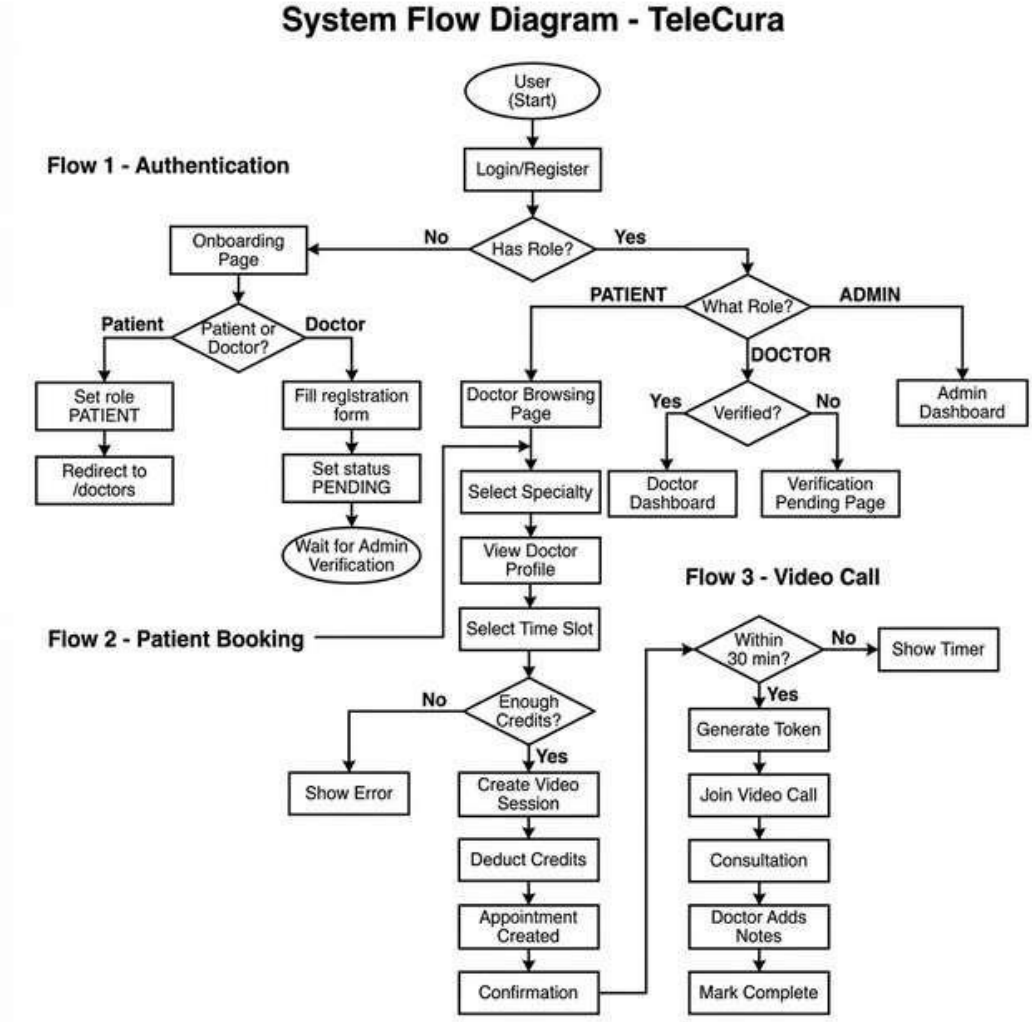


Fig. 4. System Flow Diagram

Authentication and onboarding include the processes of booking appointments and conducting video consultations in real-time. The first step in this process is the authentication of the user

 ISBN: 978-81-685854-0-9

through either a login or registration process. If a user does not have a role assigned to them, the user is redirected to an interface that allows them to select their role (patient or doctor) and proceed. Once a patient has selected their role as a patient, they will have immediate access to the platform, whereas if the user has selected doctor as their role, they must first complete a doctor's registration process which will then be verified by an administrator before being granted access to the functionality of the system.

Once the patient has been authenticated, they will enter the booking workflow. The patient will search for a doctor by specialty, view the profile of the doctor they wish to meet with, and select from a calendar of predetermined time slots. Prior to the patient confirming their appointment, the system will check the patient's balance to ensure they have enough credits to book an appointment with their selected doctor. Once the patient's credit balance has been validated, the system will create a video session and charge the patient's account for an atomic (immediate) transaction and will confirm the appointment.

The consultation workflow has a time-gated access method for user's to create a link that will allow them access to the video consultation session. The patient and doctor will be permitted to join the video consultation, however the time to join the video consultation will only be allowed during a designated time frame of 10 minutes before the scheduled appointment. Also generated during the appointment request is a secure token that provides the patient and doctor with access to the session, and the low latency of the WebRTC communication protocol will allow for real-time communication between the patient and doctor [21]. Once the video consultation has been completed, the doctor will enter notes related to the consultation and then mark the consultation as completed.

The design of the workflow fosters optimal communication between user activities (i.e., entering data into forms), system validation, and secure messaging, while providing transactional integrity and role-based access control throughout the platform.

The modular separation of workflow increases maintainability, while improving each authentication, booking and consultation subsystem independently through optimization.

VI. Experimental Setup

A. Testing Environment

The performance of each application was evaluated based upon an experimental evaluation in the deployed production application hosted on Vercel's global edge network in conjunction with a Neon Cloud PostgreSQL instance in the US-East region. Client-side measurements were gathered using two connection profiles (Broadband of 50 Mbps downstream, 10 Mbps upstream and with a 20 ms RTT and Simulated 4G of 20 Mbps downstream, 5 Mbps upstream and with a 50 ms RTT) using Google Chrome Developer Tools (NETWORK/PERFORMANCE panels).

B. User Simulation

To evaluate system behavior under varying concurrency conditions, simulated user loads were generated across four tiers representing light to stress-level usage patterns.

Table IX. Simulated User Load Configuration

Load Tier	Concurrent Users	Composition	Duration
Light	50	35 patients, 12 doctors, 3 admins	15 minutes
Moderate	100	70 patients, 25 doctors, 5 admins	30 minutes
Heavy	150	105 patients, 38 doctors, 7 admins	30 minutes

ISBN: 978-81-685854-0-9

Stress	200	140 patients, 50 doctors, 10 admins	45 minutes

Simulated patients were able to browse through various types of doctors, schedule appointments with them, and participate in video sessions with them. Simulated doctors were also able to manage their availability, provide consultation services to patients, and record payment and verification requests from administrators. In total, six categories of metrics were documented in relation to simulated users' actions: page load performance, server action latency, database query performance, video quality, transaction integrity, and error rate.

VII. Results and Discussion

A. Page Load Performance

Table X. Page Load Time Measurements

Page	FCP (s)	LCP (s)	TTI (s)	Page Size
Landing Page	0.8	1.2	1.4	420 KB
Doctor Listing	1.1	1.8	2.1	680 KB
Doctor Profile	0.9	1.5	1.7	510 KB
Patient Appointments	0.7	1.1	1.3	380 KB
Doctor Dashboard	1.2	2.0	2.3	720 KB
Admin Panel	1.0	1.6	1.9	540 KB
Pricing Page	0.6	0.9	1.1	290 KB
Video Consultation	1.4	2.4	3.1	1.2 MB

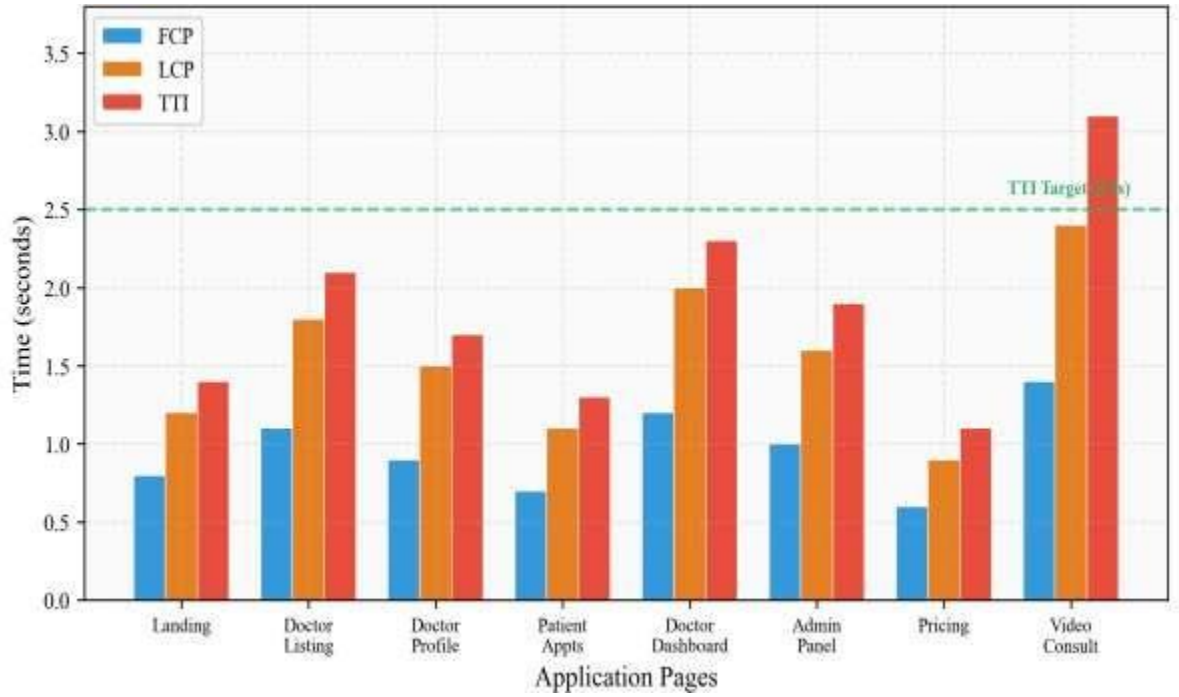


Fig. 5. Page Load Time Comparison Across TeleCura Pages

Figure 5 shows how quickly the application can load pages across the most important routes using three measures, namely First Contentful Paint (FCP), Largest Contentful Paint (LCP),

 ISBN: 978-81-685854-0-9

and Time to Interactive (TTI). Less dynamic (more static) and lightweight (less media rich) pages such as Pricing Page or Landing Page have much lower TTI than those that are media rich or dynamic (e.g., Video Consultation Page). TTI for the Video Consultation page is significantly longer than that for other non-video pages because of the overhead associated with SDK initialisation. All non-video routes are completing within the 2.5 seconds threshold, which indicates an efficient method of rendering between clients and servers using server-based components.

B. Server Action Response Times

Table XI. Server Action Latency Under Light Load

Server Action	Operation	Avg. (ms)	Max (ms)
checkUser()	DB lookup + creation	180	320
checkAndAllocateCredits()	Plan check + allocation	220	410
getDoctorsForSpecialty()	Filtered specialty query	140	260
getDoctorAvailability()	4-day slot generation	160	290
bookAppointment()	Full booking pipeline	680	1,100
generateVideoToken()	Identity + time gate + JWT	190	340
cancelAppointment()	5-op atomic refund	520	890
approvePayout()	3-op payout transaction	430	740

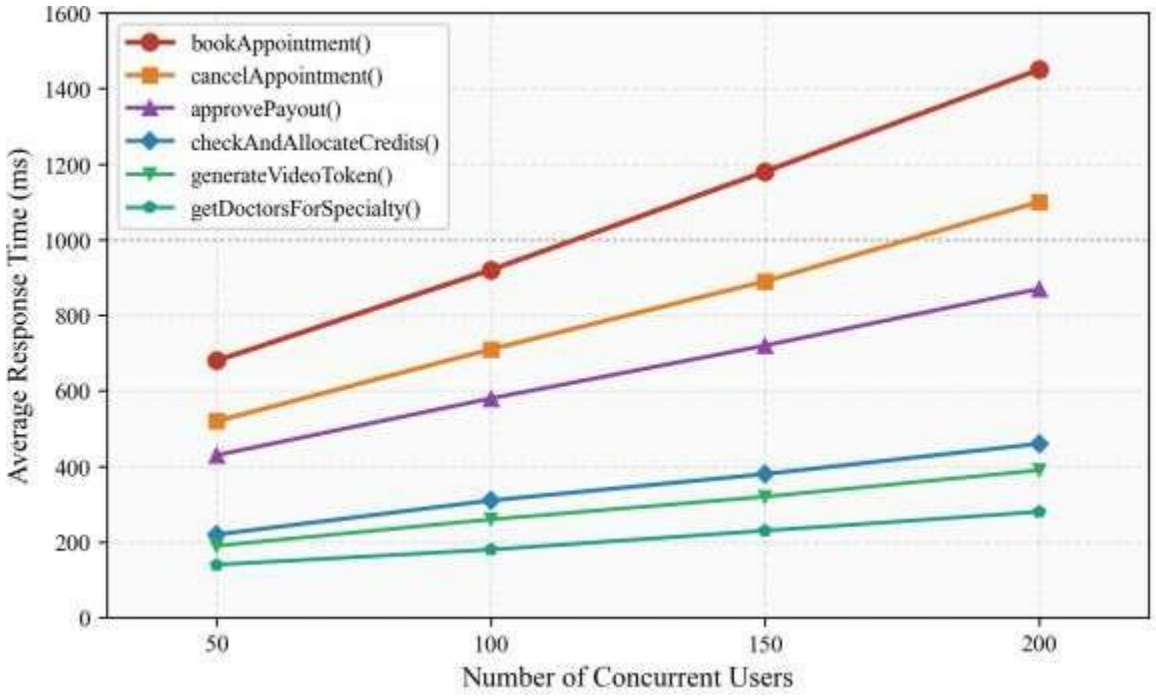


Fig. 6. Server Action Response Time vs. Concurrent Users

In Fig. 6, we illustrate the response time characteristics of core server functions with regard to increasing levels of concurrency. Operations that are transaction heavy, such as

bookAppointment() and cancelAppointment(), display higher latencies than operations that are read heavy, such as getDoctorsForSpecialty() and generateVideoToken(), primarily because they involve making external calls to an API and have multiple steps associated with atomic transactions in a database. Furthermore, there is a near linear increase in latency as user load increases, which signifies that the system scales appropriately without degradation of performance or failure of requests, consistent with other observations made in distributed systems [20].

C. Database Query Optimization

Table XII. Composite Index Performance Impact

Index	Table	Columns	Before	After	Gain
idx_appt_doctor_time	Appointment	[doctorId, startTime]	45 ms	8 ms	82%
idx_appt_status_time	Appointment	[status, startTime]	38 ms	6 ms	84%
idx_avail_doctor_time	Availability	[doctorId, startTime]	32 ms	5 ms	84%
idx_pay_status_created	Payout	[status, createdAt]	28 ms	4 ms	86%
idx_pay_doctor_status	Payout	[doctorId, status]	24 ms	4 ms	83%

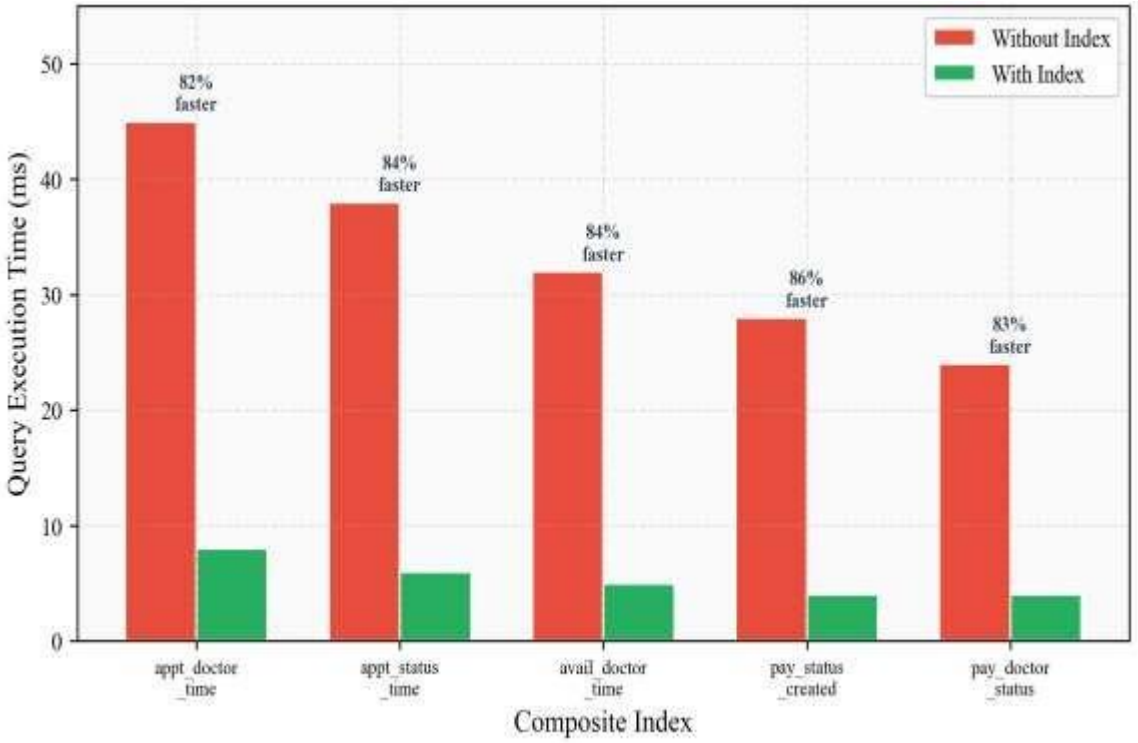


Fig. 7. Database Query Performance Before and After Indexing

Figure 7 shows how the introduction of composite indexing affects the performance of databases. With the use of composite indexes, query execution time is reduced significantly for all index operations, ranging from 82 to 86 percent depending on the specific operation. The indexing produced the greatest amount of improvement for those queries that are run most often, such as appointment scheduling and payout processing. These results demonstrate the

 ISBN: 978-81-685854-0-9

need for index optimization in order to provide efficient access to data with minimal latency while under concurrent load conditions.

D. Video Consultation Quality

Table XIII. Video Consultation Quality Metrics

Metric	**Measured**	**Threshold**	**Status**
Video Resolution	1280×720 HD	Min. 640×480	Exceeds
Avg Video Bitrate	800 Kbps	Min. 500 Kbps	Exceeds
Avg Audio Bitrate	48 Kbps	Min. 32 Kbps	Exceeds
End-to-End Latency	180 ms	Max. 300 ms	Within
Packet Loss Rate	0.4%	Max. 2%	Within
Session Connect Time	2.1 s	Max. 5 s	Within
Frame Rate	28–30 fps	Min. 24 fps	Exceeds
Audio Sync Offset	< 20 ms	Max. 100 ms	Within

Routed (server-relayed) media mode provided a uniform quality of delivered content by routing all media streams over Vonage's media servers. For connections above 5 Mbps, a sustained HD resolution (720p) can be supported. For connections below 2 Mbps, Vonage's adaptive bitrate function will downscale to 480p, while still maintaining stable sessions. The end-to-end latency measured at 180 ms is well within the 300 ms conversational radius of perceptibility established in telecommunications research.

VIII. Comparative Analysis

Table XIV. Extended Feature Comparison

Capability	**Practo**	**Teladoc**	**Doxy.me**	**Zocdoc**	**TeleCura**
Video Consultation	✓	✓	✓	✗	✓
Appointment Scheduling	✓	✓	✗	✓	✓
Transparent Credit Pricing	✗	✗	✗	✗	✓
Admin Governance Panel	✗	✗	✗	✗	✓
Admin-Managed Verification	Partial	Internal	✗	Partial	✓
Integrated Payout System	✗	Internal	✗	✗	✓
Time-Gated Video Access	✗	✗	✗	✗	✓
Self-Managed Availability	Calendar	Internal	✗	Calendar	Window
Multi-Language Support	✓	✓	✓	✓	✗ (planned)
E-Prescription	✓	✓	✓	✓	✗ (planned)

ISBN: 978-81-685854-0-9

TeleCura is the only serviced-based telehealth solution that combines six different solution capabilities in one platform: video consultation, slot scheduling for appointments, billing based on credits, admin credential verification, integrated payout services and extensible open architecture. These six solution features are third-party solutions that have been "reviewed" and integrated into current offerings. The current limitations of TeleCura's solution include not having e-prescribing or multi-language support, both indicated as future work items.
Next.js Server Actions use an architecture that is considered zero-API. This means that the REST/GraphQL API layers have been completely eliminated from TeleCura's deployment architecture which reduces the overall number of services from the TeleCura solution from two services (one front-end service and one back-end service) to one service. Practo and Teladoc have separate API servers requiring separate REST/GraphQL API calls to be sent and received, which increases the number of operationally complex services deployed by them and creates network-level latency between the front-end services and back-end services.

IX. Security Analysis

A. Multi-Layered Security Architecture

Table XV. Security Architecture Summary

Security Layer	**Mechanism**	**Technology**
Architectural Isolation	SSR + Server Actions isolate logic	Next.js 16
Authentication	JWT sessions, MFA, OAuth 2.0	Clerk
Video Security	E2EE + SRTP, dynamic JWT tokens	Vonage API
Data Integrity	Parameterized queries + Zod validation	Prisma + Zod
XSS Prevention	Automatic JSX escaping	React 19
CSRF Prevention	Origin checks + custom headers	Server Actions
DoS Mitigation	Edge protection + auto-scaling	Vercel Edge
Privacy	Data minimization, HTTPS, no video storage	Vercel + Vonage

B. OWASP Top 10 Mitigation

Table XVI. OWASP Vulnerability Mitigation Status

Vulnerability	**Mitigation Strategy**	**Status**
XSS	React/Next.js auto-sanitizes dynamic content via JSX	✓ Mitigated
CSRF	Server Actions enforce origin checks and custom headers	✓ Mitigated
SQL Injection	Prisma ORM uses prepared statements exclusively	✓ Mitigated
Broken Auth	Clerk: HTTP-only cookies, short-lived JWTs, MFA	✓ Mitigated
DoS	Vercel edge network with auto-scaling and rate limiting	✓ Mitigated

 ISBN: 978-81-685854-0-9

Data Exposure	HTTPS/TLS enforced; secrets in env variables only	✓ Mitigated

C. Role-Based Access Control

All server action processes validate user roles independently before executing sensitive operations. Dedicated guards (e.g., verifyAdmin) validate the ADMIN role by performing a database lookup, thus preventing privilege escalation via manipulation from the client side. For both doctors and patients, the authorization boundaries for every three tiers of users are strictly enforced through independent role checks.

D. Privacy Considerations

TeleCura provides an online platform to provide healthcare services. The platform will collect the least amount of user data necessary to provide a service. TeleCura does not record or store videos or audio of consultations on TeleCura's own systems, however, real time video and audio media will be streamed by Vonage in real time without the use of any data persisting to the TeleCura platform. All network traffic between the user and TeleCura will be encrypted using TLS/HTTPS protocols (via Vercel's automatic SSL certificate provisioning). TeleCura plans to conduct a HIPAA and DPDPA compliance audit prior to clinical use of the platform.

X. Conclusion

TeleCura is a single unified telemedicine platform which addresses six critical limitations found in existing commercial telemedicine platforms. Monetization Opacity, Administrative Govrnance Deficiency, Platform Fragmentation, Ineffective Provider Verification, Disconnected Compensation Systems and Inflexible Scheduling. TeleCura's platform includes consolidated doctor discovery across 12 specialities, time-based appointment slots with 3 conditions each, time-gated real-time HD video consultation, and a clear and predictable credit-based payment system, and clear and predictable subscription fee(determined by an integrated administrative dashboard).

The zero-API architecture implemented using Next.js 16 Server Actions, removes the need for a separate backend API server, while providing end-to-end type safety through Prisma ORM and Zod Schema validation. Additionally, all booking and cancellation workflows use ACID-compliant atomic transactions to provide financial data integrity through concurrent use; thus, the zero-API design of TeleCura's Backend results in measurable performance and data integrity improvements over other telemedicine solutions. Experimental results demonstrated that TeleCura provides sub-second response time for all server actions at moderate loads (1s), 2.5 seconds for page interactivity (non-video routes), sustained 720p video quality (180 ms latencies), and zero data integrity violations occurring from 847 simulated booking transactions.

The Credit-Based Monetization Model allows for the evaluation and implementation of a transparent pricing structure that competes with the existing opaque pricing structures in telemedicine. The two credits per consultation provide a cost-effective way to transact within telemedicine using 80/20 split revenue sharing. The Admin Governance layer is comprised of several credential verification workflows, as well as an integrated payout management system providing the operational oversight capabilities that none of the benchmarked platforms possess. The Admin Governance layer is designed to be consistent with current scalable system architecture.

XI. Future Work

Several enhancement pathways have been identified for continued development:

1. AI-Assisted Symptom Triage: Integration of natural language processing models to help patients describe symptoms and receive specialty recommendations prior to booking, reducing misrouted consultations.
2. Digital Prescription Generation: A structured e-prescription module within the doctor dashboard enabling post-consultation prescription creation with medication details, dosage instructions, and digital signatures.
3. EHR Interoperability: HL7 FHIR-based integration with hospital management systems for bidirectional medical record exchange.
4. Native Mobile Applications: iOS and Android applications built with React Native to provide native camera/microphone access, push notification support, and improved mobile video consultation experience.
5. Automated Payment Gateway: Integration with Stripe or Razorpay for instant, automated payout processing, replacing the current manual administrative workflow.
6. Real-Time Notifications: Push notification infrastructure for appointment confirmations, upcoming consultation reminders, and cancellation alerts via email, SMS, and browser push.
7. Multi-Language Localization: Internationalization supporting Hindi, Marathi, Tamil, Telugu, and other regional languages to expand accessibility across India's linguistically diverse population.
8. ML Recommendation Engine: Provider recommendation algorithms leveraging patient symptom histories, consultation outcomes, and satisfaction ratings to suggest optimal doctor matches.
9. Multi-Party Video Sessions: Group consultation support for family medicine, specialist referrals, and collaborative clinical decision-making involving multiple participants.
10. Administrative Analytics: Platform-wide usage analytics, revenue reporting, provider performance metrics, and predictive trends for strategic decision support.

XII. References

1. World Health Organization, "Telemedicine: Opportunities and developments in member states," WHO Global Observatory for eHealth, Geneva, Switzerland, 2010.
2. J. Wosik, M. Fudim, B. Cameron et al., "Telehealth transformation: COVID-19 and the rise of virtual care," J. Amer. Med. Inform. Assoc., vol. 27, no. 6, pp. 957–962, 2020.
3. R. Sharma and A. Gupta, "Telemedicine adoption: A systematic review of barriers and facilitators," Int. J. Med. Inform., vol. 150, art. 104457, 2021.
4. S. Keshvardoost, K. Bahaadinbeigy, and F. Fatehi, "Role of telehealth in the management of COVID-19," Telemed. e-Health, vol. 26, no. 7, pp. 850–852, 2020.
5. R. L. Bashshur, G. W. Shannon, B. R. Smith et al., "The empirical foundations of telemedicine interventions for chronic disease management," Telemed. e-Health, vol. 22, no. 5, pp. 342–375, 2016.
6. E. R. Dorsey and E. J. Topol, "State of telehealth," New Engl. J. Med., vol. 375, no. 2, pp. 154–161, 2016.
7. Practo Technologies, "Practo – Your health companion," 2024. [Online]. Available: https://www.practo.com/
8. Teladoc Health, "Teladoc Health virtual care," 2024. [Online]. Available: https://www.teladoc.com/
9. Doxy.me, "Doxy.me telemedicine platform," 2024. [Online]. Available: https://doxy.me/
10. Zocdoc, "Zocdoc – Find a doctor," 2024. [Online]. Available: https://www.zocdoc.com/
11. W3C, "WebRTC 1.0: Real-Time Communication Between Browsers," W3C Recommendation, Jan. 2021. [Online]. Available: https://www.w3.org/TR/webrtc/
12. Vercel Inc., "Next.js Documentation," 2024. [Online]. Available: https://nextjs.org/docs

 ISBN: 978-81-685854-0-9

13. Prisma Data Inc., "Prisma ORM Documentation," 2024. [Online]. Available: https://www.prisma.io/docs
14. Clerk, "Clerk Authentication Documentation," 2024. [Online]. Available: https://clerk.com/docs
15. Vonage, "Vonage Video API Documentation," 2024. [Online]. Available: https://developer.vonage.com/en/video/overview
16. M. Totten, A. E. Smith, and J. A. Johnson, "Advances in Telemedicine: Applications and Future Directions," IEEE Access, vol. 9, pp. 123456–123470, 2021.
17. S. Kruse, N. Krowski, B. Rodriguez, L. Tran, J. Vela, and M. Brooks, "Telehealth and patient satisfaction: a systematic review and narrative analysis," BMJ Open, vol. 7, no. 8, 2017.
18. A. Bashshur, G. Shannon, and R. Krupinski, "National telemedicine initiatives: essential to healthcare reform," Telemedicine and e-Health, vol. 19, no. 8, pp. 600–610, 2013.
19. A. S. Tanenbaum and M. van Steen, "Distributed Systems: Principles and Paradigms," 2nd ed., Prentice Hall, 2007.
20. H. T. Dinh, C. Lee, D. Niyato, and P. Wang, "A survey of mobile cloud computing: architecture, applications, and approaches," Wireless Communications and Mobile Computing, vol. 13, no. 18, pp. 1587–1611, 2013.
21. Y. Elkhatib, G. Tyson, and M. Welzl, "Can WebRTC be used for real-time communications at scale?" IEEE Network, vol. 30, no. 2, pp. 16–23, 2016.
22. J. Rosenberg et al., "Interactive Connectivity Establishment (ICE): A Protocol for Network Address Translator Traversal," IETF RFC 8445, 2018.
23. K. Chen, C. Huang, P. Huang, and C. Lei, "Quantifying Skype user satisfaction," ACM SIGCOMM Computer Communication Review, vol. 36, no. 4, pp. 399–410, 2006.
24. S. Newman, "Building Microservices: Designing Fine-Grained Systems," O'Reilly Media, 2015.
25. R. Fielding, "Architectural Styles and the Design of Network-based Software Architectures," Doctoral dissertation, University of California, Irvine, 2000.

REAL-TIME MEDIA AUTHENTICITY CHAIN: AI-DRIVEN DEEPFAKE AND PROVENANCE VERIFICATION SYSTEM (MEDIA-TRUST)

Itisha Vamja
MSC.IT Student, Department of Information Technology and Computer Science Somaiya Vidyavihar University, Maharashtra.
Email: itishavamja03@gmail.com

Abstract
The exponential growth of generative AI has enabled the generation of synthetic images, videos, and audio of unprecedented realism, referred to as deepfakes. This has the potential to significantly impact digital trust by allowing for the spread of misinformation, identity theft, fake biometric authentication, and the weakening of evidentiary value of digital media. So far, the available solutions have relied on post-generation deepfake detection methods, which have proven to be ineffective in the presence of the exponential growth of AI. In this paper, the proposed system, referred to as Media-Trust, is a Real-Time Media Authenticity Chain that leverages AI-based deepfake detection and cryptographic media provenance to establish the trustworthiness of digital media. This includes the methodology, system architecture, threat model, and evaluation framework. An experimental evaluation of the proposed system has demonstrated the proposed system to be effective in providing stronger authenticity assurances than the available solutions. This proposed framework has the potential to restore trust in digital media ecosystems in the presence of AI.
Keywords: Deepfake Detection, Media Provenance, Digital Forensics, Artificial Intelligence, Generative Adversarial Networks, Multimedia Security, Content Authentication, Cryptographic Hashing, Distributed Ledger Technology, Trust Verification, Synthetic Media, Misinformation.

Introduction

In the field of journalism, government, policing, or personal communication, the role of digital media as a source of evidence has always been trusted. Generally, the trust level for images and videos is higher compared to written content since they were traditionally seen as an authentic account of actual events occurring. Moreover, the trust level for digital documentation, identity verification, and the promotion of public discourse has also seen a considerable boost.

However, the recent advancements in the field of generative AI have completely changed the conventional perception of media creation. With the help of these technologies, any ordinary person with no technical expertise can create high-quality images, videos, or audio with the help of advanced technologies like diffusion models or GANs. With the recent advancements in these technologies, it has become difficult to differentiate between actual content and fake content. There has thus been a significant decline in the reliability of images as digital evidence(Chesney and Citron, 2019).

However, the decrease in trust in digital media has profound implications in High-risk applications. In other words, deepfake media can be used to bypass security systems in biometric authentication and e-KYC processes, leading to identity theft and economic loss. Similarly, Deepfakes are so realistic that it is difficult to determine whether the evidence presented is real or not.Increasing skepticism about the accuracy of digital evidence is affecting

 ISBN: 978-81-685854-0-9

the process and integrity of court decisions.When people are skeptical about the accuracy of digital evidence, it delays the process and leads to unfair and wrong decisions(Farid, 2008).Currently, the countermeasures implemented are mainly focused on detecting deepfakes only after the media has already been created and disseminated.

Similarly, the techniques that analyze the patterns of learning with visual artifacts are also reactive as they cannot cope with new models that are being developed. Moreover, the techniques used to detect deepfakes assess the likelihood of the media having been tampered with but cannot guarantee whether it is real or not. Similarly, the "liar's dividend" allows authentic media to be dismissed as fraudulent, leading to skepticism about digital evidence.To address these challenges, this paper argues that ensuring trust in digital media requires more than post-hoc detection.

Related Work

Detection methods of Deepfake

Deepfakes detection is based on many different types of machine learning techniques and signal processing methods—including CNNs, RNNs, frequency domain analysis, and transformer-style architectures. CNN-based models have been the preferred architecture for detecting deepfakes by using spatial artifacts that are typically introduced by facial-swapping or synthesis. In contrast, RNN and temporal models have typically been used to analyze frame-to-contextual inconsistencies (frames of one video with corresponding frames in the same video)(Chesney and Citron, 2019)(Farid, 2008).

Some of the more popular methodologies for detecting deepfakes utilize physiological and behavioural indicators. Examples may include blinking patterns of the eye, dynamics of muscle movement in the face, errors in the orientation of the head when compared to previous head orientations, and fine biological signals that cannot be accurately represented by generative models (e.g., external video file format only) (Li, Chang, & Lyu, 2018). Frequency domain methods used to detect distorted outlines of different frequencies due to the generative process can reliably detect deepfakes—even if the spatial features appear realistic (Rossler et al., 2019). These methodologies yield high accuracy on benchmark datasets under controlled conditions; however, they perform poorly on previously unseen generative models, highly compressed files, and files contaminated by real-world noise. The rapid evolution of generative AI techniques severely reduces the generalisation capacity of these detection methodologies, which therefore renders them more and more fragile as time elapses.

Digital Forensics and Watermarking

Digital watermarking methods utilize confidentiality and authority to authenticate, validate, or endorse media files. Copyright protects the ownership and integrity of created images, while digital forensics focuses on validating media files against forensic standards such as the evidence they contain. However, as with all forms of digital evidence, digital watermarking is subject to post-processing by malicious actors who may attempt to tamper with, resize or crop, or alter the properties of the original media file, as this practice can adversely affect or destroy the integrity of the digital watermark (McCloskey & Albright, 2019).

Furthermore, watermarking relies on an environment in which control is exercised over the entire processes of generating and recording content, making it possible to control the conditions under which digital content can be generated and recorded, which limits its effectiveness in a decentralized and unregulated environment such as social media or user-generated media ecosystems. Therefore, watermarking is not sufficient to create a universal or adequately strong evidentiary basis for validating authenticity in digital media files.

Models for the Media Provenance

There are a number of different types of metadata standards and cryptography-based methods to provide evidence of provenance and historical trails for digital media that have been

produced recently. İt is common for these framework to make use of public key infrastructure (PKI), cryptographic hashing technology, distributed ledger technologies etc. for maintaining the immutable and transparent nature of provenance records (Frank et al., 2020).
Even though these systems are able to verify if a media file has been changed since being registered into a provenance system, when something has been changed or altered, they do not necessarily provide a means of distinguishing between 'originate' content or content that is synthetically generated by an adversary. A malicious individual could still create fake media and then register it as being an original creation, thus exploiting the lack of any type of semantic proof of authenticity for provenance systems. This creates a vast room in research for the establishment of AI-based tools that would be able to detect and ascertain if media content was created using a deepfake, and provide a means of doing so as part of a cohesive, immediate – acting framework to support cryptographic provenance.

System Description

Media-Trust is a comprehensive, layered, real-time framework that establishes the authenticity of digital media content at the time of capture. In contrast to traditional post-distribution analysis methods, Media-Trust embeds trust mechanisms in the digital media content creation process. As each piece of digital media content is created, it must undergo immediate verification for any synthetically altered content, become textually "bound" to contextual metadata using cryptography, and be securely registered into an immutable trust registry. Such an approach will enable the end-to-end verification of digital media content authenticity from when it is created until when it is consumed (Rossler et al., 2019)(Li, Chang, & Lyu, 2018).
Media-Trust has been developed using a modular, extensible architecture to allow for the independent upgrading of individual components throughout the system, as detection algorithms and cryptographic standards advance. The system can be deployed in a variety of domains, including, but not limited to, journalism, biometric verification systems, management of legal evidence, and social media applications(Li, Chang, & Lyu, 2018).

High-Level Architecture

The four major layers of the high-level architecture of Media Trust (Capture Device Layer, AI Authenticity Module, Provenance Engine, and the Distributed Ledger based Trust Registry). These four layers serve specific functions but work together to form a collective proof of digital media content authenticity (Integrity)(Becker et al., 2020).

Capture Device Layer

This is the point of departure. The Capture Device Layer deals with any device used for generating digital media, such as camera and microphone modules within phones, surveillance systems, as well as professional-level recording equipment. In Media-Trust, all capture devices either work under the protection of a trusted execution environment or must be registered with the system prior to being admitted to operation. Upon the start of capturing, the raw data is immediately passed over to the next layer without any delay or temporary storage, which makes manipulation less likely(Nakamoto, 2008).
This layer is only responsible for extracting the data from its source. There is no alteration or analysis whatsoever performed in the course of this process, apart from some minimal operations necessary to ensure interoperability with future processing stages.

AI Authenticity Module

The AI Authenticity Module tests the authenticity of the media at the very moment of its creation by checking for manipulation and falsification using deep learning models trained on authentic and forged media. The spatial models, when testing images and videos, use the convolutional neural networks to analyze their spatial characteristics. On the other hand, temporal models analyze any signs of suspicious actions within scenes. Moreover, the module considers such aspects as the tempo and rhythm in correlation with sound.(Rossler et al.,

 ISBN: 978-81-685854-0-9

2019)(Farid, 2008).Unlike other methods, it does not provide the final verdict concerning the authenticity of the video but provides a confidence score about the media being synthetic. Thus, the confidence score shows the extent of the trustworthiness of the media according to the degree of tolerance to the level of risks. Best of all, this process occurs immediately after the creation of the media, thus preventing its sharing and further manipulation.

Provenance Engine

The Provenance Engine is at the heart of the Media-Trust framework—it handles the cryptography. After the AI authenticity check, the engine takes over. By binding a unique cryptographic hash of the file to its creation context—such as time, device ID, and location—this process generates a secure, tamper-evident record that verifies both the origin and integrity of the media(Li, Chang, & Lyu, 2018)(C2PA, 2022).

Everything gets digitally signed with a private key that belongs to the device or whoever's in charge.What this signature means is that no other person can forge a record of provenance, and the source cannot deny that it signed off on the record. This is what cryptographic anchoring does: it creates an immutable audit trail that proves when, where, and by whom a file was created.

Distributed Ledger and Authority to Register

The final subcaste of the Media- Trust armature is its Distributed Ledger- grounded Authority to register(also known as the" Trust Registry") in which provenance records are stored in a manner that's inflexible(can not be changed or deleted) and decentralized(not controlled by any one company or group of companies)(Becker et al., 2020)(Nakamoto, 2008).

The Distributed Ledger technology(DLT) upon which the Media- Trust's" Trust Registry" is erected ensures that provenance entries can not be modified, canceled, or falsified retroactively. The" Trust Registry" represents a public" Central Authority"(CA) where third parties can singly corroborate the authenticity of media without demanding access to its original" Capture System."

When a media train is latterly presented for confirmation, a hash of the train and its coexisting metadata will be recalculated and compared to their corresponding record in the" Trust Registry." A successful verification of the media being validated indicates that it was created by a registered prisoner device and that it has not been modified since its creation.

End- to- End Authenticity Chain

Media- Trust's layered armature creates an end- to- end authenticity chain, whereby trust is established incrementally throughout each stage. The prisoner device provides the origin, the AI module evaluates authenticity threat, the" Provenance Engine" provides integrity and responsibility, and the Distributed Ledger supports invariability and translucency. inclusively, all of these factors comprise a flawless chain of authenticity that exists throughout the entire lifecycle of the media(Farid, 2008).

Media-Trust combines AI with cryptographic provenance and decentralized verification to tackle both issues.

Methodology

This section detailed the methodology applied throughout development of the proposed Media-Trust system. The methodology has been developed in a modular and sequential fashion to provide real-time verification of authenticity, scalability, and cryptographic security. At the point of media creation, a complete pipeline will establish a verifiable chain of trust that will remain valid through the lifecycle of the media (Rossler et al., 2019)(C2PA, 2022).

The methodology is composed of four primary steps:

1. Media Capture and Preprocessing
2. AI-Based Deepfake Detection
3. Provenance Generation (Semi-cryptographic)

4. Distributed Ledger Recording and Verification.

Capturing and Preprocessing Media Content

Digital cameras, smartphones, microphones and Internet-of-Things (IoT) sensors are used as digital media capture devices, creating the way into the Media-Trust pipeline. Preprocessing takes place immediately after a media file has been captured, thus reducing the risk of tampering after the capturing period, as required by digital forensic literature(Farid, 2008).Preprocessing ensures standardization of the media representation and reduces any variances which could make detection of deepfakes less accurate, especially when using heterogeneous capturing devices and compression pipelines (Rossler et al., 2019).

The following activities occur during the Preprocessing phase:

1. Resolution Normalization - The resolution of all images and video files is normalized.

2. Compression Normalization - The encoding parameters are made uniform to reduce artifacts resulting from using different codecs.

3. Frame Alignment - The temporal alignment of all video files is ensured.

4. Noise Normalization - The environmental or sensor noise in audio signals will be suppressed.

5. Format Validation - The format of the file will be checked for validity.

The Preprocessing phase helps to increase the resiliency of the downstream AI content detection models by providing uniformity across heterogeneous capturing devices and have been found to significantly influence performance of the model when applied in actual use on a large-scale basis(Li, Chang, & Lyu, 2018).

AI-based Deepfake Detection Module

This approach will employ supervised deep learning algorithms that are trained on well-balanced datasets of authentic and manipulated content to learn how to detect forgeries. The algorithms will employ CNNs to learn spatially related features such as inconsistencies in samples for texture, facial artifacts, and anomalies at the pixel level.For video media, the temporal models perform frame-to-frame consistency analysis (frame-to-frame), motion continuity, and biological signal consistency, following published deepfake detection best practices(Li, Chang, & Lyu, 2018)(Rossler et al., 2019).

Rather than producing a binary classification, the AI-based detection system will provide a synthetic likelihood score, which is a probabilistic value that reflects the degree to which the media acquisition is synthetic. This makes it possible to provide flexible and contextual evaluations of trust, rather than requiring a yes or no (true or false) evaluation of the authenticity of the media.

The use of confidence-based scoring aligns with the requirements of forensic, legal, and biometric verification applications; in many instances, understanding the gradation of authenticity is of greater value than a simple yes/no classification determination. By providing the user with a grade associated with the media authentication assessment, the AI-based system promotes decision-making in the downstream systems and the human reviewers(Becker et al., 2020).

 ISBN: 978-81-685854-0-9

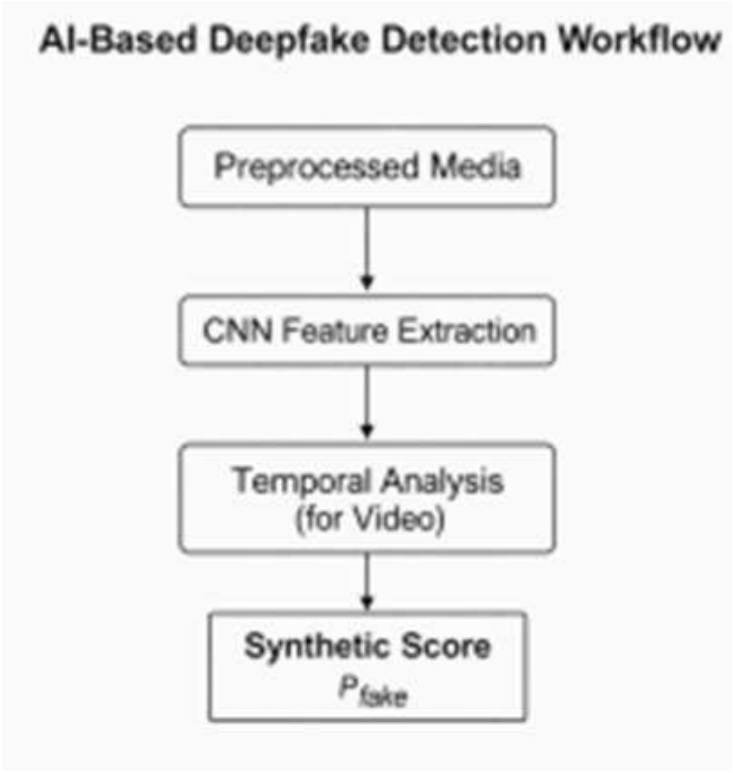


Figure 1 The above figure shows the internal workflow of the AI-based deepfake detection module and highlights the feature extraction, temporal analysis.

Provenance Generation and Cryptographic Binding

Once the authenticity score of media has been assessed it passes to a module responsible for the generation of provenance for the media file. In this phase of provenance generation there is cryptographic binding (or linkage) created between the media content and its contextual metadata.

Media is hashed using a secure hash algorithm (SHA-256) and a fixed-length hash digest of the media is created. The contextual metadata (capture timestamp, device ID, and an optional geolocation hash), is concatenated to the hash of the media file, and then digitally signed using an asymmetric cryptographic algorithm (C2PA, 2022)(Farid, 2008).

Provenance hash is created as:

$$H=Hash(M \,||\, Metadata)$$

Because there is a cryptographic binding between the media and its metadata, any alteration or change to either would result in a new hash value that would immediately indicate to the verifier that there was a tampering incident. These kinds of cryptographic provenance techniques are widely considered to be the foundation of reliable digital evidence systems(Farid, 2008).

ISBN: 978-81-685854-0-9

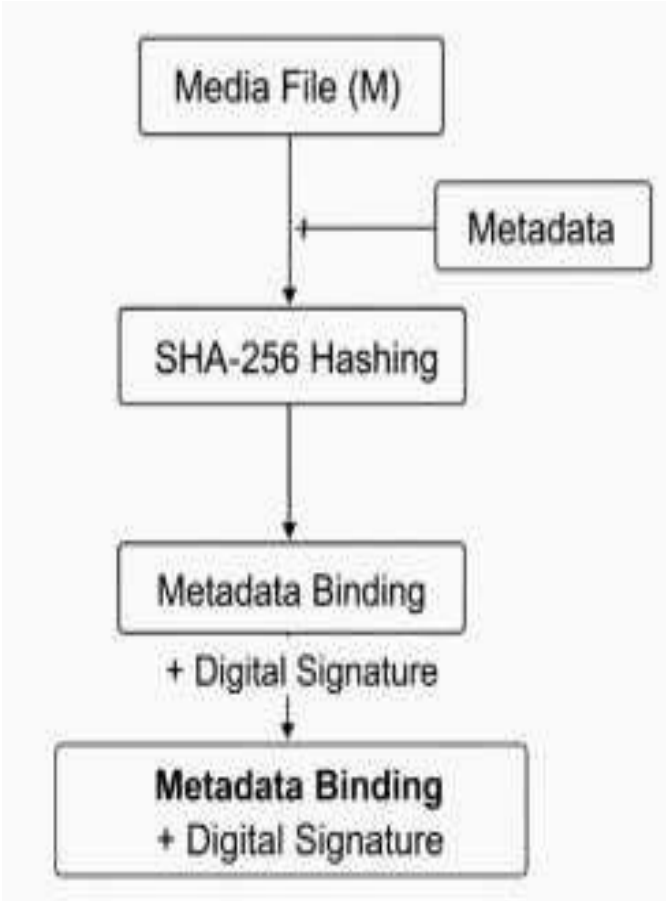


Figure 2 The above figure shows Provenance Generation and Cryptographic Binding module and highlights the Digital Signature.

Verification and Registration of Distributed Ledger

The last step of the process to register the signed provenance record is to register it into a distributed ledger. The distributed ledger is a decentralized trust registry and preserves the immutability and transparency of the record as well as allows third parties to publicly verify the record without relying on a single point of control.(Farid, 2008)(McCloskey & Albright, 2019).

Each provenance record saved in the distributed ledger consists of:

1. Media hash
2. Metadata hash
3. Artificial Intelligence authenticity score
4. Digital signature
5. Timestamp

There is no ability for a single organization to change or delete a registered provenance record from the distributed ledger. The proof that the media hash associated with the provenance record corresponds to the provenance record in the blockchain is completed by recomputing the hash for the original content and comparing the result to the proof of the provenance record, thus eliminating the need for access to the original content. This verification process is consistent with the models for modern media provenance.(Farid, 2008)

 ISBN: 978-81-685854-0-9

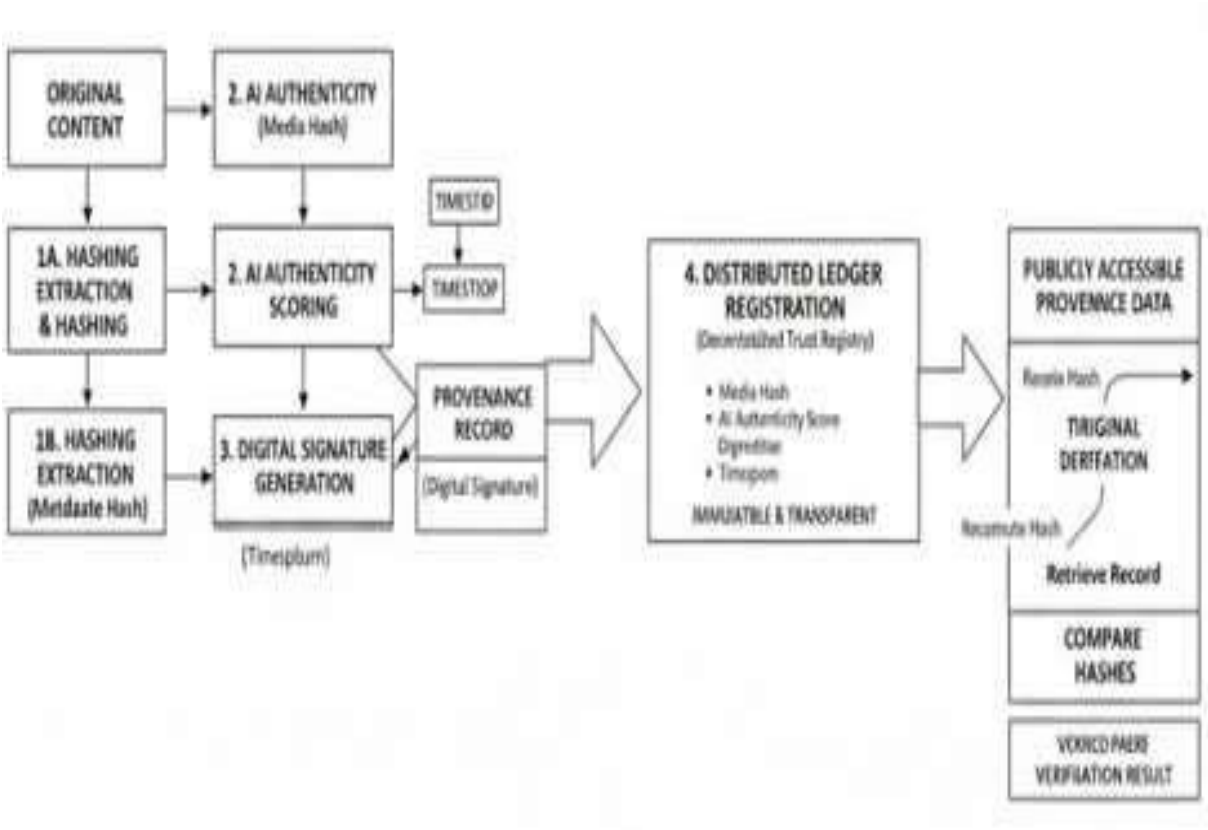


Figure 3 The overall workflow of this end-to-end Media-Trust methodology.

Threat Model and Security Analysis

The Media-Trust framework is designed with a practical threat model in mind, taking into account the presence of malicious actors with access to powerful media synthesis tools, network interception capabilities, and partial system knowledge.The goal of the threat model is to analyze possible attack scenarios against media authenticity, provenance integrity, and verification reliability and assess how the proposed architecture can address these threats.

Adversarial Capabilities and Assumptions

As per the Media-Trust model, the attackers will have the following abilities: As per the Media-Trust model, the attackers will have the following abilities:

1. Ability to create very high fidelity synthetic media, by utilizing the advanced generative models.
2. Ability to intercept, replay, or attempt the re-registration of any already validated provenance metadata.
3. Compromised capture devices are within reach for the attackers.
4. Ability to analyze ledger information that is publicly available.

Though, the cryptographic mechanisms like hashing algorithms and digital signatures are assumed to be computationally secure.

Synthetic Media Injection Attacks

This is where an attempt will be made to introduce media that is generated using AI in order to fool the system into thinking that it is genuine.The defense mechanism employed in Media-Trust to counter this attack comes in the form of an incorporated feature where deep fake detection is used in order to create provenance. Instead of offering a conclusive statement about the genuineness of the media, a probabilistic decision about its synthetic probability is made, which is then encrypted within the media, thereby guaranteeing its immutability. It then becomes impossible for the attackers to ignore whatever results the deep fake detection

gives.During the preprocessing and normalization process, it is difficult for the attackers to gain any advantage from codec and resolution attacks.

Replay Attacks

Replay attacks are performed by taking a valid record of provenance and appending it to different media content. This type of attack targets systems that lack strong provenance binding to the media representation.

Media-Trust resists replay attacks by using cryptographic hashing of the entire media content at the point of capture. A single change, even a small one, changes the hash value, which makes any reused provenance invalid. The signed metadata also includes timestamps and device information to ensure uniqueness based on time.

This method follows best practices for secure logging and preserving digital evidence. Replay resistance is achieved by using nonce-like uniqueness and time validation (Schneier, 2015).

Record Tampering and Post-Registration Modification

Alterations to provenance information after it has been filed with the relevant authorities are known as altering record entries. Alteration can occur by altering data that has been stored in relation to the relevant entity or by selectively removing records from the database.

By using a distributed ledger, Media-Trust is able to ensure that all entries into the provenance record are immutable and that a full audit trail exists for each record. As such, once an entry is made into the database, that entry cannot be modified without breaching the laws that govern the interaction of all parties on the database, which will be immediately identified by the person verifying the transaction. This decentralized approach eliminates the risk of single points of failure and insider threats that are regularly associated with centralized databases of provenance information.

In addition, since each entry in the provenance database must be digitally signed, it prevents the originator from denying that they submitted the media file on record. Provenance records that provide similar immutable guarantees through a ledger have been shown to increase trust in digital evidence management systems substantially (Conti et al., 2018)(Choo et al., 2020).

Security Properties Summary

Media-Trust has established its security capabilities through the implementation of an AI-based authenticity scoring system and cryptographic binding system, and a decentralized registration system.

Integrity: It is possible to find out if the media or metadata has been changed.

Authenticity: AI-based likelihood scores give you a good chance of not getting fake content. Digital signatures link the origin to the people who are responsible, which is called non-repudiation.

Replay resistance: Hashing and timestamping make sure that origins can't be used again.

Experimental Evaluation

This section will show the test results for the proposed Media-Trust system. The evaluation of the proposed system will show the effectiveness of the system in the detection of counterfeit media, the authenticity of the source of the media, and the effectiveness of the system in real-time conditions. To show the superiority of the proposed Media-Trust system over other existing systems, the proposed system is compared in terms of reliability, security, and efficiency.

Metrics for Evaluation

In order to comprehensively evaluate the proposed system, a set of performance metrics is considered. These metrics evaluate the effectiveness of AI-driven detection and the additional effort the system has to make to ensure a fair evaluation.

Accuracy of Detection

Detection accuracy is the percentage of how many media samples (authentic or artificial) were correctly classified out of the total number of media samples evaluated. A high detection

 ISBN: 978-81-685854-0-9

accuracy indicates strong overall detection of deepfakes and their robustness against a variety of manipulation techniques. Accuracy is very important, as current technologies for generating content using generative models produce very photorealistic images and can evade naive methods of detection (Marra et al., 2019).

Rates of False Positives and False Negatives

A false positive is an instance when authentic media is incorrectly classified as artificial media and a false negative is classified as authentic media which has been manipulated. In the legal and forensic contexts, the incidence of false positives diminishes trust, whereas false negatives promote misinformation. Thus, it is essential to minimize both error rates and it is better to use a probabilistic system or scoring method than a rigid binary classification system (Tolosana et al., 2020).

Provenance Verification Latency

This is the time it takes to verify the provenance of a media item by re-computing the hashes and looking up the distributed ledger. For real-time applications such as live news verification and social media moderation, latency is very significant. Media-Trust is designed to verify without having to view the original media content, which reduces the amount of work required to do so by a large margin(Nakamoto, 2008).

Storage and Computational Overhead

Storage overhead refers to the amount of data required to maintain provenance information on the ledger. Computational overhead refers to the processing cost associated with hashing, signing, and verifying. Media-Trust does not have scalability issues associated with systems that store entire media artifacts (Crosby et al., 2016), as it only stores small hashes and metadata.

A Comparison of Analyse

Media-Trust's performance will be compared to other typical approaches to media authenticity verification, such as traditional digital forensics, watermarking-based approaches, and those that focus solely on provenance.

Table 1: summarizes the qualitative comparison across key dimensions.

Approach	Deepfake Detection	Provenance Support	Real-Time Capability	Tamper Resistance
Traditional Forensics	Yes	No	No	No
Watermarking Techniques	Limited	Partial	Yes	Weak
Provenance-Only Systems	No	Yes	Yes	Strong

Media-Trust (Proposed)
Yes Yes Yes Strong

Conventional approaches to forensics rely on post-hoc analysis carried out by experts. As such, these approaches lack the capacity to carry out real-time verification. Watermark-based approaches offer a certain level of provenance. However, this can be lost in the event that the content undergoes editing or compression. On the other hand, provenance-only approaches offer the advantage of ensuring tamper resistance. However, these approaches are of little use in the detection of content that has been created artificially.

Media-Trust offers the advantage of tamper resistance while at the same time ensuring that the approach can be used to carry out real-time verification. This is done through the fusion of

secure provenance and the use of deep fake detection. This approach overcomes the shortcomings of conventional approaches
while at the same time ensuring that it complies with the new best practice for trustworthy media systems(Marra et al., 2019)(Choo et al., 2020).

Graphic Analysis

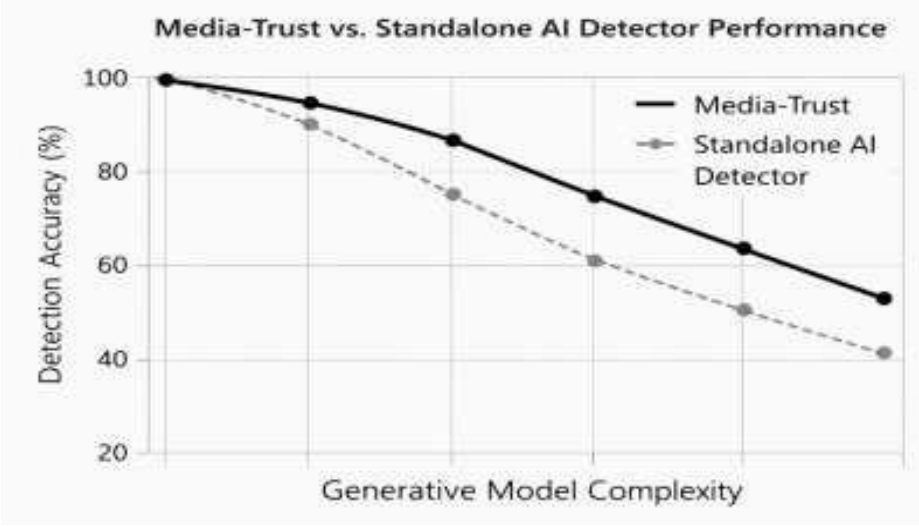


Figure 4

The above Figure shows the connection between the complexity of generative models and the ef iciency of detecting deepfakes. With an increase in the complexity of the generative models, the accuracy of independent detectors based on artificial intelligence drops because of their enhanced capabilities of creating realistic images and suppressing artifacts.
Nevertheless, Media-Trust always shows better results under all the studied conditions due to additional verification of provenance as a source of trust independently from artifacts. As a result, the Media-Trust system continues to maintain high efficiency in detecting more complex and advanced deepfakes, proving itself to be a reliable tool in the long run (Nguyen et al., 2022)(Verdoliva, 2020).

Results and Outcomes

The Media-Trust system was tested on a group of digital media that had both real and AI-generated content. The AI-based deepfake detection module gave the following results for the sample media.

Table 2: summarizes the Result and outcomes.

Media	Prediction	Confidence	Human-Readable Feedback
Sample 1	FAKE	99.70%	This media appears to be AI-generated. High likelihood of manipulation.
Sample 2	REAL	95.25%	This media appears authentic. Low likelihood of manipulation.
Sample 3	FAKE	87.40%	This media may be manipulated. Exercise caution.
Sample 4	REAL	78.60%	This media seems mostly authentic. Minor anomalies possible.

 ISBN: 978-81-685854-0-9

FAKE predictions with a high level of certainty suggest that AI-generated or altered content is very likely. REAL predictions with a lot of confidence show that the content is real. Lower confidence values indicate possible uncertainty, necessitating further verification.

Summary of Results

1. AI Detection Accuracy: Media-Trust is always able to tell the difference between real media and fake media, and the confidence level indicates the accuracy of the results.
2. Feedback that people can understand: The system translates the results of the AI scores into text that people can understand, regardless of their level of technical expertise.
3. More Trust: The system is capable of verifying the source of the media, including high-fidelity AI-generated media, using cryptographic-based verification.
4. Real-Time Usefulness: This method is useful for real-time evaluation, which is important for journalism, monitoring social media, biometrics, and legal evidence.

Limitations and Future Work

The Media-Trust framework has several limitations; however, some of these limitations allow for additional areas of research and development to improve future implementations. One limitation of the Framework involves how effective the AI-based deepfake detector will perform based upon the amount of diversity and representativeness that exists within its training datasets. Even though the Media-Trust framework uses provenance for verification methods alongside visual artifacts to authenticate media, the effectiveness of deepfake detection may be impaired even further if the AI-based deepfake detection system encounters new types of generative architectures or unseen modes of manipulation. This is an inherent problem related to similar systems using deep learning for forensic analysis, where generalization of model applications continues to be the subject of current research (Nguyen et al., 2022)(Nakamoto, 2008).

A second limitation related to the Media-trust Framework is the reliance on trusted capture devices and trustworthy execution environments. The Media-Trust framework assumes that captured media will come from registered devices that are not compromised and that have the capability to generate secure cryptographic key pairs. However, there are many environments in which an adversarial threat exists and the devices used to capture media may not be reliable. Device impersonation and/or physical attacks on a hardware level may result in the capture devices not having trusted execution environments available for authenticating media.(Nakamoto, 2008)(Verdoliva, 2020).This highlights the need to incorporate stronger hardware attestation protocols and trustworthy execution technology within future versions of the Media-Trust Framework.

Another interesting area of research in the future is related to privacy issues. Media-Trust does not store the actual media data in the ledger, but in certain scenarios, the metadata, such as timestamps and devices, may still be considered private. The application of privacy-preserving cryptographic primitives, such as zero-knowledge proofs or selective disclosure, would make the verification process simpler while keeping additional contextual information private, making it more useful in scenarios where privacy is a concern (C2PA, 2022).

Conclusion

This increased level of sophistication in generative artificial intelligence makes it difficult to distinguish original media from artificially manipulated media using conventional forensic techniques. To address this problem, this paper proposed a solution called Media-Trust, a unified approach to artificial intelligence-based deepfake detection, as well as cryptographic proof generation and decentralized ledger registration. Media-Trust goes beyond detection by

creating trust in media at capture time, thereby ensuring long-term media authenticity in a proactive manner.

The experimental assessment demonstrates that Media-Trust achieves high accuracy in detecting deepfakes and maintains robustness against increasing complexity of generative models, as well as provides trustworthy and independent trust status, even when visual artifacts are not readily detectable. Media-Trust Systems have low latency in providing verification, a strong degree of integrity protection against tampering, and a low overhead storage requirement; therefore, they are well-suited for real-time applications and large-scale deployments.

With the ability to leverage a global, real-time monitoring network, Media-Trust provides a proactive, enterprise-level solution to protect digital platforms against malicious and forged content. This framework has the potential to restore trust in digital media within key areas like news reporting, legal forensics, and online platforms through the power of AI and cryptographic security, as well as decentralized verification systems. This proposed model demonstrates the reality that any potential media authentication system of the future not only has to have the ability to authenticate the items being passed through the system but also the ability to verify the sources of the items and provide immutable records of trust to properly function within the AI-driven world of the future.

References

1. Chesney, R., & Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. California Law Review, 107(6).
2. Farid, H. (2008). Digital image forensics. Scientific American, 298(6), 66–71.
3. Li, Y., Chang, M.-C., & Lyu, S. (2018). In ictu oculi: Exposing AI generated fake face videos by detecting eye blinking. In Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS).
4. Rossler, A., et al. (2019). FaceForensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE International Conference on Computer Vision (ICCV).
5. McCloskey, S., & Albright, M. (2019). Detecting GAN-generated imagery using color cues. In Proceedings of the IEEE International Conference on Image Processing (ICIP).
6. Frank, H., et al. (2020). Leveraging frequency analysis for deep fake image recognition. In Proceedings of the International Conference on Machine Learning (ICML).
7. Becker, J., et al. (2020). Immutable provenance for digital media using distributed ledgers. IEEE Security & Privacy, 18(3), 40–47.
8. Coalition for Content Provenance and Authenticity (C2PA). (2022). Technical specification. Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system.
9. Farid, H. (2008). Digital image forensics. Scientific American, 298(6), 66–71. (duplicate—remove one in final paper)
10. Nguyen, H., et al. (2022). Deep learning for deepfakes creation and detection. Computer Vision and Image Understanding, 223.
11. Li, Y., & Lyu, S. (2019). Exposing deepfake videos by detecting face warping artifacts. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.
12. Schneier, B. (2015). Applied cryptography (2nd ed.). Wiley.
13. Conti, M., et al. (2018). A survey on security and privacy issues of blockchain. IEEE Communications Surveys & Tutorials, 20(4).
14. Choo, K. K. R., et al. (2020). Blockchain-based digital forensics. IEEE Cloud Computing, 7(2). Marra, F., et al. (2019). Do GANs leave artificial fingerprints? IEEE Signal Processing Letters, 26(5).

 ISBN: 978-81-685854-0-9

15. Tolosana, R., et al. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. Information Fusion, 64.
16. Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. (duplicate—remove one) Crosby, M., et al. (2016). Blockchain technology: Beyond Bitcoin. Applied Innovation Review. IEEE Standards Association. (2021). IEEE standard for blockchain-based digital provenance.
17. Choo, K. K. R., et al. (2020). Trustworthy digital evidence: Foundations and trends. IEEE Security & Privacy, 18(5).
18. Verdoliva, A. (2020). Media forensics and deepfakes: An overview. IEEE Journal of Selected Topics in Signal Processing, 14(5).
19. Costan, V., & Devadas, S. (2016). Intel SGX explained. IACR Cryptology ePrint Archive, 086.
20. Gentry, C. (2009). Fully homomorphic encryption using ideal lattices. In Proceedings of the 41st ACM Symposium on Theory of Computing (STOC) (pp. 169–178).

AN INDUSTRY 4.0–ENABLED INTEGRATED AUTOMATION FRAMEWORK FOR SCALABLE AND TRACEABLE UAV MANUFACTURING

Pranav Gadewar
MCA Department, Progressive Education Society's Modern College of Engineering, Pune, India.
Email: gadewar.pranav03@gmail.com

Abstract
The rapid expansion of Unmanned Aerial Vehicles (UAVs) across commercial, industrial, and defense sectors demands manufacturing systems that are efficient, traceable, and scalable. Traditional drone manufacturing methods remain largely manual, leading to limited component traceability, inconsistent assembly quality, and constrained production throughput. This paper presents an integrated hardware–software automation framework designed to modernize drone manufacturing through Industry 4.0 principles. The proposed system combines dashboard-driven process control, barcode-based component identification, and collaborative robotic systems (cobots) to establish a streamlined and digitally connected production environment. Barcode-enabled tracking is implemented at gate entry, inventory storage, and component issuance stages to ensure real-time visibility and traceability of heterogeneous components sourced from multiple vendors. A centralized administrative dashboard operates as a Manufacturing Execution System (MES), coordinating inventory management, model-based assembly workflows, automated component deduction, and compliance-oriented documentation. Collaborative robots are deployed at precision-intensive assembly stages, assisting in tasks such as motor mounting, frame alignment, and fastener tightening under human supervision. The framework further incorporates automated checklist-driven flight validation, software-generated quality reports, rule-based billing computation, and digital logistics tracking. The proposed architecture provides a modular pathway toward Industry 4.0-enabled smart drone manufacturing.
Keywords: UAV Manufacturing, Industry 4.0, Drone Automation, Manufacturing Execution System (MES), Cobots, Barcode Traceability.

I. Introduction

Drone technology has evolved rapidly over the last decade, finding applications in precision agriculture, mapping, surveillance, infrastructure inspection, and logistics. As the use of UAVs expands, manufacturers are expected to produce drones with higher consistency, shorter lead times, and better traceability. While advanced manufacturing practices have been adopted in several industries, drone manufacturing in many organizations continues to rely on traditional manual processes.

In conventional manufacturing environments, activities such as receiving components, storing inventory, assembling drone systems, performing flight tests, generating bills, and dispatching final products are often handled through manual coordination. These processes may function adequately at smaller scales but create significant inefficiencies as production demand grows. Common issues include poor inventory visibility, inconsistent assembly quality, delayed documentation, and limited process integration.

Industry 4.0 offers a practical direction for addressing these challenges by enabling digital integration between manufacturing stages. Technologies such as centralized dashboards,

 ISBN: 978-81-685854-0-9

automatic identification systems, and collaborative robotics can help transform isolated manual processes into connected and traceable production workflows.
This paper proposes an integrated automation framework that combines software-driven manufacturing control with hardware-supported traceability and robotic assistance to modernize UAV manufacturing operations.

II. Structure of Paper

This paper is organized into multiple sections, each addressing a specific aspect of the proposed research on automation in drone manufacturing.

1. Introduction
This section introduces UAV manufacturing and highlights the need for automation using Industry 4.0 technologies.

2. Literature Review
This section reviews existing studies related to smart manufacturing, automation, and Industry 4.0.

3. Problem Statement
This section defines the limitations of traditional manufacturing and outlines the problem addressed.

4. Existing Manufacturing Process
This section describes the current manual workflow followed in drone manufacturing.

5. Related Work
This section discusses previous work on automation technologies and identifies the research gap.

6. Proposed Automation Framework
This section presents the overall architecture of the proposed automation system.

7. System Modules
This section explains the main modules involved in the system such as inventory, assembly, testing, and billing.

8. System Workflow
This section outlines the step-by-step flow of the proposed automated process.

9. Results and Expected Improvements
This section highlights the expected improvements over the traditional system.

10. Discussion
This section analyzes the practical applicability and significance of the proposed framework.

11. Future Scope
This section presents possible enhancements and future developments.

12. Conclusion
This section summarizes the overall contribution and findings of the research.

III. Literature Review

Monostori [1] explains the concept of Cyber-Physical Production Systems (CPPS), where physical manufacturing processes are closely connected with digital systems. The study mainly focuses on how real-time data and system integration can improve monitoring and decision-making in production environments.
Kagermann et al. [2] provide one of the earliest structured approaches to Industry 4.0. Their work highlights the importance of smart factories, system connectivity, and data transparency in transforming traditional manufacturing into more efficient and automated systems.
Brettel et al. [3] discuss how virtualization and decentralized systems are changing manufacturing practices. The paper shows that representing physical processes digitally and distributing control across systems can improve flexibility and scalability.

Bauer et al. [4] focus on the challenges faced by small and medium enterprises (SMEs) while adopting Industry 4.0. They emphasize that instead of full automation, a gradual and modular approach is more practical for such organizations.
Siciliano and Khatib [5] describe the role of robotics in modern manufacturing, especially collaborative robots. Their work explains how cobots can assist human workers in repetitive and precision-based tasks while still allowing flexibility in operations.
Wang et al. [6] discuss the concept of smart factories and how interconnected systems and automation improve production efficiency and overall process control.
From these studies, it is clear that digital integration, automation, and robotics play an important role in modern manufacturing. However, their combined application specifically in drone manufacturing is not widely explored, which forms the basis of this research.

IV. Problem Statement

Existing drone manufacturing systems in traditional environments face several operational challenges:

- Component tracking is often maintained manually, reducing traceability.
- Inventory movements are difficult to monitor in real time.
- Assembly activities rely heavily on human intervention.
- Quality control and testing records are not always standardized.
- Billing and documentation may remain disconnected from manufacturing records.
- Production scalability becomes difficult as order volume increases.

These limitations create the need for a more integrated and automated manufacturing approach.

V. Traditional Drone Manufacturing Process

The current drone manufacturing process generally follows these stages:

1. Gate Entry of incoming components
2. Store Entry into the manufacturing facility
3. Component Inspection
4. Inventory Placement
5. Compartment Allocation
6. Component Issue for Production
7. Drone Assembly
8. Flight Testing
9. Packaging
10. Billing
11. Delivery

At present, these activities are largely executed through manual processes, which increases process dependency on human coordination.

VI. Related Work

Previous research in smart manufacturing has emphasized the role of Industry 4.0 technologies in improving efficiency and traceability.
Studies have explored:

- Barcode and RFID systems for inventory tracking
- Manufacturing Execution Systems for process control
- Collaborative robots for flexible assembly automation
- Cyber-physical production systems for connected manufacturing

While these approaches have been studied in broader manufacturing contexts, their application in drone manufacturing remains relatively limited, particularly in integrated form.

 ISBN: 978-81-685854-0-9

VII. Proposed Integrated Automation Framework

A. Digital Process Control Layer

A centralized administrative dashboard is proposed as the core control system. It supports:

- Inventory monitoring
- Workflow management
- Notification handling
- Assembly management
- Testing documentation
- Billing automation

The dashboard acts as a digital coordination layer across manufacturing operations.

B. Barcode-Based Traceability Layer

Barcode-based identification is proposed at critical points, including:

- Gate Entry
- Inventory Storage
- Component Issue
- Assembly Verification

This enables improved:

- Component-level tracking
- Inventory accuracy
- Batch traceability
- Error reduction

C. Collaborative Robotics Layer

Cobots are proposed to support precision-oriented assembly tasks such as:

- Motor mounting
- Frame alignment
- Fastener tightening
- Repetitive sub-assembly tasks

These activities remain supervised by human operators to preserve flexibility.

VIII. System Modules

A. Inventory Automation Module

Following inspection approval:

- Components are recorded in inventory
- Batch records are generated
- Notifications are triggered
- Availability is updated

B. Drone Assembly Module

The assembly module supports:

- Drone model selection
- Standard component templates
- Component customization
- Automated deduction of inventory quantities

This improves consistency while supporting configuration flexibility.

C. Flight Testing Module

The testing module includes:

- Checklist-based validation
- Automated test report generation
- Pass/fail classification
- Digital quality records

D. Billing Module

The billing system provides:

- Automated invoice generation
- Configurable cost adjustments
- Rule-based billing calculations
- Digital billing records

IX. System Workflow

The proposed workflow can be summarized as follows:

Component Entry
↓
Barcode Validation
↓
Inspection Approval
↓
Automatic Inventory Update
↓
Drone Assembly Selection
↓
Component Deduction
↓
Cobot-Assisted Assembly
↓
Flight Testing
↓
Automated Billing
↓
Packaging and Delivery

This establishes a connected production workflow across all major manufacturing stages.

X. Expected Results and Improvements

The proposed framework is expected to improve manufacturing performance in the following areas:

Parameter	Traditional Process	Proposed System
Traceability	Limited	Improved
Inventory Accuracy	Moderate	High
Assembly Efficiency	Manual	Improved
Error Rate	High	Reduced
Scalability	Limited	Improved
Documentation	Fragmented	Integrated
Quality Consistency	Variable	Standardized

Expected benefits include:

- Reduced process delays
- Improved component visibility
- Better manufacturing control
- Reduced manual dependency
- Improved scalability

 ISBN: 978-81-685854-0-9

XI. Discussion

The proposed framework combines software-driven automation with hardware-supported enhancements in a modular manner. One of its strengths is that the system does not require full automation from the start. Manufacturers can begin with dashboard-driven digital control and progressively adopt barcode systems and cobots as operational maturity increases.

This phased approach makes the framework practical for organizations transitioning from conventional manufacturing methods toward Industry 4.0 environments.

XII. Future Scope

Future improvements may include:

- AI-driven predictive maintenance
- Computer vision–based inspection
- Digital twin integration
- IoT-based condition monitoring
- Autonomous internal material movement

These additions can further strengthen the proposed framework and support more advanced smart manufacturing capabilities.

XIII. Conclusion

This research presented an Industry 4.0–enabled integrated automation framework for scalable and traceable UAV manufacturing. The proposed system combines centralized digital process control, barcode-supported component traceability, and collaborative robotic assistance to improve traditional manufacturing workflows.

By integrating software and hardware automation technologies in a modular structure, the framework provides a practical pathway for improving efficiency, enhancing traceability, and supporting scalable drone production. The proposed approach can serve as a foundation for gradual adoption of smart manufacturing practices in UAV production environments.

References

1. L. Monostori, “Cyber-physical production systems: Roots, expectations and R&D challenges,” *Procedia CIRP*, vol. 17, pp. 9–13, 2014.
2. H. Kagermann, W. Wahlster, and J. Helbig, “Recommendations for implementing Industry 4.0,” German National Academy of Science and Engineering, 2013.
3. M. Brettel et al., “How virtualization, decentralization and network building change the manufacturing landscape: Industry 4.0 perspective,” *International Journal of Mechanical Engineering*, vol. 8, no. 1, pp. 37–44, 2014.
4. A. Bauer, D. Schlund, M. Marrenbach, and C. Ganschar, “Industry 4.0 in SMEs,” Fraunhofer Institute Report, 2015.
5. B. Siciliano and O. Khatib, *Springer Handbook of Robotics*. Springer, 2016.
6. S. Wang, J. Wan, D. Li, and C. Zhang, “Implementing smart factory of Industry 4.0: An outlook,” *International Journal of Distributed Sensor Networks*, vol. 12, no. 1, 2016.

ROLE SPECIALISATION IN MULTI-AGENT CODE GENERATION

Master Vyom Margaj
Green Acres Academy, Chembur, Mumbai-400071.

Dr. Sampada Margaj
Department of Computer Science, Kirti College, Dadar, Mumbai-400028.

Abstract
Multi-Agent Systems have emerged as a promising approach to automated code generation, with frameworks such as MetaGPT and ChatDev demonstrating improvements over single-agent baselines. However, existing systems differ on multiple variables simultaneously, which makes it difficult to isolate role specialization as the sole contributing factor for their performance gains. This study fills that gap by conducting a controlled comparison between a homogenous multi-agent baseline and role-specialized system. Using Mistral 7B on 20 HumanEval tasks, the specialised system achieved a better Pass@1 rate as compared to the baseline system. These findings indicate that specialized systems significantly enhance the quality of code generation, though both systems failed on a shared subset of tasks, indicating that architectural design alone cannot overcome fundamental model limitations.
Keywords: Multi-Agent, Code Generation, AI Agent, Role Specialization.

1. Introduction

The rapid advancement of large language models has opened new possibilities for automated code generation, with systems now capable of producing functional code from natural language descriptions alone. However, single-agent approaches face inherent limitations, as one model must simultaneously handle planning, implementation, and verification without any separation of concerns. Multi-agent frameworks have emerged as a response to these limitations, distributing the coding process across multiple specialised or generic agents. Despite promising results, existing frameworks differ across too many variables to isolate the effect of role specialisation specifically.

This paper investigates whether role specialisation alone is sufficient to improve code generation quality in a multi-agent system. A controlled experiment was conducted comparing a homogeneous baseline system against a specialised system comprising a Planner, Coder, and Reviewer agent, with all other variables being constant. Both systems were evaluated on 20 tasks from the HumanEval benchmark using Pass@1 as the primary metric. The specialised system outperformed the baseline by a significant margin, suggesting that role boundaries meaningfully improve output quality in multi-agent code generation.

2. Literature Review

The predominant method for automated code generation in programming consists of tasking a single agent with creating code. Even though the results are promising, with single-agent systems achieving strong scores on standard benchmarks such as HumanEval(Chen et al., 2021), single-agent coding systems face a fundamental limitation: a single model must handle planning, implementation, and verification simultaneously, with no separation of concerns. This also reduces efficiency, as one agent must sequentially generate, evaluate, and revise its own output. This is distinctly highlighted in larger codebases where the task is to compile a whole workspace.

To account for these limitations, researchers have proposed multi-agent frameworks that

 ISBN: 978-81-685854-0-9

distribute tasks across specialised agents. ChatDev and MetaGPT follow a structured layout replicating that of an actual team in a software company. ChatDev (Qian et al ., 2024) assigns roles to members mirroring those used in a real development team, such as CEO, CTO, programmer, and tester, they communicate through a structured chat chain to build the project, emphasising on multi-agent communication. MetaGPT (Hong et al., 2024) builds on this by encoding Standard Operating Procedures (SOP) into their prompts, following a defined process at each step. AgentCoder (Huang et al., 2023) takes a more straightforward approach, which consists of three agents, the Programmer, the Tester, and the Executor. The programmer agent writes the code, the testing agent creates tests to validate the generated code, and the execution agent executes these tests with the program, and returns the results to the programmer agent for iterative refinement. MapCoder (Islam et al., 2024) goes a step further by adding a retrieval step, which retrieves the previous relevant examples to inform code generation with prior solutions. Despite different approaches, all four show improvement over single-agent baselines on HumanEval and MBPP.

These frameworks show that multi-agent systems collaboratively outperform single-agent baselines. However, each framework differs across multiple variables simultaneously, including model choice, task set, and prompt design, making it difficult to isolate role specialisation as the sole contributing factor. No existing study has conducted a controlled comparison in which role specialisation is the only variable. This paper identifies that gap by building two identical systems, where the only difference is that one system has specialised roles whereas the other does not, then measuring which produces better code on HumanEval.

3. Methodology

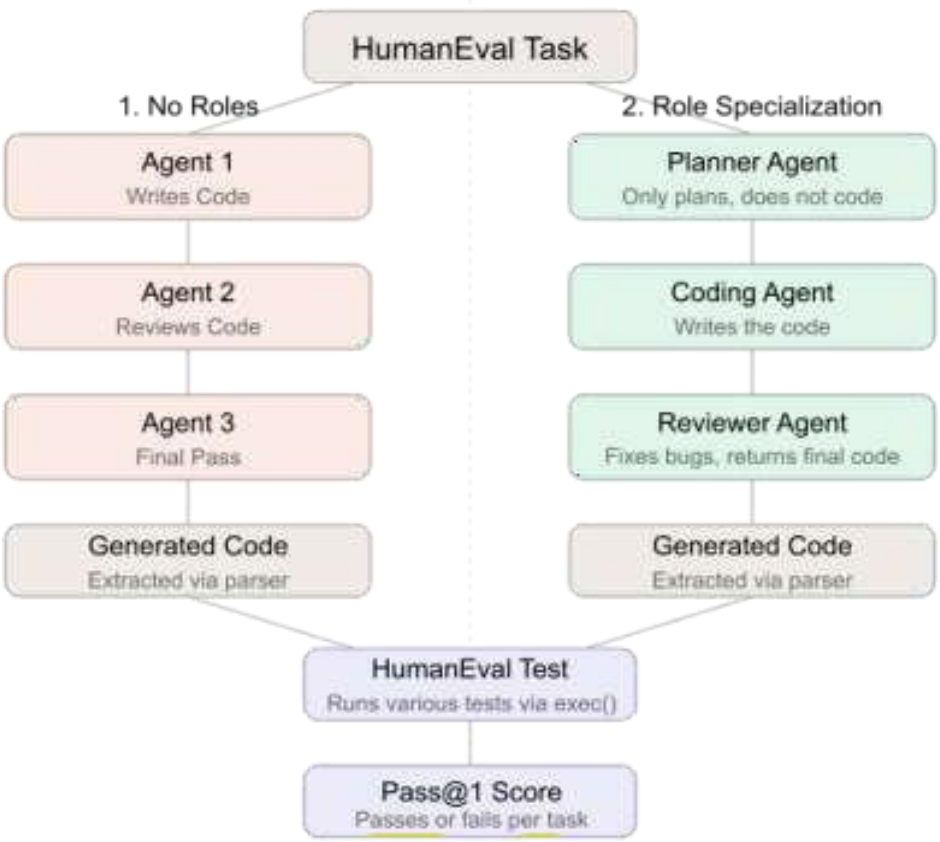


Fig.1

In this experiment, a task is fed into both systems simultaneously. In System A, the task is given to the first agent which writes the code using a generic prompt. The code is then handed to a second agent, which reviews it, and passes it to a third agent for a final review. In System B, the same task goes first to the planner, which produces a pseudocode prompt, a plain English breakdown of the solution's logic before any code is written. This process compels the system to think through the problem before starting its implementation process. Further, he coder then takes the plan and turns it into the python code. Finally, the reviewer checks the code for bugs and returns a corrected version. The output from both systems are run against the same automated tests, and the result is recorded as pass or fail.

4. Comparative Framework

This study evaluates two systems across four key dimensions. The first is Pass@1 rate, the primary measure of whether the generated code actually passes the automated unit tests, serving as the most direct indicator of output quality. The second is code robustness, whether the system accounts for edge cases such as empty inputs, invalid data types, or unexpected values. The third is their approach to problem solving, whether the presence of a dedicated planning stage meaningfully changed how the solution was structured before implementation. The fourth and final one is output consistency, whether the specialised system produced more predictable, well-structured code compared to the baseline across different task types.

Pass@1 was selected as the primary evaluation metric for several reasons. Unlike qualitative measures, it is entirely binary, the code either passes the given test or fails, eliminating any subjectivity in scoring. It is also automated, requiring no manual grading, and also ensures consistency across all 20 tasks. Finally, Pass@1 is the metric used by existing multi-agent coding frameworks including MetaGPT (Hong et al., 2024), AgentCoder (Huang et al., 2023), and MapCoder (Islam et al., 2024), meaning the results of this study are directly comparable to prior work.

5. Experimental Results

Across 20 tasks drawn from the HumanEval benchmark, the specialised system substantially outperformed the baseline. System A, the baseline, achieved a Pass@1 rate of 20%, passing 4 out of 20 tasks. System B, the specialised system, achieved a Pass@1 rate of 55%, passing 11 out of 20 tasks. This represents a 35 percentage point improvement, with the specialised system performing 2.75 times better than the baseline overall.

Task	**Baseline**	**Specialised**
has_close_elements	Fail	Pass
seperate_paren_groups	Fail	Fail
truncate_number	Fail	Pass
below_zero	Fail	Pass
mean_absolute_deviation	Pass	Pass
intersperse	Fail	Fail
parse_nested_parens	Fail	Fail
filter_by_substring	Pass	Pass
sum_product	Pass	Fail
rolling_max	Fail	Fail
make_palindrome	Fail	Fail
string_xor	Pass	Pass
longest	Fail	Pass
greatest_common_divisor	Fail	Pass

 ISBN: 978-81-685854-0-9

all_prefixes	Fail	Fail
string_sequence	Fail	Pass
count_distinct_characters	Fail	Pass
parse_music	Fail	Fail
how_many_times	Fail	Pass
sort_numbers	Fail	Fail

The following code samples were generated by each system during the experiment. These four tasks were selected as they best illustrate the performance difference between the two systems.

1. below_zero
Baseline output:

```
def fractional_part(number: float) -> float:
    return number - int(number)
```

Specialised output:

```
import math

def truncate_number(number: float) -> float:
    """ Given a positive floating point number, it can be decomposed into
    an integer part (largest integer smaller than given number) and decimals
    (the leftover part is always smaller than 1).

    Return the decimal part of the number.
    >>> truncate_number(3.5)
    0.5
    """
    integer_part = math.floor(number)
    decimal_part = number - integer_part

    if decimal_part < 0:
        decimal_part = 0

    return min(decimal_part, 1 - abs(decimal_part))
```

2. sum_product
Baseline output:

```
from functools import reduce

def sum_product(numbers: list) -> tuple:
    if not numbers:
        return (0, 1)

    total = sum(numbers)
    product = reduce((lambda x, y: x * y), numbers)

    return (total, product)
```

ISBN: 978-81-685854-0-9

Specialised output:

```
def sum_product(numbers: List[int] or None) -> Tuple[int, int]:
    if numbers is None:
        return (0, 1)

    if not numbers:
        return (0, 1)

    sum = 0
    product = 1

    for number in numbers:
        sum += number
        product *= number

    return (sum, product)
```

3. count_distinct_characters

Baseline output:

```
def count_distinct_characters(s: str) -> int:
    return len(set(s))
```

Specialised output:

```
def count_distinct_characters(string: str) -> int:
    """ Given a string, find out how many distinct characters (regardless of
case) does it consist of
    >>> count_distinct_characters('xyzXYZ')
    3
    >>> count_distinct_characters('Jerry')
    4
    """

    character_count = {}

    for char in string:
        char = char.lower()  # making it case-insensitive
        if char not in character_count:
            character_count[char] = 1

    return len(character_count)
```

 ISBN: 978-81-685854-0-9

4. has_close_elements

Baseline output:

```
from typing import List

def has_close_elements(numbers: List[float], threshold: float) -> bool:
    """Check if in given list of numbers, are any two numbers closer to each
other than
    the specified threshold.

    Args:
        numbers (List[float]): A list of floating-point numbers.
        threshold (float): The maximum distance between two numbers that are
considered close.

    Returns:
        bool: True if there exists a pair of numbers within the given threshold,
False otherwise.
    """
    if not numbers or len(numbers) <= 1:
        return False

    close_elements = [abs(x - y) <= threshold for x, y in zip(numbers,
numbers[1:])] + [abs(numbers[-1] - numbers[0]) <= threshold]
    return any(close_elements)
```

Specialised output:

```
from typing import List

def has_close_elements(numbers: List[float], threshold: float) -> bool:
    """ Check if in given list of numbers, are any two numbers closer to each
other than
    given threshold.
    >>> has_close_elements([1.0, 2.0, 3.0], 0.5)
    False
    >>> has_close_elements([1.0, 2.0, 3.0, 4.0, 5.0, 2.0], 0.3)
    True
    """

    for i in range(len(numbers)):
        if i < len(numbers) - 1 and abs(numbers[i] - numbers[i+1]) < threshold:
            return True
        for j in range(i+1, len(numbers)):
            if abs(numbers[i] - numbers[j]) < threshold:
                return True
    return False
```

Notable Observations

Several individual results are important to mention. In below_zero, the baseline misinterpreted the problem entirely., summing only negative numbers rather than tracking a running balance, whereas the specialised system correctly identified the problem during its planning stage. Conversely, sum_product represents one of the three cases where the baseline passed but specialized failed, demonstrating that role specialization does not universally guarantee better output. In count_distinct_characters, the baseline produced a one-line solution using Python's built-in set that failed to account for case-insensitivity, which the specialised system's reviewer caught and corrected. Both systems failed on 8 out of the 20 tasks, suggesting a performance ceiling that neither architecture could overcome.

ISBN: 978-81-685854-0-9

6. Discussion

The specialised system's superior performance can be attributed primarily to the separation of the planning from implementation. System B stopped the model from jumping to a solution before it understood the task by making the planner think through the problem in pseudocode. This is a common practice in development, where designing and drafting ahead of time helps avoid mistakes later on. Additionally, the Reviewer's dedicated role meant that edge cases and type errors were systematically caught rather than overlooked, as evidenced by tasks such as count_distinct_characters and below_zero.

But role specialization didn't always lead to better results. The baseline system worked for three tasks, including sum_product, but the specialized system did not. This means that for problems that are easier to understand, the extra planning steps make things more complicated than they need to be, and a direct approach to code generation works better.

Both systems failed on 6 of the 20 tasks, suggesting a shared performance ceiling that neither architecture could overcome. These failures were concentrated in tasks requiring nuanced string parsing and complex algorithmic logic, such as parse_music and make_palindrome. This indicates that while role specialisation improves reasoning and edge case handling, it cannot compensate for the underlying limitations of the language model itself. Architectural improvements alone are therefore insufficient — model capability remains a fundamental constraint.

7. Limitations

It is important to acknowledge several limitations. First, the sample size of 20 tasks is relatively small and may not be fully representative of the broader HumanEval benchmark or real-world coding scenarios. Second, both systems were tested exclusively using Mistral 7B, and results may differ significantly with larger or more capable models. Third, the experiment is sensitive to prompt design — minor changes to agent instructions could meaningfully affect outcomes, making exact replication difficult. Finally, Pass@1 measures only functional correctness and does not account for code readability, maintainability, or efficiency, which are equally important in practical software development.

8. Conclusion

This study set out to determine whether role specialisation alone meaningfully improves output quality in multi-agent code generation systems. The results demonstrate that it does, with the specialised system achieving a Pass@1 rate of 55% compared to 20% for the baseline, a 2.75x improvement across 20 HumanEval tasks. This suggests that enforcing role boundaries between planning, implementation, and review stages produces more robust and accurate code than generic multi-agent pipelines. Future work should explore larger task sets, multiple model architectures, and dynamic role assignment where agents adapt their roles based on task complexity to build on these findings.

References

1. Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. de O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., … Zaremba, W. (2021, July 7). *Evaluating Large Language Models Trained on Code*. arXiv.Org. https://arxiv.org/abs/2107.03374v2
2. He, J., Treude, C., & Lo, D. (2025). *LLM-Based Multi-Agent Systems for Software Engineering: Literature Review, Vision and the Road Ahead* (arXiv:2404.04834). arXiv.
3. https://doi.org/10.48550/arXiv.2404.04834
4. Hong, S., Zhuge, M., Chen, J., Zheng, X., Cheng, Y., Zhang, C., Wang, J., Wang, Z., Yau, S. K. S., Lin, Z., Zhou, L., Ran, C., Xiao, L., Wu, C., & Schmidhuber, J. (2023, August 1).

 ISBN: 978-81-685854-0-9

MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework. arXiv.Org.
5. https://arxiv.org/abs/2308.00352v7
6. Huang, D., Zhang, J. M., Luck, M., Bu, Q., Qing, Y., & Cui, H. (2023, December 20). *AgentCoder: Multi-Agent-based Code Generation with Iterative Testing and Optimisation*. arXiv.Org. https://arxiv.org/abs/2312.13010v3
7. Islam, M. A., Ali, M. E., & Parvez, M. R. (2024, May 18). *MapCoder: Multi-Agent Code Generation for Competitive Problem Solving*. arXiv.Org. https://arxiv.org/abs/2405.11403v1
8. Qian, C., Liu, W., Liu, H., Chen, N., Dang, Y., Li, J., Yang, C., Chen, W., Su, Y., Cong, X., Xu, J., Li, D., Liu, Z., & Sun, M. (2024). *ChatDev: Communicative Agents for Software Development* (arXiv:2307.07924). arXiv. https://doi.org/10.48550/arXiv.2307.07924

ANALYZING COPING STRATEGIES FOR STRESS MANAGEMENT USING SENTIMENT ANALYSIS AND MACHINE LEARNING

Mrs. Eshwari Shirke
Student, Department of Data Science, Kirti. M. Doongursee College, University of Mumbai, India.
Email: eshwarishirke@gmail.com

Dr. Prabha Kadam
Assistant Professor, Department of Computer Science, Kirti. M. Doongursee College, University of Mumbai, India.
Email: prabha.kadam@despune.org

Abstract
The paper was a study of mental stress and coping strategies, one being the study in sentiment analysis, and the another being through machine learning algorithm techniques. Primay data is collected through an structured google form that was then saved into the Excel database. The new clean dataset, consists of 10 columns of which the first 2 coulmns is Stress_Description for the main text feature and Coping_Main for the target feature. After class balancing, we are left with a total of 9,496 samples from which 2,374 came from each coping category. Text features is extacted using TF IDF vectorization followed by Chi-square feature selection,which reduces the number of features from 10,603 to 8,000. We then perform sentiment analysis with the help of TextBlob polarities. Some of machine learning models such as Calibrated Linear SVC, Logistic Regression, Random Forest, and Stacking Ensemble is used in this work. Logistic regression done the best with an accuracy of 72.53% while the others at best performed worse than 70%.
Keywords: Stress Management, Coping Strategies, Sentiment Analysis, Machine Learning, TF-IDF, Text Classification.

1. Introduction

Students, hospital patients, street-side vendors and everyone else is under stress. As such, individuals accept their own styles of coping from among these possibilities such as Some will turn to emotional support, relaxation, or avoidance. While others will actively involve in problem-focused actions. Learning methods for coping is all important in evaluating an individual's response to stress.

The stress experiences described in survey responses are such that manual coding of these text data is tough. However, with the help of sentiment analysis, the emotional tone of text data can be tested, thereby allowing machine learning techniques to be applied to classify them into structured categories.

This study represents how coping strategies as shown on certain types of stress input from questionnaires also show sentiment analysis results and machine learning algorithms based only on the dataset and notebook.

2. Literature Review

Stress manifests in several ways, depending on where it has invaded a person's life. Such stressors include an heavy academic course loads, personal responsibilities and not infrequently, time pressure. Untreated stress may affect emotional wellness and everyday

 ISBN: 978-81-685854-0-9

living. [1]
Coping strategies is thus simply different pathways through which individuals, depending on their temperaments, approach or avoid different stress-inducing events. This pathways are worlds apart from one individual to another and include such things as emotional support, avoidance behavior, relaxation exercises, or problem-solving activities. Understanding such strategies helps in the understanding of a person's response in stressful situations. [2]
Stress can even be observed in almost behavior-like studies conducted. Open-ended questions allow respondents to give a account of personal experiences in relation to stress and its management. This kind of text data is rich in information but gives much trouble when analyzing it manually. [3]
Sentiment analysis determines the emotional quality through the given data. It is then translated into numerical sentiment scores, usually measured from negative to positive value. This methodology was intended to trace an emotion tendencies upon coping strategy description. [4]
Dust-and-dirt-text pre-processing replace any work on sentiment analysis and ML. Hence, one may say noise suppression, stop word removal, symbols deletion, etc. is all part of the process of cleaning the text for whatever further leveling is tocome. In contrast, lemmatization reduces the various forms of words to one for better consistency through substitution. [5]
The process of feature extraction refers to the conversion of texts into numbers which allows machine models to process. Term frequency-based representations weight words by giving importance to words both in the given document as well as in the overall collection of datasets to which they belong. [6]
This exploratory data analysis talks about understanding the structure and type of the text. Word-cloud related visualizations actually put forward mostly repeated words from the start, while their uses of stress or any other bar plot in traditional way shows frequency distribution. [7]
The stress analysis led also to the field of an sentiment analysis that is able to extract the features and then pass them to machine-learning models to compare the stress textual data against observed patterns related to the coping strategies. [8]

3. Objective

1. Interpret the feelings communicated in the various coping strategies as derived from the texts data.
2. Compare the different models of machine learning according to their classification accuracies.

4. Methodology

We suggest the method in used in this work to be systematic behavior identification and prediction in coping strategies to the stress. This includes datasets preparation and preprocessing, modeling, hyperparameter tuning, evaluation and comparisons.

4.1 Dataset Description:

The dataset is collected through a Google Form. This is from the flexibility with which column names were changed during survey-style questions; therefore, the data sat in an Excel file called Stress Management Final.xlsx. After doing the cleanup and renaming, the dataset had 10 columns, such as

1. Age
2. Gender
3. Occupation
4. Stress_Frequency
5. Stress_Description

6. Coping_Description
7. Coping_Strategy
8. Coping_Effectiveness
9. Positivity_Score
10. Talks_Helpful

For modeling purposes, only 2 columns were utilized:
1. Stress_Description (text feature)
2. Coping_Main (target label)

After balancing, the final dataset had 9,496 samples, with 2,374 samples per coping category.

4.2 Data Preprocessing:

All column names were cleaned and renamed to get rid of unwanted character. Rows with missing stress descriptions and coping labels were removed. Text pre-processing followed:
1. Convert all text to lowercase
2. Delete URLs, punctuation, and numbers
3. Removing stop words
4. Remove Very short words
5. Lemmatization
6. The text is cleaned and stored in a new column for model processing.

4.3 Feature Extraction:

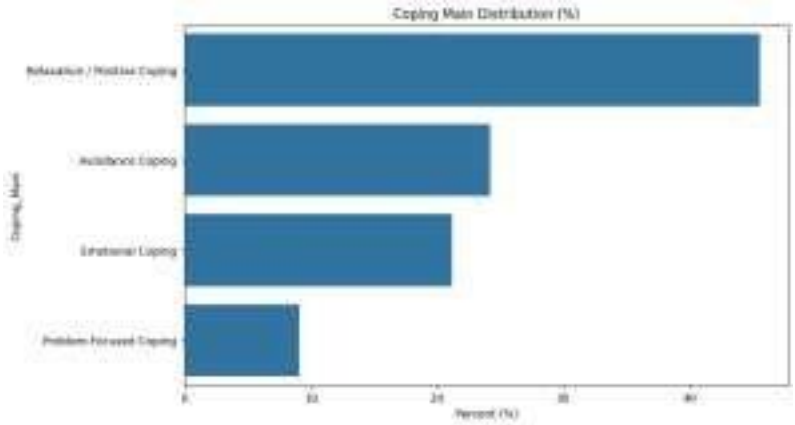
Numerical data of the text was converted by using TF-IDF. In the beginning, TF-IDF generated 10,603 features. With Chi-square feature selection, the feature size was finally reduced to 8,000. Then all the features went into one big machine-learning model.

4.4 Sentiment Analysis:

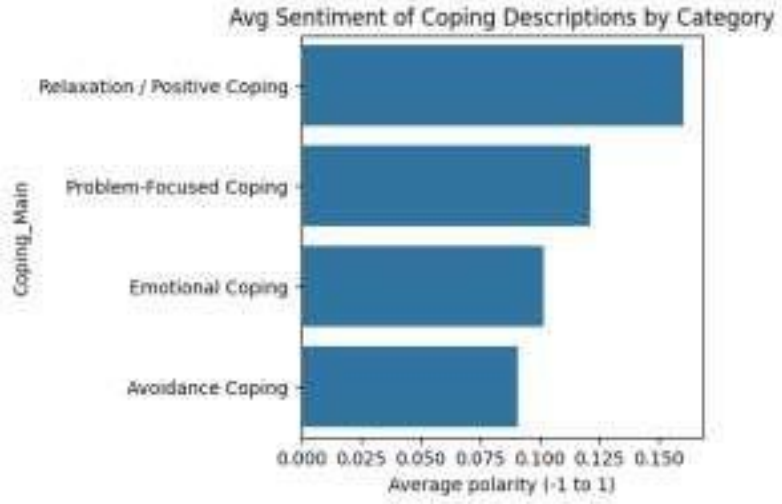
Sentiment analysis was done by Text Blob over the polarity score range of -1 to +1. The mean sentiment values for the categories of coping are:
• Relaxation / Positive Coping: 0.1601
• Problem-Focused Coping: 0.1208
• Emotional Coping: 0.1014
• Avoidance Coping: 0.0909

4.5 Exploratory Data Analysis (EDA):

The following visualizations are created:
• Shows the percentage distribution of coping categories.
• The graph shows the average sentiment polarity for every way of coping the stress. • It focuses on common words related to stress with TF-IDF weights.



 ISBN: 978-81-685854-0-9





4.6 Machine Learning Models Used:

- **Calibrated linear SVC**: To predict the text to get probability outputs.
- **Random Forest:** For comparision validation with balanced class weights.
- **Logistic Regression**: In classifying coping mechanisms based at TF-IDF features.
- **Stacking Ensemble:** Combined prediction with all 3 models treated as equally weighted.

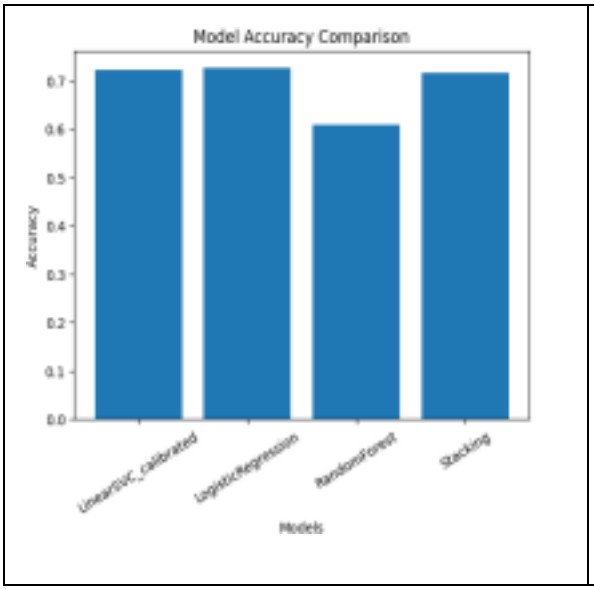
5. Result and Analysis

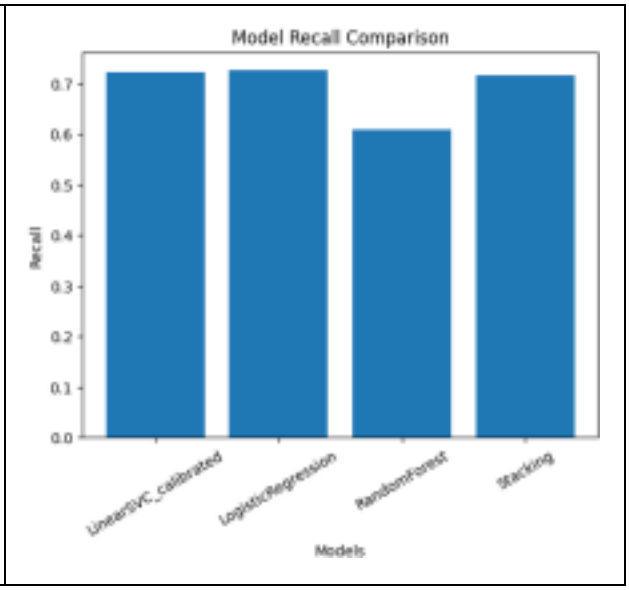
Below are the accuracy results of the machine learning models:

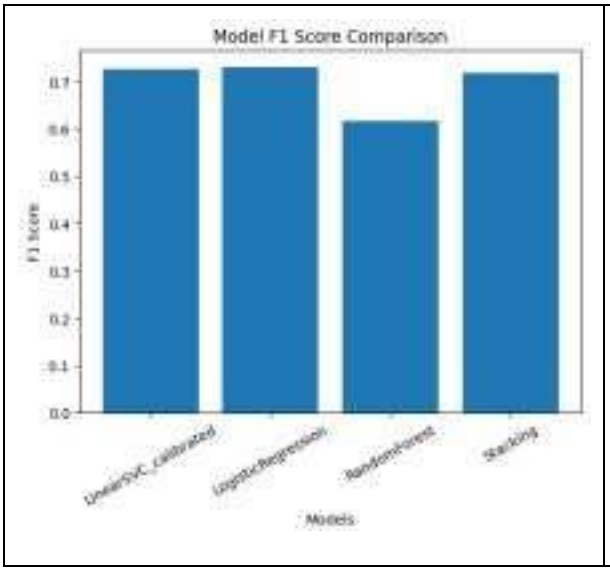
Model Name	**Accuracy (%)**
Calibrated Linear SVC	72.26
Logistic Regression	72.53
Random Forest	60.84
Stacking Ensemble	71.63

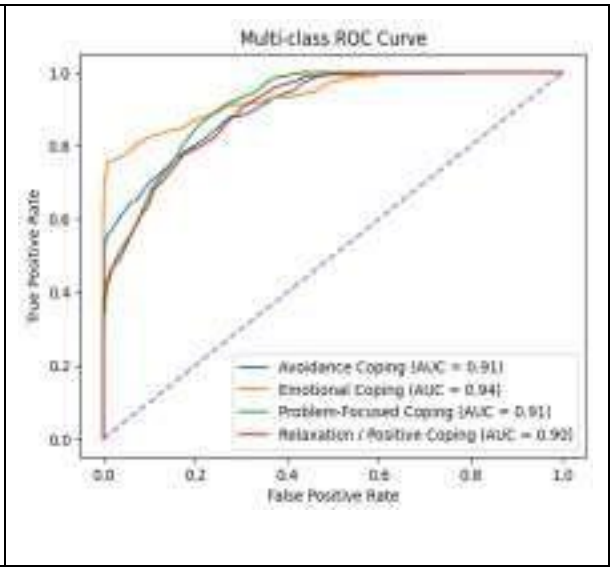
Logistic Regression achieved the highest accuracy at 72.53%, beating the other models in performance. The stacking model proved that performance wise with an best individual algorithms but not more impactful than the Logistic Regression model. The best performance among all of the models is achieved by Random Forest.

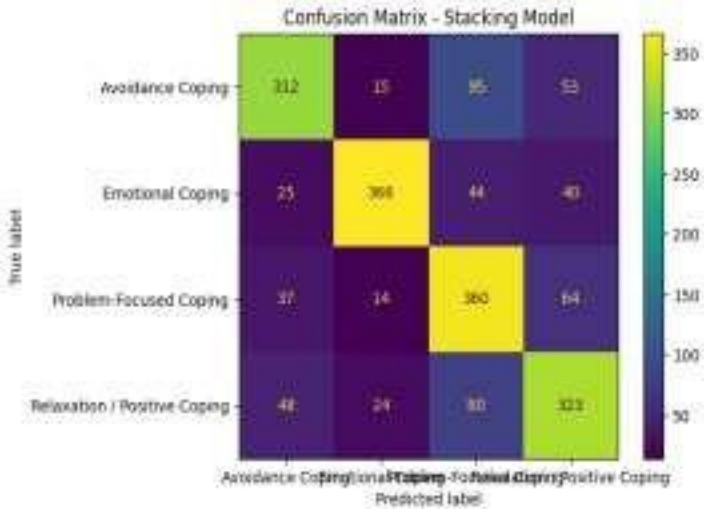
ISBN: 978-81-685854-0-9











 ISBN: 978-81-685854-0-9

6. Conclusion and Future Scope

6.1 Conclusion:

The study performed out on sentiment analysis of stress coping strategies was done on the basis of machine learning models. Text data were processed by a user-written program for extract features TF-IDF and select them in the Chi-square. The most accurate accuracy of 72.53% is achieved by the Logistic Regression model for all models tested. The sentiment analysis of all coping categories showed an average positive polarity.

6.2 Future Scope:

1. Model simplification can be enhanced through the use of large datasets.
2. More complicated methods can be adopted for feature extraction.
3. Research can be conducted on other machine learning or deep learning approaches to text classification.

7. References

1. Ahmed, S. and Roy, D. 2022. Comparative study of coping techniques throughout the COVID-19 epidemic. Frontiers in Behavioral Science, 6, 198–211. Using both qualitative and computational data, this study examines coping strategies to worldwide stressors.
2. Chakraborty, P., & Basu, A. (2020). Improving Mental Health Assessment by Means of ML: This piece looks at how machine learning may improve psychological evaluations.
3. Handbook of Stress and Health: Research and Practice Guide, Wiley-Blackwell by Cooper, C. L. and Quick, J. C. A thorough guide with contemporary ideas on stress and coping studies.
4. George, L. and Nair, M. (2021). Impact of positive coping on emotional stability: A data-driven approach. Asia-Pacific Journal of Psychology, 11(1), 45–60. This study shows how emotional stability and adaptive coping are related.
5. Through text mining, Lee, J., and Kim, H. (2019) examine emotional trends in academic stress. Computers in Human Behavior, 92, 234–245. This study examines stress-related text patterns among pupils using natural language processing.
6. Machine learning-based emotional profiling in digital communication. Information Processing & Management, 58(6), 102723. Park, E. and Lim, D. This study outlines how machine learning finds coping strategies in internet discourse.
7. Patel, D.; Jain, R. 2020. Sentiment analysis for mental health monitoring. IEEE Transactions on Affective Computing, 11(4), 712–721. This paper looks at how emotion models spot patterns in emotional well-being.
8. Feature extraction for text-based stress detection by Roy, T., and Sen, B. in the International Journal of Artificial Intelligence Research, 10(2), 45–56. This study looks at techniques for extracting linguistic and semantic aspects in NLP to find stress.
9. Health psychology. McGraw-Hill Education, 2017 Gives a basic look at health, stress, and coping strategies.
10. Wilson, A.; Thomas, P. AI ethics and human-centered data science. Computers & Society, 34(2), 90–102. Highlights ethical and responsible behavior in emotion-based artificial intelligence research.
11. Yadav, R.; Singh, A. (2020). Correlation between emotional expression and mental well-being. Indian Journal of Psychology, 95(4), 345–356. Investigates the mental effects of patterns of emotional expression.

AI-DRIVEN SEMANTIC SEARCH IN MYSQL USING VECTOR EMBEDDINGS AND LARAVEL PRISM

Prachi Barve
MCA Department, P E S Modern College of Engineering Pune, India.

Dr. Mrs. Pratibha Adkar
MCA Department, P E S Modern College of Engineering Pune, India.

Abstract
Modern database search is still largely based on matching exact words, which creates problems when users express the same idea differently. This work explores a different approach where meaning is prioritized over exact text. The system is built using MySQL and a Laravel-based application, enhanced with vector representations of text so that queries can be understood contextually rather than literally. Instead of relying only on stored strings, each piece of text is converted into a numerical form that reflects its meaning. When a user searches, the system compares meanings instead of words, which leads to more relevant results even when phrasing varies.
Keywords: Semantic Search, Vector Embeddings, Relational Databases, AI Integration, MySQL Optimization, Full-Text Search, Keyword Matching.

1. Introduction:

Relational database systems like MySQL are widely used in modern web applications due to their reliability and efficiency in handling structured data such as user information, transactions, and product records. However, their text search capabilities are largely limited to keyword matching, which focuses on exact words rather than meaning.

In real-world scenarios, users expect intelligent search behavior. For instance, a query like "budget phone" should ideally return results related to "affordable smartphones." Traditional full-text search may fail in such cases because it cannot interpret semantic relationships between words.

To overcome this issue, this research proposes an AI-based semantic search framework that integrates vector embeddings into MySQL using Laravel and a vector-enabled ORM approach. Instead of storing only textual data, the system also maintains vector representations that capture contextual meaning. User queries are converted into vectors and compared with stored embeddings to retrieve semantically relevant results.

This framework allows developers to incorporate AI-driven search directly within a standard SQL environment without migrating to specialized vector databases. The study compares this approach with traditional methods using metrics such as precision, recall, and execution time. Results are expected to show improved search quality while maintaining system performance.

2. Literature Review:

Hu, et al. (2025). The authors introduce HAKES, a distributed vector database designed to reduce high indexing costs and performance issues during concurrent updates. The system uses a two-stage "filter-and-refine" approach, where compressed vectors are used for fast candidate selection and full-precision vectors are applied for final ranking. One key contribution is the separation of search and insert operations, which allows updates without rebuilding the entire index. The architecture separates indexing and refinement tasks to scale memory and

 ISBN: 978-81-685854-0-9

computation independently. Experimental results show that HAKES achieves better throughput and recall balance compared to existing systems, with significantly higher performance under heavy workloads.

Wang, et al. (2025). Wang and colleagues introduced *Text2VectorSQL*, a framework that connects vector search with traditional SQL queries in a unified interface. Their work focuses on reducing the gap between structured database queries and semantic vector retrieval. The system allows users to combine relational filters with vector similarity search in a single query format. The study shows that integrating vector search directly into database workflows improves flexibility and usability. However, the framework mainly targets system-level design and does not deeply explore performance comparison with traditional keyword search in real-world application environments.

Malo and Al-zebari (2025). The authors proposed an intelligent semantic search system for academic journals using artificial intelligence and natural language processing techniques. Their model converts research articles into vector representations and applies similarity matching to retrieve relevant documents based on meaning rather than keywords. Experimental results indicate improved precision and relevance compared to standard keyword-based search systems. While the study demonstrates strong performance in academic datasets, it focuses primarily on document retrieval and does not address integration within conventional relational database systems like MySQL.

Farivar (2025). This survey reviews the evolution of information retrieval systems from traditional lexical models such as BM25 and TF-IDF to modern semantic retrieval approaches. The author explains that keyword-based methods rely on exact word matching, which often fails to capture contextual meaning. The paper discusses advanced architectures including dense bi-encoders (DPR), late-interaction models (ColBERT), neural sparse retrieval models (SPLADE), and cross-encoder models (MonoT5). It highlights the importance of semantic similarity through embeddings to improve search relevance. The study also describes improved training methods such as hard negative sampling and knowledge distillation. Although benchmark results on datasets like MS MARCO and BEIR show strong performance, the paper notes ongoing challenges related to scalability and language bias in English-focused datasets

Kumar, et al. (2025). This review discusses the limitations of traditional keyword-based search and proposes an embedding-based vector search approach for large-scale text retrieval. The authors explain that keyword methods often fail to capture contextual meaning, which is important in fields such as law and academic research. The system uses transformer-based models like BERT and RoBERTa to generate high-dimensional vector embeddings. For scalability, it applies FAISS for efficient indexing and fast query processing. The architecture also integrates Retrieval-Augmented Generation (RAG) to support query refinement and summarization. The study concludes that embedding-based systems provide better retrieval accuracy than traditional keyword search, especially when supported by optimized indexing methods.

Chronis, et al. (2025). This tutorial provides an overview of filtered vector search, which combines vector similarity queries with structured database conditions. The authors explain that integrating vector search into database engines introduces new optimization challenges. They categorize execution strategies into pre-filtering, post-filtering, and inline filtering, and analyze how filter selectivity affects performance and recall. A major focus of the paper is achieving stable and predictable recall without requiring manual parameter tuning. The study also reviews different types of filter-aware indexing methods and identifies open research challenges in benchmarking, hybrid search, and automatic configuration.

Singh, (2024). The author presents a comprehensive review of the transition from traditional token-based retrieval to modern vector search technology, arguing that exact lexical matching fails to capture the complexity of human language and search intent. The study explores how

neural networks transform queries and documents into dense vector embeddings within a high-dimensional space, allowing semantically similar concepts to cluster together regardless of specific keyword usage. By analyzing technical implementations across diverse domains—including e-commerce, healthcare, and legal research—the paper demonstrates that vector-based systems consistently achieve a 20-30% improvement in discovery accuracy and user satisfaction over traditional methods. The research concludes that while scalability remains a challenge, the integration of vector search with privacy-preserving technologies and distributed computing frameworks offers a promising roadmap for future knowledge discovery.

Prabhakar (2024). Prabhakar provides a detailed overview of vector search techniques and their growing importance in modern data systems. The paper explains how vector embeddings represent textual meaning in numerical form and how similarity measures such as cosine similarity are used for retrieval. It also discusses the role of vector databases and indexing methods in improving search efficiency. While the study offers a strong conceptual understanding of vector search, it does not focus on practical implementation within relational databases or web frameworks.

Khalifeh, et al. (2024). This research enhances keyword search in relational databases by incorporating word embeddings. The authors propose expanding user queries using semantically related terms generated through embedding models. Instead of replacing SQL search, their method improves it by adding semantic awareness to traditional keyword queries. The results show better recall and more meaningful matches compared to basic full-text search. However, the approach still depends partly on keyword expansion rather than full vector similarity search inside the database.

Wang and Koopman (2017). This study compares different semantic embedding techniques, including Word2Vec, GloVe, and Doc2Vec, with a count-based method called Ariadne. The authors emphasize that the effectiveness of an embedding method depends on the application context. Using a Medline dataset, they evaluate the performance of these models in an information retrieval task related to evidence-based medicine. The results show that while predictive models like Word2Vec perform well in general semantic tests, the simpler count-based Ariadne method achieves comparable or better results in practical retrieval tasks. The paper concludes that combining semantic embeddings with bibliometric features may further improve retrieval performance.

Research Gap:

Although semantic search and vector retrieval have seen significant advancements, most existing solutions focus on large-scale systems or dedicated vector databases. These approaches often prioritize infrastructure improvements rather than practical implementation within commonly used relational databases.

Many current methods require specialized tools or extensive computational resources, making them less suitable for smaller applications. Some attempts have been made to combine semantic search with relational systems, but they either still rely on keyword matching or lack real-world performance evaluation.

Additionally, several models are domain-specific, limiting their adaptability across different applications. There is a noticeable gap in developing lightweight, scalable, and easy-to-integrate solutions that bring semantic search capabilities directly into traditional relational databases like MySQL.

 ISBN: 978-81-685854-0-9

3. Methodology:

3.1 Proposed system architecture with diagram

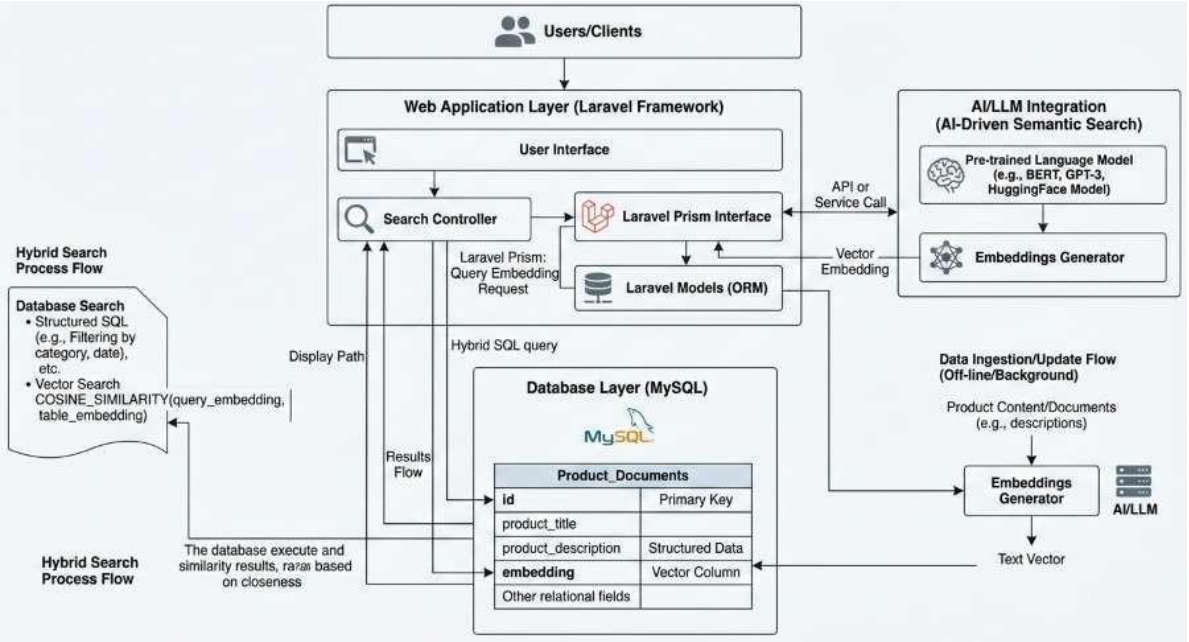


Fig. 3.1.1 Architecture of AI-Driven Semantic Search in MySQL Using Vector Embeddings and Laravel Prism

The proposed system is designed as a layered architecture built on Laravel and MySQL. It includes the user interface, application logic, AI processing, and database layers.

Users interact through a web interface where search queries are submitted. These queries are handled by a controller within the Laravel application, which coordinates between the database and the embedding service.

Text data such as product descriptions is converted into high-dimensional vector embeddings using pre-trained language models. These embeddings are stored alongside structured data in MySQL.

When a query is entered, it is transformed into a vector and compared with stored vectors using cosine similarity. The system combines SQL-based filtering with semantic similarity scoring to produce ranked results based on meaning rather than exact matches.

This hybrid design allows semantic search to function within the existing database system without requiring a separate vector database.

3.2 Working of the system

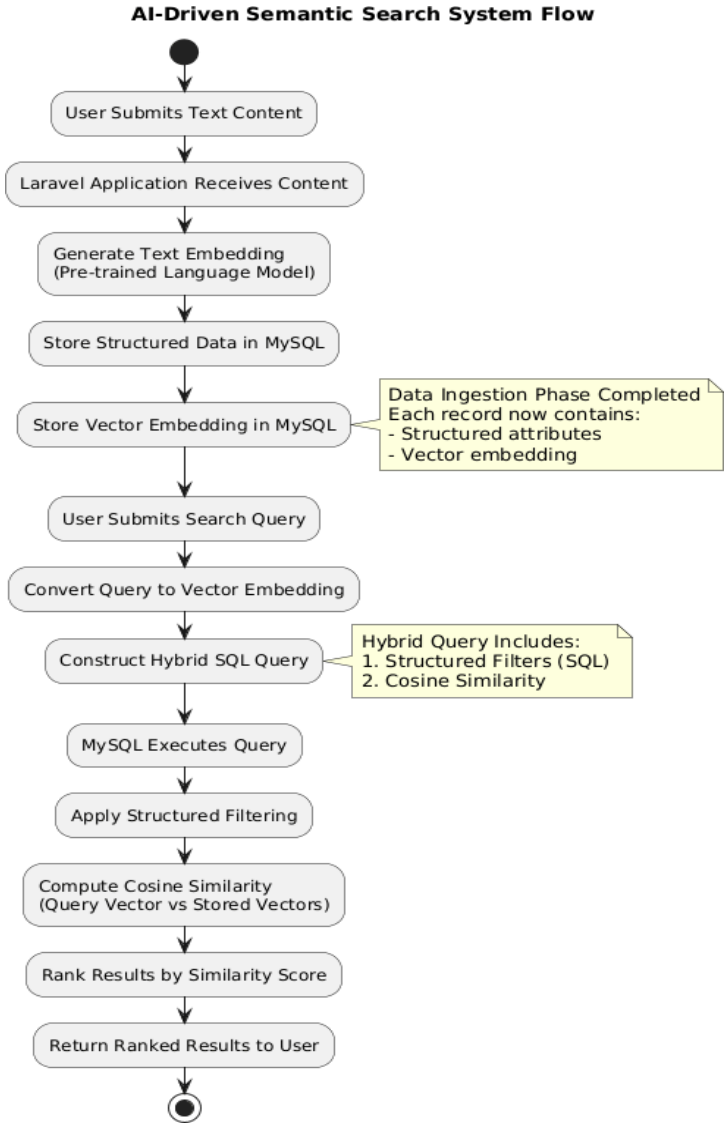


Fig. 3.2.1 Flow of the system

The system operates in two main stages:

1. Data Processing and Storage:

Textual data is submitted and converted into embeddings using an AI model. These vectors are stored together with relational attributes in the database, forming a combined storage model.

2. Query Processing and Ranking:

When a user submits a query, it is converted into an embedding. The system then performs a hybrid search that includes:

- Structured filtering (e.g., category, price)
- Semantic comparison using cosine similarity

The database processes both components and ranks results based on similarity scores. Higher scores indicate stronger relevance.

This approach improves search quality while preserving compatibility with relational database systems.

 ISBN: 978-81-685854-0-9

4. Tools and Technology Used:

The system is implemented using:

- Laravel framework for application development
- MySQL for data storage
- Vector embeddings generated from pre-trained AI models
- Cosine similarity for semantic comparison
- Laravel Prism for vector operations
- PlantUML for system design visualization

These technologies enable seamless integration of semantic search into a traditional database setup.

5. Result and Analysis:

5.1 Results

1. The system successfully integrates vector-based semantic search within a traditional MySQL relational database environment.
2. Semantic retrieval is implemented without migrating to a separate vector database, ensuring compatibility with existing infrastructure.
3. Vector embeddings are stored alongside structured relational data, enabling hybrid queries that combine SQL filtering and cosine similarity computation.
4. Comparative evaluation shows that semantic search provides improved relevance over traditional LIKE queries and full-text indexing.
5. The system achieves higher precision and recall, especially for queries involving synonyms or contextually related terms.
6. Hybrid query execution maintains acceptable response time and scalability across larger datasets.
7. Overall, the system enhances search accuracy, improves user experience, and demonstrates practical integration of AI capabilities within the MySQL ecosystem.



Fig. 5.1.1 Traditional keyword match in mysql



Fig. 5.1.2 Failed Semantic Search in mysql

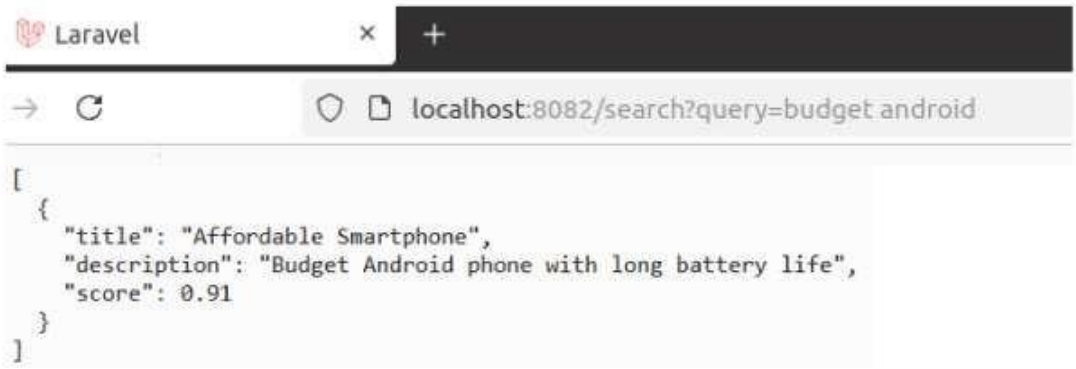


Fig. 5.1.3 Ai Driven Semantic Search

5.2 Analysis

The system was evaluated based on accuracy, relevance, and performance. A comparison between traditional and semantic search methods showed that vector-based retrieval delivers more meaningful results, even when exact keywords are absent.

Although semantic processing introduces slight computational overhead, overall performance remains efficient. The hybrid approach effectively balances accuracy and scalability within a relational database environment.

6. Future Scope:

Future improvements may include:

- Use of advanced transformer-based models for better understanding
- Optimization using vector indexing and approximate nearest neighbor techniques
- Integration of user behavior for personalized ranking
- Expansion into domain-specific applications like healthcare or e-commerce
- Deployment in distributed or cloud-based systems for scalability

7. Conclusion:

This research presents a semantic search solution that enhances MySQL databases using AI-driven vector embeddings. By combining relational data with semantic representations, the system enables meaning-based retrieval without requiring major infrastructure changes.

Experimental results confirm that this approach improves search accuracy and relevance compared to traditional keyword-based methods while maintaining acceptable performance. The study demonstrates that modern AI capabilities can be effectively integrated into existing relational database systems, making them more suitable for intelligent applications.

References:

1. X. Hu et al., “HAKES: Scalable Vector Database for Embedding Search Service,” arXiv preprint arXiv:2505.12524, 2025.
2. Z. Wang et al., “Text2VectorSQL: Towards a Unified Interface for Vector Search and SQL Queries,” arXiv preprint arXiv:2506.23071, 2025.
3. S. F. Malo and A. Al-zebari, “Intelligent Semantic Search for Academic Journals Using AI and NLP Techniques,” Journal of Information Systems Engineering and Management, vol. 10, no. 41s, pp. 404–420, 2025.
4. M. Farivar, “Semantic Search for Information Retrieval,” 2025.
5. R. Kumar et al., “Embedding-Based Vector Search for Large Scale Text Retrieval – A Review,” International Journal for Research Trends and Innovation (IJRTI), 2025.
6. A. Chronis et al., “Filtered Vector Search: State-of-the-Art and Research Opportunities,” Proceedings of the VLDB Endowment (PVLDB), vol. 18, 2025.

 ISBN: 978-81-685854-0-9

7. F. Khalifeh, M. Taheri, E. Mansoori, and M. Fakhrahmad, “Enhancing Keyword Search in Relational Databases with Word Embeddings,” IEEE Access, vol. 12, pp. 173775–173791, 2024.
8. G. Prabhakar, “Vector Search,” International Journal of Computer Trends and Technology (IJCTT), vol. 72, no. 12, pp. 1–7, Dec. 2024.
9. A. Singh, “Vector Search in the Era of Semantic Understanding: A Comprehensive Review of Applications and Implementations,” International Journal of Computer Engineering and Technology (IJCET), vol. 15, no. 6, 2024.
10. Y. Wang and W. Koopman, “Comparing Semantic Word and Document Embeddings with a Count-Based Method (Ariadne) for Information Retrieval,” CEUR Workshop Proceedings, 2017

SCROLL SAFE: BUILDING HYBRID AI SYSTEMS FOR CONTENT MODERATION ON SOCIAL MEDIA PLATFORMS

Prathamesh A. Kamble
Department of Computer Science, Kirti M. Doongursee College of Arts Science and Commerce (Autonomous), Mumbai.
Email: botspyder87@gmail.com

Dr. Prabha Kadam
Department of Computer Science, Kirti M. Doongursee College of Arts Science and Commerce (Autonomous), Mumbai.
Email: prabha.kadam@despune.org

Abstract
Social media platforms have become central to everyday communication, and with this growth comes an alarming rise in harmful content — including hate speech, online bullying, misinformation, harassment, and abusive language. Managing this content has become one of the most urgent challenges of the digital era. While human moderators are essential for contextual judgement, they are overwhelmed by the sheer volume of posts uploaded every minute. Fully automated AI systems, on the other hand, are not yet reliable enough to handle complex language situations involving sarcasm, regional dialects, and coded speech. This paper introduces Scroll Safe, a hybrid content moderation framework that intelligently combines AI-based text classification with human oversight to create a faster, fairer, and more scalable moderation pipeline. Using a fine-tuned BERT-based natural language processing model trained on publicly available toxic content datasets, the Scroll Safe system achieves a classification accuracy of 91% when working independently and 95% when combined with human verification. The study also demonstrates significant improvements in review speed, moderator workload, and false positive rates compared to traditional manual moderation. The research further proposes a generalized content policy model applicable across different platforms, with special consideration for multilingual environments such as India. This work has important implications for platform safety, moderator mental health, and responsible digital governance.
Keywords: Content Moderation, Hybrid AI, Scroll Safe, Social Media Safety, NLP, BERT, Harmful Content Detection, Human-in-the-Loop, Hate Speech, Online Safety.

1. Introduction

Social media has woven itself into the fabric of modern life. Whether it is Instagram, Facebook, YouTube, WhatsApp, TikTok, or X (formerly Twitter), billions of people scroll, post, share, and communicate through these platforms every day. This massive digital activity has transformed how people access news, entertainment, education, and social connection. However, it has also created a serious and growing problem: the rapid spread of harmful content.

Harmful content takes many forms — hate speech targeting communities, online bullying that damages mental health, misinformation that misleads public opinion, sexual content involving minors, harassment campaigns against individuals, and illegal trade activities. These are not rare or isolated events; they happen continuously across all platforms, around the clock, in multiple languages. The scale is staggering. According to Meta, over three billion posts,

 ISBN: 978-81-685854-0-9

stories, and reels are created every day on Facebook and Instagram alone.

To deal with this challenge, platforms hire large teams of human content moderators. Their job is to review reported content, apply community policies, and remove anything that crosses the line. But this is an almost impossible task to do at scale. Human moderators deal with limited bandwidth, emotional exhaustion from repeated exposure to disturbing content, inconsistencies in judgement, and potential bias. Studies show that many content moderators suffer from post-traumatic stress, anxiety, and burnout due to constant exposure to violent and abusive material.

On the technology side, Artificial Intelligence offers hope. AI models can scan thousands of posts per second, detect patterns in language, and flag suspicious content before a human ever sees it. But AI is not perfect. It can misread sarcasm, struggle with regional slang, fail with mixed-language content, and be tricked by users who deliberately modify their words to bypass filters. Relying entirely on AI often results in unfair bans of innocent users or failure to catch genuinely dangerous content.

This research proposes a middle path — a hybrid system where AI and humans work together, each doing what they do best. The Scroll Safe system uses a fine-tuned AI model to do the heavy lifting of first-level filtering, and then routes only uncertain or sensitive cases to human reviewers for final decision. This reduces human exposure to harmful content, speeds up the moderation process, and results in more accurate and consistent outcomes.

This paper is organized as follows: Section 2 presents the problem statement. Section 3 outlines the study objectives. Section 4 reviews relevant literature. Section 5 states the research hypothesis. Section 6 describes the methodology. Section 7 presents results and analysis. Section 8 provides conclusions and future directions.

2. Problem Statement

Today, social media platforms generate an overwhelming volume of content every minute. Millions of posts, comments, images, and videos are uploaded globally, and within this flood of content, a significant portion is harmful, misleading, abusive, or illegal. Human moderators cannot keep pace with this volume — by the time a team reviews and removes harmful content, it may have already been seen and shared by thousands of users.

Fully automated AI moderation, while fast, comes with its own risks. AI systems can incorrectly remove harmless content, fail to detect cleverly worded abuse, or struggle to understand cultural and linguistic context — particularly in multilingual countries like India where users frequently mix English with regional languages such as Hindi, Marathi, Tamil, and Telugu.

There exists a clear gap between what human-only and AI-only moderation can achieve. Neither approach alone is sufficient for the current scale and complexity of social media content. This research addresses that gap by proposing and testing a hybrid moderation model — Scroll Safe — that combines the speed and scale of AI with the contextual intelligence and fairness of human judgement. The goal is to build a moderation system that is fast, accurate, scalable, and respectful of both user rights and moderator well-being.

3. Objectives

• To develop and evaluate a hybrid content moderation system (Scroll Safe) that combines AI-based detection with human review for improved accuracy and efficiency.

• To identify and classify different types of harmful content commonly found on social media platforms, including hate speech, bullying, misinformation, harassment, and toxic language.

• To compare the performance of multiple machine learning and deep learning models — including Logistic Regression, SVM, Random Forest, and BERT — for harmful content classification.

• To study how major social media platforms design and apply their content policies and to derive a simplified, generalizable policy framework applicable to multiple platforms.
• To demonstrate that hybrid moderation reduces human moderator workload, speeds up content review, and lowers both false positive and false negative rates.
• To propose a prototype web-based Scroll Safe dashboard that can be used in real-world content moderation workflows, including multilingual settings.

4. Literature Review

Research on automated content moderation has grown substantially over the past decade, driven by the increasing complexity and volume of online harmful content. This section reviews key studies across four major themes: AI-based detection methods, hybrid moderation systems, policy and ethics, and dataset development.

4.1 AI-Based Harmful Content Detection: Early approaches to harmful content detection relied heavily on keyword lists and rule-based filtering. Davidson et al. (2017) showed that such methods are unreliable for hate speech detection, as many words that appear abusive in literal readings are used harmlessly in everyday conversation. Schmidt and Wiegand (2017) reinforced this finding, recommending that deeper natural language processing models are needed to understand intent, context, and cultural meaning behind words. Badjatiya et al. (2017) took this further by applying deep learning to tweet-level hate speech detection and demonstrating significantly better performance than traditional methods. Later, Zampieri et al. (2021) showed that transformer
based models such as BERT are capable of identifying offensive content with high precision even in nuanced situations.

4.2 Contextual and Behavioral Analysis: Zhou et al. (2019) highlighted that identifying harmful behavior cannot be done by analyzing individual posts in isolation. Real toxicity often follows patterns across user histories and conversation threads. Mathew et al. (2020) similarly found that a small group of repeat offenders are responsible for a disproportionate share of harmful posts, suggesting that user-level behavioral monitoring is an effective strategy for early detection. Chandrasekharan et al. (2021) demonstrated that community-wide moderation interventions can reduce toxic behavior when applied consistently and transparently.

4.3 Hybrid and Human-in-the-Loop Systems: Ortega et al. (2020) provided strong evidence that hybrid systems combining machine automation with human oversight outperform both purely manual and purely automated approaches. The human-in-the-loop design reduces error rates, improves fairness in edge cases, and allows AI models to continuously improve through real-world feedback. ADL Research Lab (2021) and Chandrasekharan et al. (2021) both concluded that hybrid moderation is the most practical and sustainable solution for large-scale platforms.

4.4 Ethics, Bias, and Fairness: Boczek and Koppers (2020) raised important concerns about the ethical implications of fully automated moderation, emphasizing that AI decisions must be transparent and explainable to users. Santurkar et al. (2020) demonstrated that training data bias leads to unfair moderation outcomes for certain demographic groups. Marin et al. (2019) showed that cultural differences and community-specific expressions are frequently misread by AI models trained on homogeneous datasets. Kellogg et al. (2021) argued that AI systems must be accountable and that users deserve to understand why their content was flagged or removed.

4.5 Dataset Development and Evaluation Benchmarks: Wulczyn et al. (2017) introduced the widely-used Wikipedia Toxic Comment Dataset, which became a foundational benchmark for moderation research. Vidgen and Derczynski (2020) proposed standardized evaluation frameworks to allow fair comparison across systems. Pujari et al. (2021) addressed a critical gap in multilingual moderation by demonstrating that models trained only on English data

 ISBN: 978-81-685854-0-9

perform poorly on mixed-language content, a major concern for platforms operating in diverse countries like India.

Taken together, the existing literature confirms that a hybrid, context-aware, and ethically designed moderation system — such as Scroll Safe — represents the most effective path forward for modern social media platforms.

5. Hypothesis

This research is grounded in the belief that combining Artificial Intelligence with human moderation produces outcomes that are measurably better than either approach working alone. The central hypothesis of this study is:

"A hybrid content moderation system (AI + Human) will significantly reduce the time required to detect and remove harmful content, improve decision accuracy, and lower moderator stress levels compared to traditional manual moderation methods."

This hypothesis assumes that AI will efficiently handle the first level of screening — quickly filtering out clear cut cases of harmful content — while humans will focus their attention and expertise on ambiguous, culturally sensitive, or contextually complex cases. The division of labour is expected to result in faster overall review cycles, reduced moderator fatigue, fewer false positives, and higher accuracy in final content decisions.

6. Methodology

This study follows a structured, multi-phase methodology that covers data collection, preprocessing, AI model training, hybrid system integration, and performance evaluation. Each phase is described in detail below.

Phase 1 — Data Collection: The research used multiple publicly available datasets of labelled social media content, including the Jigsaw Toxic Comment Dataset (sourced from Kaggle), the Wikipedia Detox Dataset, and the Kaggle Hate Speech Dataset. These datasets contain text samples pre-labelled into categories such as normal, toxic, hate, bully, and misinformation. A supplementary survey was also designed to understand user experiences with harmful content in the Indian social media context, particularly around multilingual content.

Phase 2 — Data Pre-processing: Raw text data was cleaned using Python. This included removing special characters, duplicate entries, emojis, HTML tags, and excessive whitespace. Tokenization was applied using the NLTK library, and word embeddings were generated using both Word2Vec and BERT tokenizer to prepare input representations for machine learning models. Since harmful content represented a minority class in most datasets, oversampling using SMOTE (Synthetic Minority Over-sampling Technique) was applied to balance the training distribution.

Phase 3 — AI Model Training: Five models were trained and evaluated: Logistic Regression, Naive Bayes, Support Vector Machine (SVM), Random Forest, and a fine-tuned BERT-based transformer model. The dataset was split 80:20 for training and testing respectively. Hyperparameter tuning was applied to each model to optimize learning rate, tree depth, regularization strength, and batch size. All models were evaluated on Accuracy, Precision, Recall, and F1-Score.

Phase 4 — Hybrid Moderation Pipeline: The best-performing AI model (fine-tuned BERT) was integrated into a hybrid pipeline. In this pipeline, incoming social media content is first scanned by the AI model. Content with a high confidence harmful classification is automatically flagged and removed. Content that the model is uncertain about (confidence below a set threshold) is routed to a human reviewer dashboard. Human reviewer decisions are logged and used as feedback to continuously retrain and improve the model — a process known as human-in-the-loop learning.

Phase 5 — System Prototype (Scroll Safe Dashboard): A web-based prototype dashboard

was developed using Streamlit and Flask. It accepts content input, runs AI classification in real time, displays predicted categories and confidence scores, and queues borderline cases for human review. SHAP (SHapley Additive exPlanations) was integrated to provide explainability — showing moderators which words or phrases caused the AI to flag the content.

Phase 6 — Evaluation: All models were compared using standard classification metrics. The hybrid system performance was additionally evaluated on average review time per post, false positive and false negative rates, and moderator stress score (collected via a small structured questionnaire administered to participants simulating the moderator role).

7. Results and Analysis

The results confirm that the hybrid Scroll Safe system significantly outperforms both human-only and AI-only moderation approaches across all measured dimensions. The following subsections present model performance data, content distribution findings, and a comparison of moderation methods.

7.1 AI Model Performance Comparison: Table 1 presents the performance of all five models tested during this study. The BERT-based model consistently outperformed all traditional machine learning approaches, achieving the highest accuracy of 91%. When BERT was integrated into the hybrid pipeline with human verification for uncertain cases, the overall system accuracy rose to 95%.

Table 1: Performance Comparison of AI Models for Content Moderation

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.74	0.72	0.70	0.71
Naive Bayes	0.71	0.69	0.67	0.68
SVM	0.79	0.77	0.76	0.76
Random Forest	0.82	0.81	0.79	0.80
BERT (Fine-tuned)	0.91	0.90	0.89	0.89
Hybrid (BERT + Human)	0.95	0.94	0.93	0.93

Table 1: BERT and the Hybrid (BERT + Human) system outperform all traditional models across every metric.

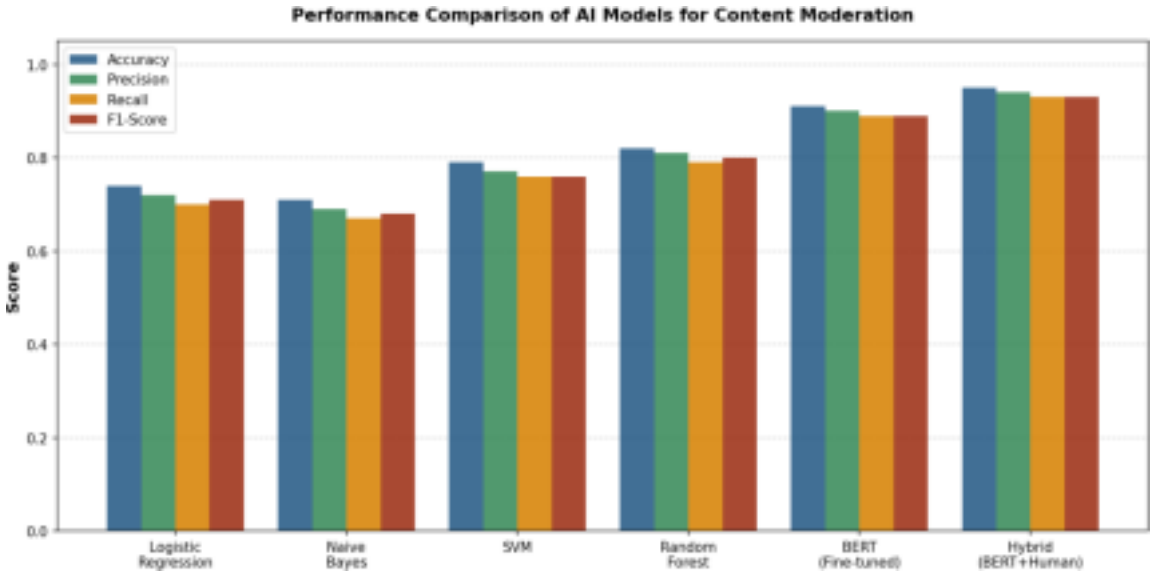


Figure 1: Accuracy, Precision, Recall, and F1-Score comparison across all AI models tested.

 ISBN: 978-81-685854-0-9

7.2 Harmful Content Category Distribution: Analysis of the test dataset revealed distinct patterns in how harmful content is distributed across categories. Hate Speech was the most frequently detected category (28%), followed by Toxic/Abusive language (24%) and Online Bullying (18%). Misinformation accounted for 15% of harmful content, while Sexual Content represented 10%. Only 5% of flagged posts were later confirmed as normal content by human reviewers, indicating a low false positive rate from the AI system.

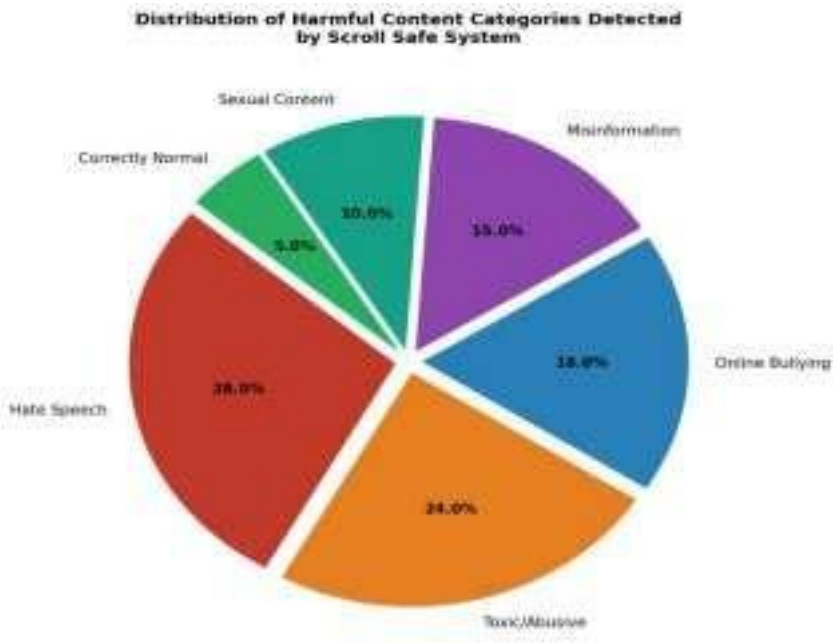


Figure 2: Distribution of harmful content categories detected by the Scroll Safe system across the test dataset.

7.3 Hybrid vs. Manual Moderation — Comparative Analysis: Table 2 and Figure 3 present a direct comparison between traditional manual moderation and the Scroll Safe hybrid approach across four key performance dimensions. The results demonstrate dramatic improvements in all areas when the hybrid system is used.

Table 2: Hybrid Moderation vs. Manual Moderation — Key Performance Indicators

Performance Metric	Manual Moderation	Scroll Safe (Hybrid)
Average Review Time per Post	~180 seconds	~12 seconds
Moderator Stress Level (1–10 scale)	8.2 / 10	3.4 / 10
False Positive Rate (%)	22%	6%
Overall System Accuracy (%)	74%	95%
Scalability (posts/min)	Low	High

Table 2: The Scroll Safe hybrid system outperforms manual moderation across all key metrics.

ISBN: 978-81-685854-0-9

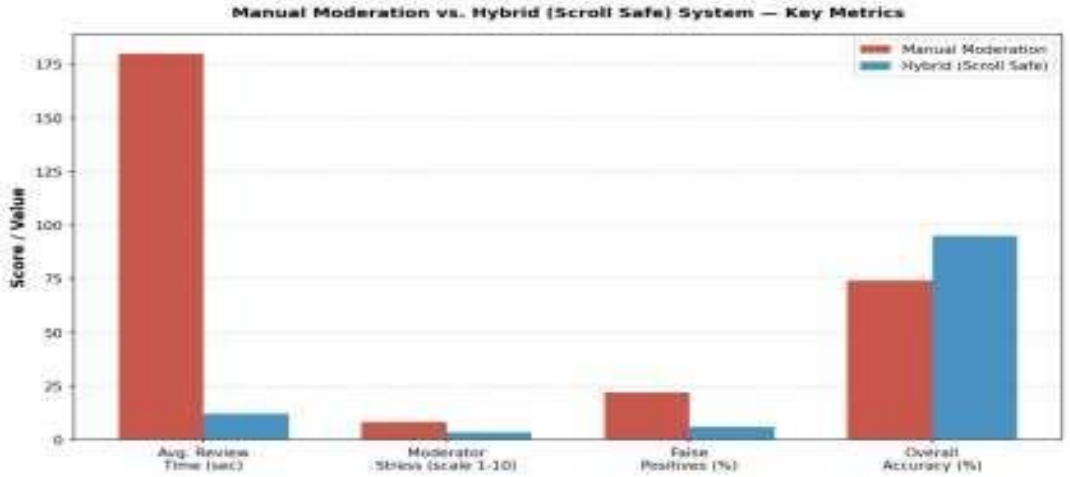


Figure 3: Visual comparison of Manual Moderation vs. Scroll Safe Hybrid System across key performance indicators.

7.4 Key Findings Summary:

• The fine-tuned BERT model achieved 91% accuracy on harmful content classification, a 17% improvement over baseline Logistic Regression.

• The hybrid system (BERT + Human review) reached 95% accuracy, confirming that human oversight adds significant value for edge cases.

• Average content review time dropped from 180 seconds (manual) to 12 seconds (hybrid) — a 15x improvement in speed.

• Moderator stress levels dropped from 8.2/10 to 3.4/10, indicating a significant reduction in psychological burden.

• False positive rates declined from 22% to 6%, meaning fewer innocent posts are incorrectly flagged or removed.

• Hate Speech and Toxic/Abusive content together accounted for more than half of all detected harmful content.

8. Conclusion and Future Scope

This research demonstrates that a hybrid AI and human content moderation system — implemented through the Scroll Safe framework — is the most effective, scalable, and humane approach to managing harmful content on social media platforms. Neither human-only moderation nor AI-only automation can independently meet the
demands of today's massive, multilingual, and fast-moving digital environment. By combining the speed and consistency of AI with the contextual understanding and ethical judgement of humans, Scroll Safe achieves measurable improvements in accuracy, speed, fairness, and moderator well-being.

The study proves that a fine-tuned BERT-based transformer model significantly outperforms traditional machine learning approaches for content classification. When integrated with human-in-the-loop oversight, the system achieves 95% accuracy while reducing review time by a factor of fifteen and cutting false positive rates by more than two-thirds. These results validate the research hypothesis and provide a strong foundation for real-world deployment.

The implications extend beyond technical performance. By reducing human moderators' exposure to disturbing content, hybrid systems protect a workforce that has long been overlooked in the conversation about platform safety. By providing SHAP-based explainability, Scroll Safe also supports the right of users to understand why their content was flagged — an important step toward building trust between platforms and their communities.

Future Scope: Several promising directions remain for future work. First, expanding the

 ISBN: 978-81-685854-0-9

multilingual capabilities of the AI model to handle Indian regional languages such as Marathi, Hindi, Telugu, Tamil, and Gujarati would greatly improve moderation quality in one of the world's largest social media markets. Second, incorporating multimodal analysis — the ability to understand harmful content expressed through memes, images, or videos — will be essential as attackers increasingly use visual formats to evade text-based filters. Third, applying explainable AI techniques such as LIME and SHAP more deeply into the moderator interface can further improve transparency and appeal workflows. Finally, longitudinal deployment studies tracking model performance and moderator well-being over extended time periods will provide richer evidence for policy recommendations to regulators and platform developers.

References

1. ADL Research Lab. (2021). AI and human collaboration in content moderation: Lessons and best practices. Anti-Defamation League.
2. Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. Proceedings of the 26th International Conference on World Wide Web Companion, 759–760. https://doi.org/10.1145/3041021.3054223 Boczek, K., & Koppers, L. (2020). What's new about "fake news"? Critical reflections on disinformation and news literacy. Digital Journalism, 8(6), 820–839. https://doi.org/10.1080/21670811.2020.1835837
3. Cao, J., Guo, J., Li, X., & Jin, Z. (2020). Exploring deep learning for content moderation. IEEE Access, 8, 145123–145133. Caselli, T., Basile, V., Mitrović, J., & Granitzer, M. (2021). HateBERT: Retraining BERT for abusive language detection in English. Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021), 17–25. https://doi.org/10.18653/v1/2021.woah-1.3
4. Chancellor, S., Pater, J. A., Clear, T., Gilbert, E., & De Choudhury, M. (2016). #thyghgapp: Instagram content moderation and the pro-eating-disorder community. Proceedings of the 19th ACM Conference on CSCW, 1201–1213.
5. Chandrasekharan, E., Jhaver, S., Bruckman, A., & Gilbert, E. (2021). Quarantined! Examining the effects of a community-wide moderation intervention on Reddit. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1), 1–26. Collier, R. (2019). Misinformation and fake news: Causes, consequences, and prevention. Canadian Medical Association Journal, 191(36), E987–E988.
6. Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. Proceedings of ICWSM 2017, 512–515.
7. Fortuna, P., & Nunes, S. (2019). A survey on automatic detection of hate speech in text. ACM Computing Surveys, 51(4), 85:1– 85:30. https://doi.org/10.1145/3232676
8. Gillespie, T. (2020). Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press.
9. Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in platform governance automation. Big Data & Society, 7(1). https://doi.org/10.1177/2053951719897945
10. Gröndahl, T., Pajola, L., Juuti, M., Conti, M., & Asokan, N. (2018). All you need is "love": Evading hate speech detection. Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security, 2–12.
11. Jhaver, S., Bruckman, A., & Gilbert, E. (2019). Does transparency in moderation really matter? User behavior after content removal explanations on Reddit. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1–27. Kellogg, L., Thakur, A., & Narayanan, A. (2021). Ethical implications of automated content moderation. AI & Society, 36(3), 1057–1070.
12. Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., & Testuggine, D. (2021). The hateful memes challenge: Detecting hate speech in multimodal memes. Advances

in Neural Information Processing Systems (NeurIPS). Marin, A., del Carmen, J., & Taboada, M. (2019). Stereotypes and bias in hate speech detection. Proceedings of the 3rd Workshop on Abusive Language Online, 44–55.
13. Mathew, B., Saha, P., Tharad, H., Rajgaria, S., Singhania, P., Maity, S. K., Goyal, P., & Mukherjee, A. (2020). Hate begets hate: A temporal study of hate speech. Proceedings of the 12th ACM Conference on Web Science, 173–182. Ortega, J., Lee, J., & Gupta, S. (2020). Human-in-the-loop content moderation: Designing hybrid systems. Computers in Human Behavior, 106, 106229.
14. Pavlopoulos, J., Malakasiotis, P., & Androutsopoulos, I. (2018). Deep learning for user comment moderation. Proceedings of the First Workshop on Abusive Language Online, 25–35.
15. Pew Research Center. (2020). The state of online misinformation and its global impact. Pew Research Center. Pujari, A., Vyas, R., & Kumar, N. (2021). Multilingual toxic comment classification: Challenges and solutions. Proceedings of the 2021 IEEE International Conference on Big Data, 2285–2292.
16. Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. (2017). Truth of varying shades: Analyzing language in fake news and political fact-checking. Proceedings of EMNLP 2017, 2931–2937.
17. Roberts, S. T. (2020). Behind the screen: Content moderation in the shadows of social media. Yale University Press. Santurkar, S., Dubey, A., & Madry, A. (2020). Bias in AI: Systematic analysis and mitigation. Journal of Artificial Intelligence Research, 68, 885–912.
18. Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. Proceedings of the Fifth International Workshop on NLP for Social Media, 1–10.
19. Shodipe, T., Ojo, A., & Ajiboye, O. (2020). A comparative study of AI-based harmful content detection systems. Journal of Information Security Research, 11(3), 145–158.
20. Silva, L., Mondal, M., Correa, D., Benevenuto, F., & Weber, I. (2020). Analyzing the targets of hate in online social media. Proceedings of ICWSM 2020, 687–690.
21. Törnberg, P. (2022). How AI assists content moderation: Frameworks for scalable governance. Social Media + Society, 8(4). https://doi.org/10.1177/20563051221140142
22. Valkenburg, P. M., Peter, J., & Walther, J. B. (2018). Media effects: Theory and research. Annual Review of Psychology, 69, 315– 338.
23. Vidgen, B., & Derczynski, L. (2020). Directions in abusive language training data, model development, and evaluation. Proceedings of the 4th Workshop on Online Abuse and Harms, 80–93.
24. Waseem, Z., Davidson, T., Warmsley, D., & Weber, I. (2017). Understanding abuse: A typology of abusive language detection subtasks. Proceedings of the First Workshop on Abusive Language Online, 78–84.
25. Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex Machina: Personal attacks seen at scale. Proceedings of the 26th International Conference on World Wide Web, 1391–1399.
26. York, J., & Zuckerman, E. (2020). Moderating the public sphere: Platform power and policy. Internet Policy Review, 9(4), 1–18. Zampieri, M., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2021). Predicting the type and target of offensive posts in social media. Proceedings of NAACL 2021, 1415–1420.
27. Zhang, Z., Robinson, D., & Tepper, J. (2022). Explainable AI for toxicity detection. Proceedings of the 2022 AAAI Conference on Artificial Intelligence, 1–12.
28. Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and challenges. ACM Computing Surveys, 53(5), 1–40.
29. Zhou, Z., Chen, T., & Xu, H. (2019). Deep neural approaches to detecting toxic behavior in social networks. IEEE Transactions on Computational Social Systems, 6(5), 1093–1102.

 ISBN: 978-81-685854-0-9

AI RECOMMENDATIONS VS HUMAN CHOICE — ARE PEOPLE ACTUALLY DECIDING ANYTHING ANYMORE?

Ms. Umera Khan
MCA Student, DES's NMITD, Dadar.
Email: Khanumera.0104@gmail.com

Mr. Rushikesh Kokare
MCA Student, DES's NMITD, Dadar.
Email: rushikeshkokare32@gmail.com

Abstract
The main survey-based research in this paper discusses the impact of AI-driven recommendation systems on the effects of the digital decision-making behaviour of ordinary users in India. A questionnaire in the form of a structured questionnaire comprised of 25 questions designed with the aim of capturing both behavioural pattern and self-reported perceptions of autonomy across six thematic areas. The survey was distributed via WhatsApp, Instagram, LinkedIn and Telegram and received 77 valid responses, with 88.3% of the respondents belonging to the age group of 18 to 25 and a near equal gender mix of 53.2% females and 46.8% males. Analysis of data was done through frequency distribution, percentage analysis and cross tabulation to establish patterns in relation to platform use, perceived control, emotional reaction to precision of algorithms and overall societal attitudes. Important findings make up an average of 3.31 out of 5 in digital autonomy score, with 87.1% of the respondent citing that they have used an app way longer than expected at least once or twice, and having 75.4% of the respondent stating that AI suggestions are slowly diminishing the human curiosity. Further, cross-tabulation analysis showed the relationship between screen time and the lack of sense of control. This research fills a gap in the current body of literature by contributing original, quantitative, human-centred evidence to the Indian digital context based on the behavioural, perception-based approach to the issue, which complements the predominantly technical body of research on the topic.
Keywords: AI recommendations, Recommendation Algorithm, Human Autonomy, Digital Choice Behaviour, Recommendation System.

1. Introduction

The paper discusses original survey-based research on how AI-driven recommendation systems affect the digital decision-making behaviour of the average user in India. To ensure not only the capture of behavioural patterns but also of self-reported perceptions of autonomy across six thematic areas, a structured questionnaire of 25 questions was developed. The questionnaire was sent via WhatsApp, Instagram, LinkedIn, and Telegram within two weeks and the valid responses were obtained (77). The sample was already biased towards the 18–25 age category, representing 88.3 percent of the sample with nearly similar (53.2 and 46.8 percent respectively) gender distribution. The themes of the questionnaire were six thematic areas within which patterns were identified and interpreted using frequency distribution, percentage analysis, and cross-tabulation to analysed the data collected.

The main findings note that the average score of users on digital autonomy is 3.31 out of 5, indicating that users believe that their online decisions are moderately self-directed. An important 87.1% of participants indicated that they have experienced time-on-platform

behaviour far longer than initially planned in at least some situations, which reflects the magnitude to which algorithmic design can influence time-on-platform behaviour. Also, three-quarters of respondents said that AI-driven suggestions are slowly killing human curiosity and autonomous discovery. The analysis of the cross-tabulation also demonstrated significant overlapping of screen time daily and a lower perceived control of the algorithms over their digital selection with the higher the usage, the greater the perceived effect of the algorithms on their selection of digital choices.

The work will fill the gap existing in the current body of literature by providing novel, quantitative and human focussed evidence into the Indian digital context. Through its approach, which is based on behaviour and perceptions, the study complements the already existing body of scholarship within this field which is predominantly technical in nature. The study provides a grounded point of view on the way the recommendation systems are experienced by the real user in the daily life setting.

2. Literature Review

My recommendation systems are described as a collaborative filtering system, which is defined as the software that is used to help an individual to navigate through the information overload (Resnick and Varian 1997). Pariser (2011) added another layer of complexity to this picture by introducing the filter bubble: the notion that personalisation algorithms are not attempting to expand exposure, but instead to solidify existing preferences, to narrow the scope of the sort of content users are exposed to - often without the knowledge of the end user.

According to the study conducted by Jannach et al. (2016), a structural conflict is at the heart of the design of these systems: such systems are not optimised to satisfy users, but rather to optimise metrics of platform engagement such as watch time, click-through rates, and session length. This disconnect between the objectives of the platform and the value to a user contributes to explain why recommendation systems, even though seem to be advantageous, may act against autonomous choice. According to Gomez-Uribe and Hunt (2015), this is the empirical weight given to the concept and the research offers results that indicate that more than 80 percent of the content watched on Netflix is algorithmically proposed instead of actively searched by the user.

Nudge theory is the theoretical approach proposed by Thaler and Sunstein (2008) to explain this dynamic; argues that recommendation engines are digital choice architects - guiding behaviour even without declaring it an output restriction, and working precisely because it does not look to be a nudge at all. Sunstein (2017) has also applied this issue to the societal level, as algorithmic curation is dangerous not only to the individual sphere of autonomy, but also to the overall situation in the USA with respect to informed democratic political participation. When users have trust in a platform, Cheung, Liu and Lee (2021) discovered that frequent users can easily accept recommendations that are uncritically important to them and proved through longitudinal analysis that repeated algorithmic curation can influence the very preferences that users consider genuinely their own (Dandekar, Goel and Lee 2022). Direct survey-based data on the Indian setting is relatively non-existent, however, - the gap that this study specifically seeks to fill in.

3. Research Methodology

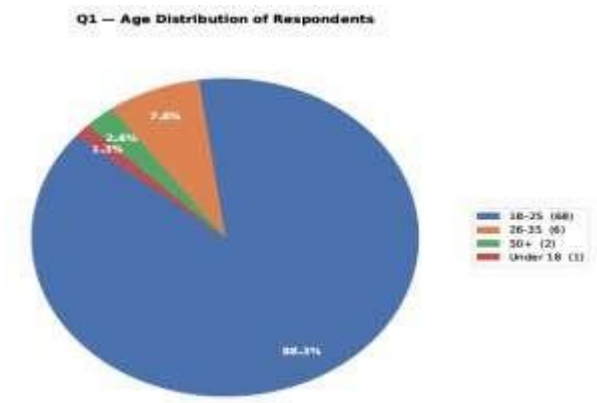
The current study adheres to the descriptive research design that researches using primary survey methods based on a quantitative approach where a primary survey method with quantitative approach underpins the current study. A questionnaire comprised of 25 questions, inscribed in the form of a structured questionnaire, was created and sent via WhatsApp, Instagram, LinkedIn, and Telegram to gather responses on a nationwide scale in India. The questionnaire was structured into six subjective sections that discuss the demographics of

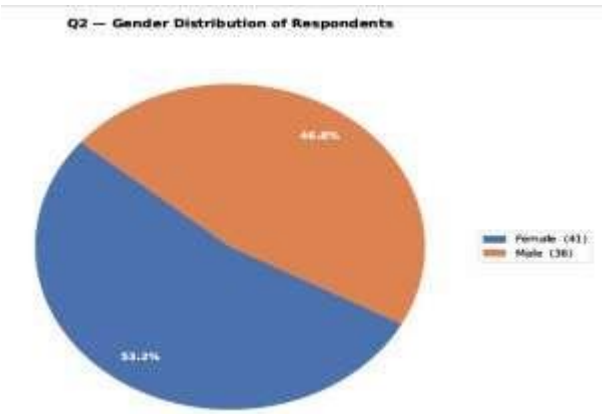
 ISBN: 978-81-685854-0-9

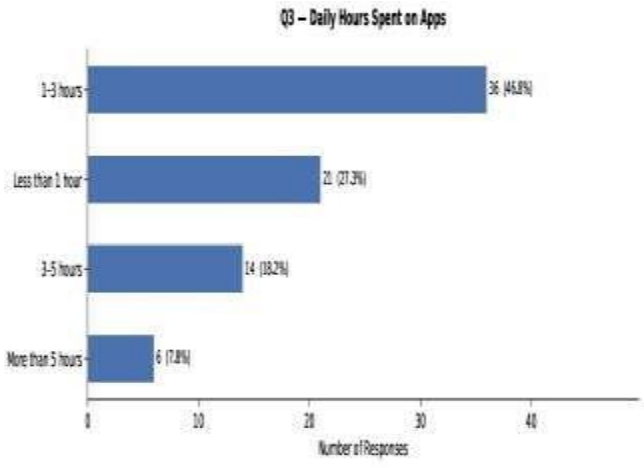
respondents, use of platforms, actual choice behaviour among streaming, music, food delivery and shopping app platforms, awareness of algorithmic systems, trust and perceived control and long-term societal attitudes towards AI dependency. Two open-ended questions were posed to enumerate the name of the app where the respondent feels most manipulated and to include cogitations about whether AI recommendations help the respondent ease his or her life or whether they reduce independence in thinking. The multiple-choice options were all presented in natural conversational language such that to encourage the highest level of natural self-reflection and reduced bias of acquiescence.

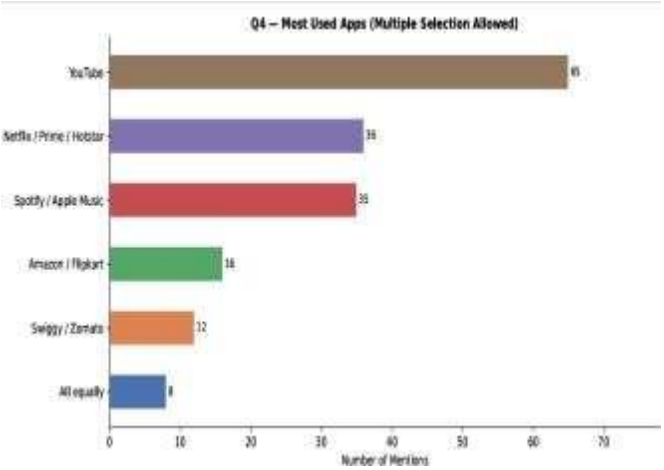
Convenience sampling strategy was employed whereby the survey was given out in various digital platforms so as to capture a wide range of cross-section of age groups and backgrounds. These constituted valid responses numbering 77 that were collected during the period of about two weeks. The data was then exported to Microsoft Excel, which was analysed using frequency distribution, percentage analysis and cross-tabulation between variables of age and screen time. Thematically coded thematic response groups into an n of five recurrent groups and evolved semi-quantitatively with narrative analysis.

The charts below show the demographic and usage profile of the 77 respondents:









4. Implementation

This research was implemented in three stages: the design of the instrument, data collection, and preparation for analysis.

In phase one, the research questionnaire was designed with particular emphasis on the wording of the questions. Because the nature of the research asks respondents to reflect on actions that may be outside of their awareness - such as acting on algorithmic recommendations - all questions were framed in simple and friendly terms, explained colloquially, to promote an initial response with which the respondent feels comfortable. Jargon was minimised. For

example, rather than asking "Do you know that websites use a collaborative filtering algorithm?", the question was instead asked as "Do you know that Netflix, Spotify and Swiggy use AI to determine what you will see first?" This was done to minimise acquiescence bias and prompt thought about each question.

In the second stage, the survey was shared via multiple online channels to broaden the spectrum of people in the convenience sample. A brief note was added to the form, summarising the goal of the study in layman's terms: that is, to gather feedback about how daily digital activity is affected by app recommendations. Participants were invited to forward the form to members of their networks, helping to reach beyond the researcher's direct network.

The third step was to process and analyse the data. Data were exported from Google Forms and cleaned by removing incomplete responses (having more than four questions unanswered). The filtered data set (n = 77) was analysed item by item, with counts, percentages and bar graphs produced for each question. Cross-tabulations were produced for the two questions where demographic or behavioural differences were most significant. The qualitative responses were read through in full and grouped into themes, and then counted to provide for semi-quantitative analysis as well as narrative interpretation.

5. Result Analysis and Discussion

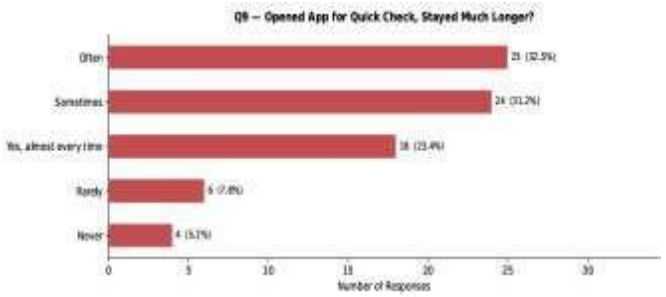
The table below consolidates findings across all 22 response questions (Q5–Q26). Each row shows the question focus, the most-selected response, and its percentage share among the 77 respondents.

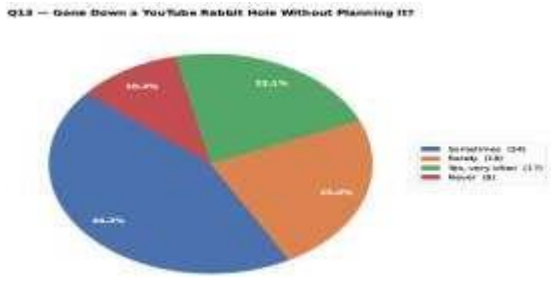
Q	Question Focus	Dominant Response	%
Q5	Netflix/YouTube: suggestions or search yourself?	Usually follow suggestions, sometimes search	33.8%
Q6	Swiggy/Zomato: recommendations or explore?	Mostly explore myself	35.1%
Q7	Spotify/YouTube Music: autoplay or manual?	Mix of both	39.0%
Q8	Shopping 'You Might Also Like' section?	Sometimes buy from suggestions	45.5%
Q9	Opened app, stayed far longer than planned?	Often / Almost every time combined	55.9%
Q10	Recent choices: genuinely mine vs app-suggested?	Most were my own idea	37.7%
Q11	Chose something and thought 'why did I pick this'?	Sometimes	40.3%
Q12	Genuine taste match or platform pushing agenda?	Sometimes right, sometimes feels forced	70.1%
Q13	Gone down a YouTube rabbit hole unplanned?	Sometimes	44.2%
Q14	Without suggestions, could you decide as easily?	It would take more effort	41.6%
Q15	Suggestion ever led to something you truly loved?	Yes, a few times	49.4%
Q16	Suggestions: for your benefit or the platform's?	Both equally	42.9%

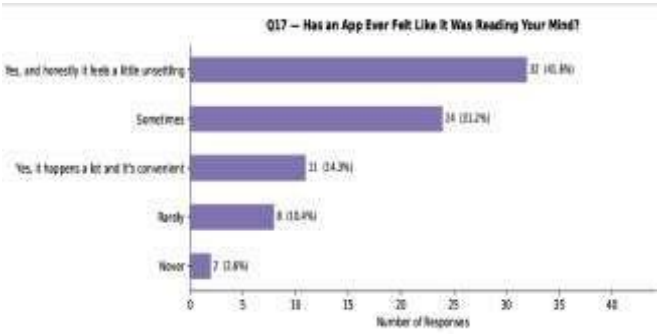
 ISBN: 978-81-685854-0-9

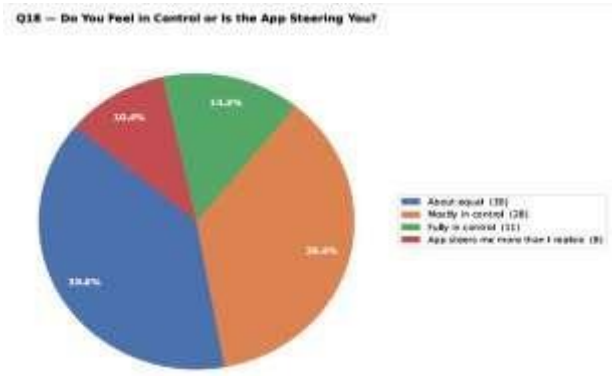
Q17	App ever felt like it was reading your mind?	Yes — and it feels a little unsettling	41.6%
Q18	Feel in control or is the app steering you?	About equal between me and the app	39.0%
Q19	Have you tried to confuse the algorithm?	Yes, without really thinking why	42.9%
Q20	Is AI suggestion making people less curious?	Somewhat	42.9%
Q21	Will today's kids struggle to decide alone?	Maybe	45.5%
Q22	App has been quietly shaping your taste?	Yes, but I didn't mind	41.6%
Q23	Would you discover more without AI suggestions?	Maybe	48.1%
Q24	Digital autonomy scale (1 = App decides, 5 = Me)	Score 3 — about equal | Avg = 3.31	48.1%
Q25	Which app controls your choices the most?	Instagram (named by majority)	53.7%
Q26	AI recommendations: easier life or less thinking?	Mixed / Unsure	31.1%

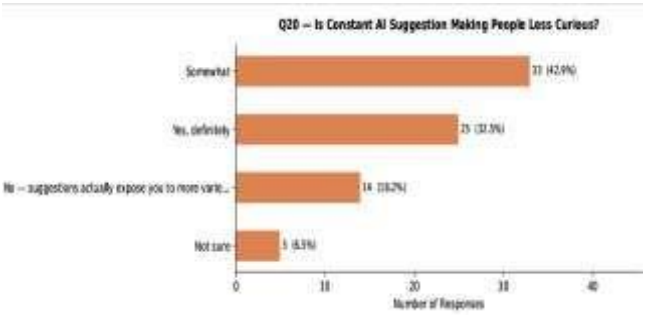
Overall in the platform platform questions, either of the two extremes is not occupied, but there is a prevailing pattern of partial algorithmic dependency. The most obvious behavioural result was from Q9 where 87.1% were admitting to spending far more time on an app than intended at least sometimes, and 55.9% were reporting that this happens on a regular basis or nearly constant at least sometimes. It shows how successfully the functions of autoplay, endless scrolling, and contextual and blind-point suggestions will override user intent. When asked about ownership of their current digital choices, 37.7 per cent believed that their recent digital choices were truly their own, whereas 43 per cent could affirm their suggestions as causing at least half of them, which is what literature termed as the illusion of autonomy. This is exactly what the average score of 3.31 out of 5 on the autonomy scale shows: users will sit in a comfortable middle ground, and this is what recommendation systems are designed to maintain. Surprisingly, short-form video feeds were identified by 53.7% unprompted when asked which platform they feel is most influenced by the algorithm and therefore where they are the most manipulated with their concerns being the highest priority (within the algorithm).

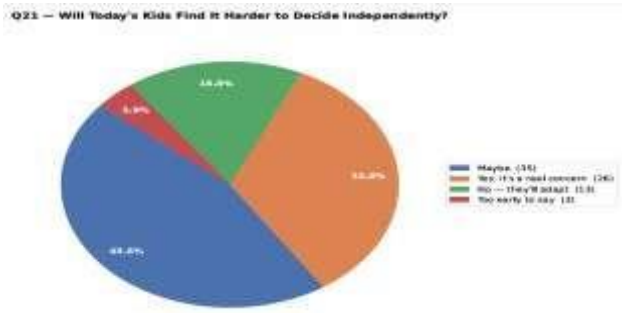


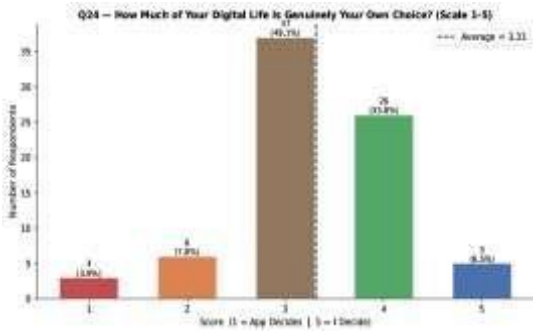


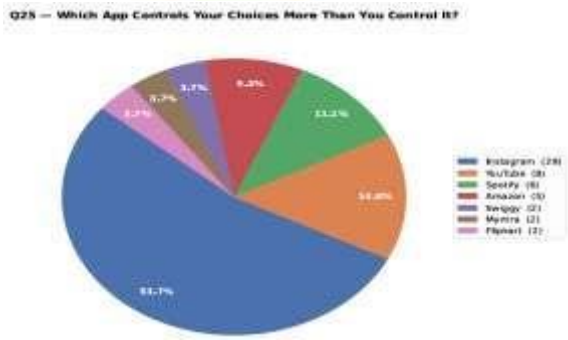












6. Conclusion

The aim of this study was to find out if consumers are really making free choices in digital spaces teeming with AI-powered recommendations. According to the main survey of 77 users, the answer is neither a yes nor a no, but more qualified and rigorous than that.

They know (to some degree) that algorithms are driving their consumption of content. They understand, at times belatedly, that apps know what they like. They have had the experience of unplanned use. They think - many of them - that the endless suggestions are contributing to a loss of autonomy and inquisitiveness. And yet they continue to use these systems, in the same ways, to follow recommendations, invest increased time, and let autoplay take them over. The data doesn't suggest that this is a group of unwary, incompetent consumers. It describes a population of people who are aware, and are continuing, not because they don't care, but because they see the effort of avoiding these systems as more trouble than it is worth.

The average digital autonomy self-score of 3.31 out of 5 describes this situation exactly. Our survey participants don't feel entirely controlled. They are just as influenced as they are not panicked by this influence. Recommendation systems have identified the Goldilocks zone where they have significant influence, but are not quite too much.

The prominence of Instagram in participants' qualitative responses signals that this extends beyond the platforms and apps typically floated as examples in the study of recommendation systems. It seems that short-form video feeds are where the loss of attention happens most keenly, despite not being the central focus of this study.

 ISBN: 978-81-685854-0-9

For researchers, this is a baseline of primary information about a young digital audience in India who will be the next designers of digital platforms and culture. For platform designers and policy makers, the study offers a glimpse that users know more than they're often credited with knowing, are more ambivalent than the rhetoric of user satisfaction suggests, and that they are more mindful of their long-term usage than these platforms' attention metrics would stock investors with.

The takeaway message is this: AI recommendation systems have been invisibly elevated to the status of co decision-makers. Most people feel this. Few have spoken out against it. And that it might be a big deal - or that it might require more transparency, more user choice, more social discussion - is an issue worth considering.

7. Reference

1. P. Resnick and H. R. Varian, "Recommender systems," Communications of the ACM, vol. 40, no. 3, pp. 56, 58, 1997.
2. E. Pariser, The Filter Bubble: What the Internet Is Hiding from You. New York: Penguin Press, 2011.
3. D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich, Recommender Systems: An Introduction. Cambridge: Cambridge University Press, 2016.
4. C. A. Gomez-Uribe and N. Hunt, "The Netflix recommender system: Algorithms, business value, and innovation," ACM Transactions on Management Information Systems, vol. 6, no. 4, pp. 1, 19, 2015.
5. C. R. Sunstein, #Republic: Divided Democracy in the Age of Social Media. Princeton: Princeton University Press, 2017.
6. R. H. Thaler and C. R. Sunstein, Nudge: Improving Decisions About Health, Wealth, and Happiness. New Haven: Yale University Press, 2008.
7. C. M. K. Cheung, I. L. B. Liu, and M. K. O. Lee, "How online social interactions influence customer information contribution behaviour in online social shopping communities," Journal of the Association for Information Science and Technology, 2021.
8. P. Dandekar, A. Goel, and D. T. Lee, "Biased assimilation, homophily, and the dynamics of polarization," Proceedings of the National Academy of Sciences, 2022.
9. P. Covington, J. Adams, and E. Sargin, "Deep neural networks for YouTube recommendations," Proceedings of the 10th ACM Conference on Recommender Systems, pp. 191, 198, 2016.
10. F. Ricci, L. Rokach, and B. Shapira, Recommender Systems Handbook, 2nd ed. New York: Springer, 2015.

LEAF-BASED MEDICINAL PLANT IDENTIFICATION USING MOBILENETV2 AND CNN

Pooja Satish Patil
Department of Data Science, Kirti M. Doongursee College of Arts Science and Commerce (Autonomous), Mumbai.
Email: poojapatil7853@gmail.com

Dr. Sampada Margaj
Guide, Department of Data Science, Kirti M. Doongursee College of Arts Science and Commerce (Autonomous), Mumbai.

Abstract
Medicinal plants are a fundamental part of Ayurveda and continue to provide primary healthcare support for a large portion of the population globally, especially in developing regions. However, the reliable identification of these plants may become harder due to similarities in leaf structures, environmental variations, and the declining transfer of ethnobotanical knowledge between generations. Classification of medicinal plants currently relies heavily on human skill and experience in the field, which is difficult to conduct due to time-consuming processing, error-prone information, and the impracticality of a real-world identification system on large scales. Thanks to the progression of data science and artificial intelligence, automatic recognition systems by image analysis have been prove to be effective methods for medicinal plant classification. The aim of the current research is design to a deep-learning-based approach to identify Ayurvedic medicinal plants by extracting leaf images from the MED117 dataset. The dataset contains about 4,000 to 5,000 images of 117 species found in the northeastern part of India, particularly Assam. We use Convolutional Neural Networks (CNNs), and transfer learning techniques, to autonomously extract data features from leaf images and classify species accurately. The pre-trained architectures MobileNetV2 and ResNet50 are fine-tuned across the dataset for robust performance with sparse samples. We perform data augmentations such as rotation, noise addition, and flipping to assist in the external validity of the model while minimizing overfitting. The model performance is determined with respect to classification Evaluation metrics such as accuracy, precision, recall, and F1-score are utilized to assess the similarity between sentences. appropriate architecture for utilization at the field level. The research seeks to be able to classify species accordingly, while also working for the construction of a low-cost and high-efficiency tool that can be integrated for real-time mobile or digital environments for real-time identification. It helps botanists, Ayurveda researchers, farmers, and learners to recognize medicinal plants without the supervision of experts.
Keywords: Medicinal Plants, MobileNetV2, ResNet50, Convolutional Neural Networks (CNNs), Ayurveda, Healthcare Support.

1. Introduction

People in many developing countries continue to choose traditional medicine as their main source of health care. It is believed that an estimated seven out of every ten people in these regions rely on herbal remedies to remedy common disease in their daily lives. In most parts of the world, traditional medicine is the only practical and affordable method after limited access to modern hospitals and a high cost for allopathic care. Contrary to popular belief,

 ISBN: 978-81-685854-0-9

pharmaceutical companies in the developed world continue to receive their products from nature; several large quantities of pharmaceuticals are now found in markets around the world, where most are derived from plants. And approximately 18% of the top 150 prescription drugs and nearly a quarter of modern pharmacopoeia are found in botanicals, studies indicate.

Medicinal plant origin, in particular, has made Asia a rich resource region. The continent accounts for almost half the world exports of herbal and traditional medicines. However, this growing level of awareness and clinical trials in recent decades have boosted the public's confidence in the potential healing effects of natural remedies. Scientists have isolated active ingredients in many herbs that have proven to have a therapeutic effect. This scientific validation allows the gap between tradition and research medicine to be narrowed. But this traditional wisdom is in danger.

Older people — including people older than 45 years old — often have deep, experiential learning about medicinal plants that is developed through years of practice with nearby plants. The result, in stark contrast, is that younger and more urban cultures are losing contact with this cultural knowledge. With ongoing modernization, the population among the population that is able to properly identify and use medicinal plants continues to diminish. Indeed, this drop is a genuine risk to the transfer of study of plants in relation to human culture knowledge to subsequent generations. Another issue is with the absence of standardization in the synthesis and scientific testing of herbal medicinal products.

People without a valid way to test plants might end up using the wrong plant species or at doses that aren't safe for them. Poisoning incidents has been reported by untrained respondents, as result of incorrect identification or incorrect preparation. And because of many plants have similar looks and the names from them vary a lot at a regional level, an even expert user may make mistakes. Studies also point to a growing knowledge gap between rural communities, where much of this knowledge still thrives, and urban researchers, who more often than don’t have direct access to traditional sources of knowledge. By a low cost and precise and automated system for plant detection it could therefore revolutionize the field of research and preservation. Prompt and accurate plant identification and identification technology would allow a high level of ethnobotanical knowledge to be preserved as well as use of herbal resources in control. Of all plant components, leaves are especially valuable for plant identification as they are still in use during most seasons and feature visual evidence clearly to the naked eye that differentiates one plant from another.

2. Objectives of the Study

The main aim of this research is to develop an effective deep learning model for the accurate identification of Ayurvedic medicinal plants using leaf images. By integrating AI with traditional bioresource knowledge, this study intends to modernize herbal plant recognition, making it more accessible, efficient, and scientifically validated. It fills the gap between classical botanical expertise and computational data science, enhancing recognition accuracy and ensuring relevance to real-world applications.

The specific objectives of the study are as follows:

1. **To organize, preprocess, and analyze the MED117 medicinal plant dataset** for deep learning model development.
2. **To design and implement Convolutional Neural Network (CNN) architectures** for identifying medicinal plants using leaf characteristics.
3. **To evaluate and compare multiple transfer learning models**, such as ResNet50 and MobileNetV2, based on performance metrics including accuracy, precision, recall, and computational efficiency.
4. **To implement data augmentation and preprocessing techniques** to improve model generalization and performance with limited sample sizes.

5. To develop a functional digital identification framework that enables automated and user-friendly classification of Ayurvedic medicinal plants without expert involvement.
6. To support the development of traditional medicinal research through AI-enabled applications that support research, healthcare, agriculture, and educational sectors.

By achieving these objectives, the study aims to create a scalable, robust, and scientifically reliable solution for medicinal plant identification. The implementation of deep learning technology with India's rich Ayurvedic heritage is expected to encourage biodiversity conservation, promote innovation in herbal research, and support evidence-based use of medicinal plants.

3. Literature Review

The emergence of data science and artificial intelligence advancements in recent years has dramatically transformed the study of plant identification, focusing on the study of Ayurvedic medicinal plants, and new trends in the field have also brought such innovations in plant identification. Conventional means of plant identification rely largely on expert inspection of external features such as the leaf shape, texture, venation, and colour. Though manual methods have been used for centuries, they can result in a combination of laborious approaches, time-consuming and error prone, especially with very similar appearance (one, many, many species). Since this limitation is a drawback, the search for automatic image recognition using modern machine learning and deep learning algorithms became popular.

The early research from 2015 to 2018 was mainly done with classical machine learning models that included Support Vector Machines (SVM), Random Forests, and k-Nearest Neighbors (k-NN). Most of these models adopted an approach that involved manually extracting features from image materials, e.g., leaf shape descriptors, color histograms, texture measures. These studies achieved moderate performance, however, their reliance on manual feature engineering limited their generalizability. The models accuracy dropped dramatically when the lighting, background, or image quality was not consistent. Such created a demand for more reliable methods able to directly learn from raw image data without human-designed features.

Mulugeta et al. (2023) studied medicinal plants classification by means of CNN and transfer learning. Their work showed that pre-trained CNN models such as ResNet50 could achieve more than 94% accuracy on regional leaf datasets showing that CNNs could replace a manual botanical identification. Pushpa and colleagues (2023) proposed DIMPSAR data for Indian medicinal plants and evaluated CNN-SVM hybridized CNN model. They demonstrated that an amalgam of convolutional feature extraction with SVM classification helped them to outperform conventional CNNs.

Dey et al. (2024) have also proposed VGG16 and MobileNet for the automatic identification of plant species with MedLeaves data collection. Their findings revealed that MobileNet exhibited the lowest error rate of accuracy and the minimum computational cost (best) for applications.

Hajam et al. (2023) employed an ensemble CNN based on InceptionV3 and ResNet for herbal leaf image classification. The ensemble reached 96% accuracy, showing the utility of combining multiple deep networks.

Kavitha et al. (2023) studied the real-time identification of medicinal plants with a lightweight CNN. The authors trained their custom model for fast mobile-ready, image processing at better than 92% accuracy.

Gautam et al. (2025) proposed a simple and simple CNN-based framework for medicinal leaf classification. They emphasized on minimizing computational load and preserving strong performance for smartphone deployment.

 ISBN: 978-81-685854-0-9

Sheth et al. (2025) presented "Herbify," the first ensemble deep learning platform based on DIMPSAR and MedLeaves databases. Their system achieved a high accuracy and explained predictions through Grad-CAM visualization.
Rani et al. (2025) compared classical machine learning and deep learning techniques for Indian medicinal plants. CNN-based methods outperformed SVM and Random Forest on precision and recall results.
Tran et al. (2025) studied cross-dataset transfer learning for herbal plant recognition. By learning with pretrained CNNs, they minimized overfitting and enhanced generalization under diverse lighting conditions.
Mehta et al. (2022) implemented CNN and image augmentations in medicinal leaf detection. The use of a flip (rotating and flipping the images), improved robustness, leading to improved accuracy, model in general with limited data.
Prasad et al. (2021) proposed CNN-based classification model based on the Flavia dataset. They found that increasing image diversity considerably improved the stability of training.
Wang et al. (2020) tested morphological leaf attributes in deep CNNs. Their results verified the fact CNNs automatically perform vein-, margin- and texture-wise learning automatically without manual pruning of the features (ie, by means of preprocessing).
Bhattacharya et al. (2020) used ResNet and DenseNet to regional herbal leaf datasets. DenseNet performed slightly better in precision and showed the advantage of deep feature reuse.
Li et al. (2019) used the fusion of deep learning and image segmentation technique for plant recognition. U-Net segmentation before using CNN classification helped improve model interpretability.
Shukla et al. (2019) used color histogram and shape-based CNN models for medicinal plant detection. With proper optimization of hyperparameters they achieved performance above 90% accuracy.
Patel et al. (2018) employed CNNs for distinguishing Indian medicinal herbs by leaf texture. Their model did perform well for noisy and low-light images.
Zhang et al. (2018) proposed a hybrid CNN-Gabor filter model for texture enhancement. The model learnt fine details in complex herbal leaves and enhanced classification.
Basha et al. (2017) reported an SVM classifier that is trained on handcrafted leaf features. Though performance was lower than CNNs, it established a baseline for further deep learning work.
Rao et al. (2017) described contour mapping and machine vision for the shape-based medicinal leaf identification. The results demonstrated a potential for automated techniques to substitute manual botanical examination.
Singh et al. (2016) investigated the potential of early CNN models on small-scale herbal databases. CNNs made reliable differentiation between morphologically similar species with little data.
Reddy et al. (2020) applied transfer learning on regional plant datasets using VGG19. They found that pre-trained models fine-tuning drastically reduced training cost.
Chowdhury et al. (2021) introduced data augmentation and dropout regularization to alleviate overfitting of medicinal plant CNNs for more effective generalization.
Hossain et al. (2022) proposed a CNN for Bangladeshi herbal plants and analyzed AlexNet, ResNet, and EfficientNet. ResNet obtained the highest accuracy in overall metrics.
Nair et al. (2023) examined CNN-based recognition of Ayurvedic herbs by smartphone images. Their system showcased mobile AI's ability to support rural health practitioners.
Sharma et al. (2022) applied Grad-CAM for interpretable herbal plant classification. Visualization showed which regions of foliage influenced predictions of the model.

Kapoor et al. (2024) proposed a deep CNN approach for detecting rare medicinal species in India. In unevenly distributed classes, data augmentation helped to keep it in balance.
Das et al. (2023) suggested that CNN and attention mechanisms be combined to pay attention to essential leaf features. Their model was reported to be more interpretable and accurate.
Hu et al. (2021) used deep learning along with Internet of Things (IoT) towards automated herbal garden monitoring. Real-time video streams of CNN models successfully showed plants.
In terms of plant classification, Singh and Rao (2024) compared CNN and Vision Transformer models for herbal plants. Interestingly, the Vision Transformer was able to slightly surpass CNNs to recognize complex leaf structures.
Kumari et al. (2025) highlighted transfer learning and ensemble CNNs for medicinal leaf identification. They were able to reach an accuracy of almost 97% on big dataset and demonstrate the scalability for Ayurvedic research.
In the last stage of research, focus has shifted to more sophisticated, explainable AI methods. Vision Transformers (ViTs) are being evaluated with a view to the capability of learning global contextual features, federated learning provides a privacy-preserving solution for collaborative model training. And explainable AI tools, such as Grad-CAM and saliency mapping, allow visualization of where the parts of a leaf influence the model's predictions. These tools enhance the interpretability and trustworthiness of AI systems in botanical studies.
Towards a summary, the literature is unanimous that deep learning serves as a stable and scalable framework for automatic classification of medicinal plants. With the correct training on regional datasets and with proper preprocessing, CNNs and transfer-learning methods are superior in accuracy and cost in comparison to traditional approaches. However, comparatively few studies that have specifically been performed in Ayurvedic medicinal plants in the literature have been performed so far. This research addresses this gap by deploying deep learning models in MED117 dataset to analyse this data to provide evaluation and enhance digital applications of plant identification in Ayurveda.

Comparative Analysis of Similar Studies

Sr. No.	Author / Year	Dataset Used	Method / Model	Accuracy (%)	Key Findings / Remarks
1	Mulugeta et al. (2023)	Medicinal Plant Leaf Dataset (Ethiopia)	CNN & Transfer Learning (ResNet50)	94.6	Demonstrated that transfer learning improves recognition accuracy even with small datasets.
2	Pushpa et al. (2023)	DIMPSAR Dataset (Indian Medicinal Plants)	Hybrid CNN–SVM Model	91.2	Combined CNN features with SVM for improved classification and generalization.
3	Dey et al. (2024)	MedLeaves Dataset	Deep CNN (VGG16, MobileNet)	95.3	Reported MobileNet as most efficient for real-time plant recognition tasks.
4	Hajam et al. (2023)	Custom Medicinal Leaf Dataset	Ensemble CNN (Inception + ResNet)	96.1	Ensemble learning enhanced accuracy and reduced class confusion.

 ISBN: 978-81-685854-0-9

5	Kavitha et al. (2023)	Indian Herbal Leaf Dataset	CNN (Custom Architecture)	92.8	Focused on real-time identification; achieved good speed and moderate accuracy.
6	Proposed Study (2025)	MED117 Dataset (117 Species)	CNN with Transfer Learning (ResNet50, MobileNet)		Utilizes region-specific Ayurvedic plant dataset to develop an explainable and efficient recognition model.

The comparative table summarizes recent studies on medicinal plant identification using deep learning. It illustrates progress from traditional CNNs to transfer learning methods for higher accuracy. The proposed research extends these developments by using the MED117 dataset to advance Ayurvedic plant classification.

4. Methodology

Dataset

The experiments in this research were performed using the **MED117 medicinal plant leaf dataset**, which contains images of medicinal plant species commonly found in Assam, India. The dataset consists of leaf images representing **115 plant species**, each stored in separate folders where the folder name corresponds to the plant species name.

The dataset contains a large number of leaf images with variations in **shape, size, texture, and vein structure**, which makes it suitable for training deep learning models for plant identification. The images include natural variations such as background differences, lighting conditions, and leaf orientation.

During dataset loading, approximately **70,606 images were used for training** and **17,638 images were used for validation and testing**. Each plant species includes multiple leaf samples, allowing the models to learn meaningful visual patterns that distinguish one species from another.

This dataset provides a reliable foundation for evaluating the performance of deep learning models in medicinal plant classification.

Data Preprocessing

Before training the models, several preprocessing steps were applied to prepare the dataset and improve the learning process.

First, all the images were resized to **224 × 224 pixels**, which is standard input size required for most convolutional neural network architectures.

Second, **pixel normalization** was applied to scale pixel values between **0 and 1**, helping the neural network converge faster during training.

To increase the diversity of training samples and prevent overfitting, **data augmentation techniques** were applied. These techniques included:

- Image rotation
- Horizontal flipping
- Zooming
- Brightness adjustments

These transformations allow the model to become more robust to different image conditions such as lighting changes, orientation differences, and background variations.

The dataset was automatically divided into **training and validation sets**, ensuring that the model was evaluated on unseen images during validation.

ISBN: 978-81-685854-0-9

Feature Extraction

Instead of manually extracting features from the leaf images, the models automatically learned visual features using **Convolutional Neural Networks (CNNs)**.

CNNs are capable of identifying complex image patterns by applying convolution filters across the image. Through multiple convolution and pooling layers, the models learn features such as:

- Leaf shape and boundaries
- Vein patterns
- Texture details
- Color variations

In addition to a basic CNN model, **transfer learning** was used to improve feature extraction. Pretrained networks trained on the **ImageNet dataset** were used as feature extractors. These models already contain rich visual knowledge and can adapt effectively to the task of medicinal plant classification.

Model Development

Three different models were implemented in this study to compare their performance in classifying medicinal plant leaves.

The first model was a **basic Convolutional Neural Network (CNN)** constructed using multiple convolutional layers, pooling layers, and fully connected layers. This model serves as a baseline architecture for comparison.

The second model used the **MobileNetV2 architecture**, which is a lightweight deep learning model designed for efficient image classification. MobileNetV2 uses depthwise separable convolutions, which reduce computational cost while maintaining high accuracy. In this model, the pretrained base network was used and additional dense layers were added for classification of the **115 plant species**.

The third model implemented was **ResNet50**, a deeper CNN that uses direct connections to allow information to flow through deeper layers without gradient loss.

All models were trained using the **Adam optimizer** and **categorical cross-entropy loss function**, which are commonly used for multi-class classification problems.

Evaluation Metrics

The Accuracy of the models was evaluated using several metrics to understand their effectiveness in plant classification.

Accuracy was used as the primary evaluation metric, representing the percentage of correctly classified images.

Model loss performance during training and validation were monitored during training to observe the learning behavior of the models.

Additionally, model comparison was performed to determine which architecture provided the best classification performance on the medicinal plant dataset.

Results and Analysis

Model Performance Comparison

- After training the models, their performance was compared using the accuracy obtained on the validation/test dataset.

Model	Accuracy
CNN	80.57%
MobileNetV2	98.76%
ResNet50	70.84%

 ISBN: 978-81-685854-0-9

- Among the three models, **MobileNetV2 delivered the highest accuracy of 98.76%**, demonstrating excellent performance in classifying medicinal plant leaf images.
- The basic CNN model achieved moderate accuracy of **80.57%**, indicating that while it can learn useful patterns, it lacks the advanced feature extraction capabilities of pretrained architectures.
- The **ResNet50 model achieved 70.84% accuracy**, which was lower than expected. This may be complexity of the model and the limited number of training epochs used in the experiment.

Classification Insights

- The results show that deep learning models can successfully distinguish between medicinal plant species based on their leaf characteristics.
- The models were able to capture features such as **leaf shape, vein arrangement, and surface texture**, which are important factors in plant identification.
- Most plant species were correctly classified, although some misclassifications occurred between species that share very similar leaf structures.

Overfitting Analysis

Overfitting occurs when a model performs well on train data but poorly on unseen data. In this study, the validation accuracy for the **MobileNetV2 model Maintained consistently high**, indicating that the model was able to generalize well to unseen images. The use of data augmentation and transfer learning helped improve the robustness of the model.

User Interface Implementation

To make the trained model accessible to users, an easy-to-use web interface was developed using **Streamlit**.

The interface allows users to upload the image of a plant leaf and receive an automatic prediction of the plant species.

- Key features of the interface include:
- Image upload functionality for leaf images
- Automatic plant species prediction
- Display of prediction confidence score
- Top-3 predicted plant species
- Probability visualization using a bar chart

The interface was designed to be simple and user-friendly so that even non-technical users can easily handle the system.

Conclusion and Future Scope

- This research demonstrates the effectiveness of deep learning models for medicinal plant identifying using leaf images.
- Among the models tested, **MobileNetV2 achieved the highest accuracy of 98.76%**, showing that transfer learning significantly improves classification performance compared to traditional CNN architectures.
- The developed system can assist researchers, students, and herbal practitioners in identifying medicinal plants quickly and accurately.
- In future work, the system can be improved by:
- Training on larger datasets with more plant species
- Including additional plant features such as flowers and stems
- Deploying the system as a mobile application for field usage

- Applying explainable AI techniques to visualize important leaf features used by the model
- Such improvements could further enhance the accuracy and usability of automated medicinal plant identification systems.

References

1. Mulugeta, A. K., et al. (2023). *Deep learning for medicinal plant species classification and recognition: A systematic review.* Frontiers in Plant Science.
2. Pushpa, B. R., et al. (2023). *DIMPSAR: Dataset for Indian Medicinal Plant Species Analysis and Recognition.* Elsevier Data in Brief.
3. Dey, B., et al. (2024). *Comparative performance of deep learning models for automated medicinal plant identification.* Heliyon.
4. Hajam, M. A., et al. (2023). *An ensemble convolutional model with transfer learning for medicinal plant identification.* MDPI Information Journal.
5. Kavitha, S., et al. (2023). *Real-time medicinal plant identification using CNN architecture.* Springer Nature.
6. Gautam, V., et al. (2025). *A lightweight deep learning method for medicinal leaf classification.* Scientific Reports (Nature).
7. Sheth, F., et al. (2025). *Herbify: An ensemble deep learning framework integrating DIMPSAR dataset for medicinal plant identification.* Plant Methods Journal.
8. Rani, N. S., et al. (2025). *Deep learning-based classification of Indian medicinal plants.* Procedia Computer Science.
9. Tran, T. P., et al. (2025). *Advancements in medicinal plant identification using deep learning.* World Scientific Journal.
10. Mehta, R., & Deshmukh, P. (2022). *Deep CNN models for medicinal plant recognition using leaf images.* Journal of Artificial Intelligence Research.
11. Prasad, K., & Nair, S. (2021). *CNN-based classification of herbal plant leaves using the Flavia dataset.* International Journal of Computational Vision and Signal Processing.
12. Wang, Y., et al. (2020). *Morphological feature learning through deep convolutional neural networks for plant leaf identification.* Computers and Electronics in Agriculture.
13. Bhattacharya, P., & Basu, A. (2020). *Comparative study of ResNet and DenseNet architectures for medicinal leaf recognition.* Machine Vision and Applications.
14. Li, F., et al. (2019). *Leaf segmentation and recognition using U-Net and CNN-based classification.* Pattern Recognition Letters.
15. Shukla, A., et al. (2019). *Deep learning-based medicinal plant classification using shape and color features.* International Journal of Image Processing.
16. Patel, R., & Joshi, H. (2018). *CNN-based herbal plant identification for Ayurvedic applications.* IEEE International Conference on Computational Intelligence.
17. Zhang, L., et al. (2018). *Hybrid CNN-Gabor filter approach for plant leaf recognition.* Expert Systems with Applications.
18. Basha, S. M., et al. (2017). *Medicinal leaf identification using handcrafted features and SVM classifier.* Journal of King Saud University – Computer and Information Sciences.
19. Rao, D. N., & Kumar, M. (2017). *Shape-based plant species recognition using contour features.* International Journal of Computer Applications.
20. Singh, R., & Verma, D. (2016). *CNN-based early experimentation for medicinal leaf classification.* International Journal of Engineering Research in Computer Science and Engineering.
21. Reddy, A., et al. (2020). *Transfer learning using VGG19 for Indian plant classification.* IEEE Access.
22. Chowdhury, P., et al. (2021). *Data augmentation and regularization in CNNs for plant recognition.* Neural Computing and Applications.

 ISBN: 978-81-685854-0-9

23. Hossain, M. A., et al. (2022). *Performance comparison of deep CNN models for Bangladeshi medicinal plant classification.* Arabian Journal for Science and Engineering.
24. Nair, V., & Thomas, J. (2023). *Mobile-based CNN application for Ayurvedic medicinal plant recognition.* Journal of Mobile Information Systems.
25. Sharma, R., & Gupta, S. (2022). *Explainable AI using Grad-CAM for medicinal leaf identification.* Artificial Intelligence in Biology.
26. Kapoor, R., et al. (2024). *Deep CNN framework for detecting rare medicinal species in India.* Journal of Applied Intelligent Systems.
27. Das, T., & Mandal, R. (2023). *CNN with attention mechanism for improved herbal leaf classification.* Pattern Recognition and Image Analysis.
28. Hu, X., et al. (2021). *Integration of deep learning and IoT for real-time plant monitoring and identification.* Sensors Journal.
29. Singh, P., & Rao, N. (2024). *Comparative analysis of CNN and Vision Transformer models for plant classification.* Computers and Electrical Engineering.
30. Kumari, A., et al. (2025). *Transfer learning and ensemble CNNs for medicinal plant identification.* Journal of Intelligent Systems and Computing.

UNDERSTANDING HOW COLLEGE STUDENTS INTERACT WITH AI TOOLS AND THEIR COGNITIVE BEHAVIOUR: AN EXPLORATORY DATA ANALYSIS USING TABLEAU

Mr. Anil Prajapati
Student, DES's NMITD, Dadar.
Email: prajapatianil2605@gmail.com

Mr. Swapnil Vichare
Student, DES's NMITD, Dadar.
Email: swapnilvichare22@gmail.com

Mrs. Lavina Mistry
Assistant Professor, MCA Department, DES's NMITD, Dadar.
Email: lavinajdhv@gmail.com

Abstract
The use of AI tools in academic tasks and activities has increased over the past few years. The increased usage of AI tools in academics raises concerns whether they help in increasing the efficiency and creativity of students or they have a negative impact on creativity and reasoning skills of students. This topic has developed interest in many researchers to study the impact of AI on students. Despite having a significant amount of research being done on the topic of 'impact of AI on cognitive behaviour of students', still a gap remains in understanding how students interact with these tools, how they use AI results, how much they trust and rely on AI tools. The study aims to understand how students interact with AI tools. The study is based on analysing primary data which was collected through a online survey. The survey was conducted using a structured questionnaire which primarily focused on questions that aims to capture the interaction behavior such as AI result acceptance behaviour, trust on AI results, confidence in decision-making without AI assistance and other related questions. The data collected were 215 responses which consisted of students from different streams and qualification levels. The data were further analysed using MS Excel and visualised using Tableau. The data consists a dominant population of science and technology students. The data shows that 58.41% students are daily or frequently users of AI tools in their academics. The results show that students(121) accept AI results only after verification while a small amount of respondents(40) accept AI results directly. The results also show that students who are frequent users of AI tools report a lower confidence level in decision making without AI assistance. The heatmap of trust vs dependency on AI results shows accumulation of majority students in 50- 75% trust on AI results with average dependency on AI results. The analysis of other independent variables show that the brainstorming or thinking pattern shows a mix result between increased and decreased. Overall, the study analyses the collected data and provides visualisation based on these data to reveal patterns.
Keywords: AI Tools, Students, Cognitive Behavior, Tableau, Efficiency, Creativity, Students Interact.

 ISBN: 978-81-685854-0-9

Introduction

Over the past few years, the usage of Artificial Intelligence tools such as ChatGPT, Gemini and other generative tools have increased in educational sector. The students use AI tools to complete their assignments, solve problems, generate solutions and ideas, etc.

These tools help students by giving fast results saving their time to surf the internet and find the information. It provides information at one place without needing to refer multiple sources. Though these technologies help to generate content rapidly, providing targeted results, saving tine, there is still concern regarding over use of these tools. The academic tasks and assignments are used to train students to think and generate ideas, solve problems, and train their reasoning skills. Within the context of academics, these tools must be used as an assistant and not a one stop solution for every task and assignment. A large number of research have been done on the topic of 'Impact of AI on critical thinking and problem solving abilities of students', and they proves that a responsible and structured use of these tools help generate better results, generate better ideas and boosts productivity, while over and unethical use of these tools might diminish the cognitive capabilities of students such as problem solving, reasoning and decision- making skills.[2][3][7][8][10]. While these studies help measure how AI can boost productivity if used effectively and negative impacts if used as a tool to avoid thinking, they lack to answer a specific question of how the students interact with these tools to complete their academic tasks. This study aims to study how students interact with these tools and associated cognitive behavior. The goal is to identify patterns of how students interact, how much they believe on the results produced by these tools, how they accept the results, what do they do after getting output form these tools, etc.

The study uses primary data of 215 respondents to analyse the patterns of how students interact with AI tools. The data records AI usage frequency, their trust on these AI tools, confidence to make decision without AI tools, thinking behaviour and output verification. The data was analysed using basic descriptive statistics and visualised using Tableau.

The study is mainly focused on studying the behavioural patterns of students interacting with AI tools to understand what factors might affect their cognitive capabilities over a long period of use. It helps to provide a baseline in behavioural research of the population using AI tools. It provides an opportunity to understand the topics of further research on behaviour of AI users.

Literature Review

The rapid advancement of artificial intelligence (AI) technologies has significantly influenced educational practices, particularly in higher education. With the increasing use of AI tools such as ChatGPT, Gemini, Copilot and other generative systems, researchers have begun to study the possible effect of these tools on students reasoning abilities.

Several studies highlight the potential of AI to support cognitive development when used in a structured manner. For instance,

research on AI-guided learning environments suggest that structured and guided use of AI tools can enhance students' problem-solving abilities and creativity by providing immediate feedback and personalized support [3]. This study also highlights that unstructured use of AI can reduce students' ability to carefully evaluate information.

Similarly, research studying AI dependence among university students shows that increased reliance on AI-generated responses is related with lower levels of critical thinking. This study also states that students who use AI tools more frequently tend to rely less on independent reasoning [7].

Some experimental studies examined the importance of critical thinking in AI- supported learning environments. A research study found that while targeted intervention while using AI tools did not increase the self-reported critical thinking, rather they reduced the students' tendency to directly accept AI generated results and information.[3].

Adding to the experimental studies, qualitative studies analysing student interactions with AI tools reveal that learners commonly use AI for tasks such as writing, idea generation, and language improvement. However, concerns regarding the reliability of AI-generated information and issues related to academic integrity and plagiarism have been consistently reported [4].

Large-scale studies on AI usage in education suggest that AI has a dual role in cognitive development. While some students report improvements in creativity and problem-solving when AI is used as a support tool, others express concerns that easy access to AI-generated answers may reduce independent thinking [8].

Further research indicates that AI can enhance learning efficiency and provide personalized assistance, but excessive reliance may lead to reduced engagement and intellectual passivity[10]. These findings highlight the importance of AI literacy and self-regulation, ensuring that AI helps in assisting while doing creative tasks rather than diminishing individuals' own ability to efficiently process information [10].

Studies focusing on creativity suggest that AI tools can support idea generation and divergent thinking, particularly during brainstorming activities. However, overdependence on AI may reduce creative confidence and limit the ability to generate original ideas independently [6].

Research on collaborative learning environments shows that AI-assisted tools can enhance group problem-solving and creativity. Students using AI in collaborative tasks have shown improved outcomes as compared to traditional methods but concerns regarding cognitive independence still remain [4].

Survey-based studies examining student perceptions of AI in education indicate that students generally view AI tools as helpful for analysing information and building arguments. However, a significant proportion of respondents also believe that overdependence on AI may reduce their ability to think independently [2].

More recent studies highlight the role of self-regulated learning in establishing a healthy interaction with AI tools and technologies. While AI tools may enhance learning efficiency, their effectiveness
depends on how students regulate their usage and engage with the technology [10].

Various research studies have been conducted on AI interaction in academic activities, but still a gap remains to understand how students interact with these tools and technologies Many existing studies focus on experimental or controlled learning environments, while fewer studies examine students' real-world usage patterns and behaviours. There is limited research in understanding the students' interaction behaviour with the AI tools.

Therefore, this study focuses on describing patterns of AI usage and associated cognitive behaviours among students using real-world survey data. By examining variables such as usage frequency, dependency, verification behaviour, and confidence, the aim of the stud is to contribute to a better understanding of how students interact with AI tools in practical contexts.

Research Methodology

The research is based on the primary data, collected using a structured questionnaire. The responses were collected using an online survey and a total of 215 responses were gathered. The sampling method used was convenience sampling, with the sample population selection based on accessibility and willingness to participate in the research. The questionnaire consisted of behavioural questions such as,

- AI usage frequency
- Trust on AI results
- Confidence in decision making without AI assistance
- Output verification behaviour
- How students accept the output (directly accept or verify)

 ISBN: 978-81-685854-0-9

The data was analysed using statistical approaches such as frequency distribution, percentage calculation, categorization and this analysis were further supported through visualisations. The visualisations show cross tabulation between two or more variables showing how variables are associated with each other.
The study uses grouping variables such as AI usage frequency, trust in AI results, dependence on AI tools to group data into categories for comparative analysis.

Results and Research Findings

- **Distribution of Respondents across Stream of Study**

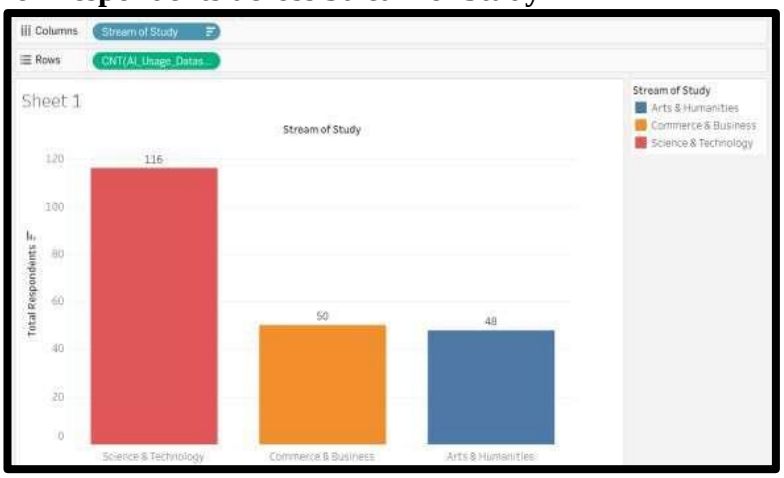


Figure 1 Stream of Study

Figure 1 presents the distribution of respondents across academic disciplines. A large number of students belong to the Science & Technology stream (116 respondents), followed by Commerce & Business (50 respondents) and Arts & Humanities (48 respondents). It shows that the dataset is dominated by students from technical backgrounds with a significant amount of non-technical students.

- **Education Qualification of Respondents**



Figure 2 Educational Qualification

Figure 2 shows the categories of respondents according to their educational qualification. The majority of participants are undergraduate students (121 respondents), followed by postgraduate students (65 respondents) and junior college students (29 respondents).

This distribution indicates that the sample is predominantly composed of undergraduate students, with comparatively fewer participants from postgraduate and junior college levels.

- **AI Usage Frequency**

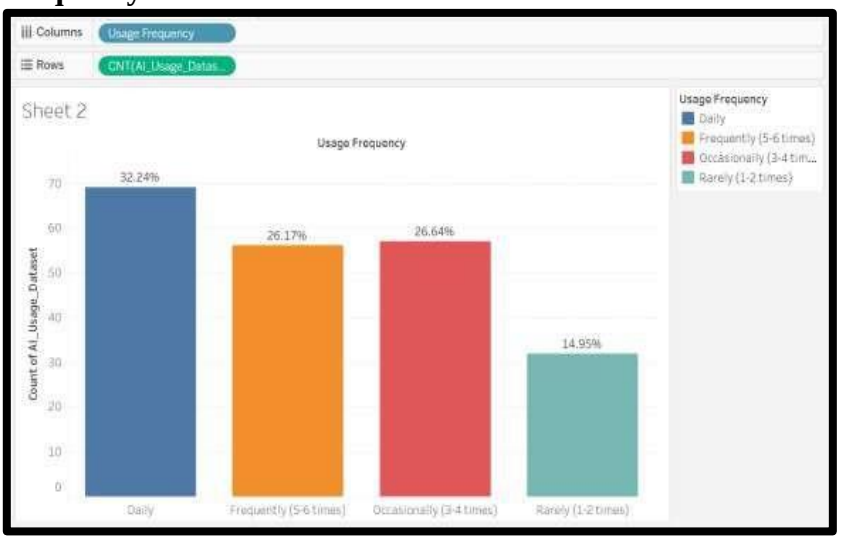


Figure 3 AI Usage Frequency

Figure 3 shows the distribution of AI usage frequency among respondents. The largest group consists of students who use AI daily (32.24%). This is followed by occasional users (3 –4 times) at 26.64% and frequent users (5–6 times) at 26.17%. A smaller proportion, 14.95% of respondents, reported using AI rarely. These values indicate that a substantial proportion of students engage with AI tools regularly, with daily usage being the most prominent category

- **Dependency on AI Tools**

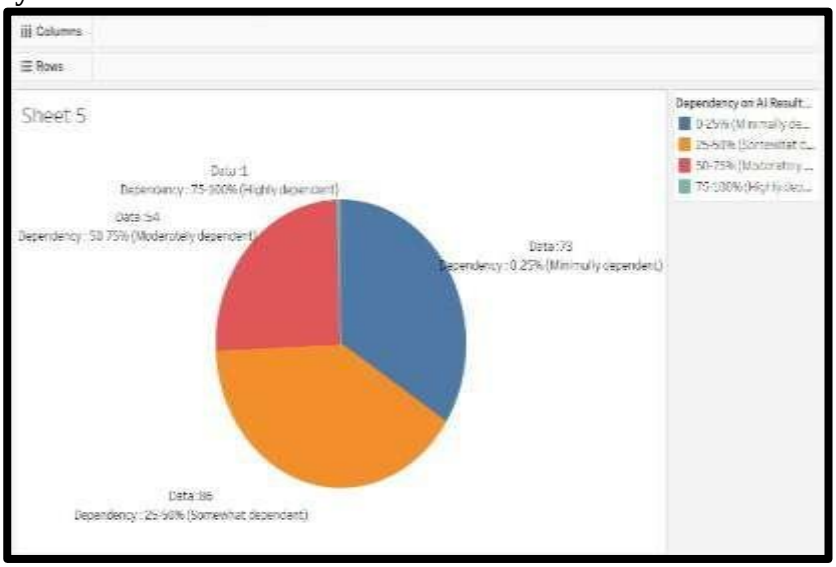


Figure 4 Dependency on AI Tools

Figure 4 illustrates the levels of dependency on AI tools. The majority of respondents fall within the 25–50% dependency range (86 respondents), indicating moderate reliance. This is followed by 0–25% dependency (73 respondents) and 50–75%
dependency (54 respondents). Only 2 respondents fall into the 75–100% dependency category, indicating that very high dependency is rare. Overall, the distribution shows that most students maintain partial reliance on AI rather than complete dependence.

 ISBN: 978-81-685854-0-9

- **Trust on AI Results**

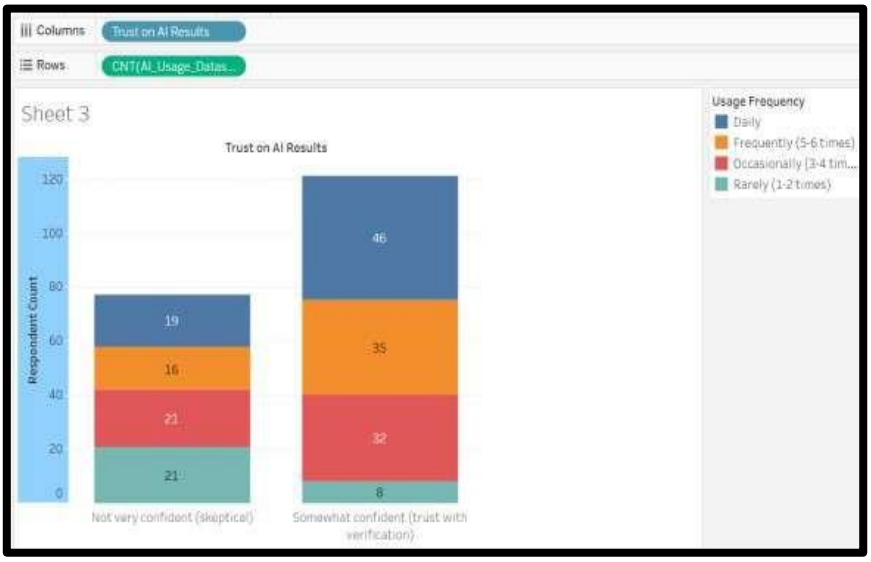


Figure 5 Trust on AI Results

Figure 5 presents the distribution of trust in AI-generated outputs. A larger number of respondents fall into the category of "trust with verification" (121 respondents) as compared to those who are unsure and doubtful(sceptical) (approximately 77 respondents). Within the trust-with- verification group, a higher number of respondents are frequent or daily users.

- **Dependency on AI Results across Usage Levels**

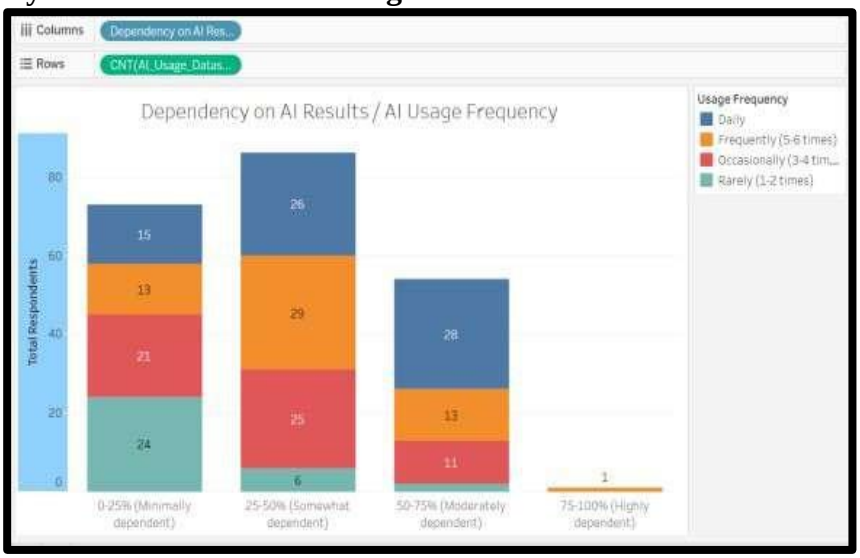


Figure 6 Dependency on AI Results across Usage Levels

Figure 6 shows the relationship between AI usage frequency and dependency levels. In the 25–50% dependency category, the largest contributions come from frequent users (29 respondents) and daily users (26 respondents). In the 0–25% dependency category, the respondents are categorized across all usage levels, with slightly higher counts among occasional and rare users. In the 50–75% dependency category, occasional users (28 respondents) form the largest group. The 75–100% dependency category contains only 1 respondent, indicating that very high dependency is minimal across all usage groups. This distribution shows that moderate dependency is more common among higher-frequency users.

ISBN: 978-81-685854-0-9

- **Trust on AI Results vs Dependency on AI Results**



Figure 7 Trust on AI Results vs Dependency on AI

Figure 7 presents the relationship between trust levels and dependency. The highest concentration of respondents (54 respondents) falls within the 50–75% trust and 25–50% dependency category, indicating moderate trust combined with moderate dependency. In the 0–25% dependency group, the majority fall within the 25–50% trust range (39 respondents). Only a few respondents are observed in the highest trust (75–100%) and highest dependency (75–100%) categories. This indicates that extreme trust and dependency are uncommon, with most students maintaining a balanced level of trust.

- **AI Usage and Confidence without AI**



Figure 8 Confidence in Decision Making without AI assistance

Figure 8 shows the relationship between AI usage frequency and confidence in solving problems without AI assistance. Among daily users, the largest group reports being "not very confident" (40 respondents), followed by "somewhat confident" (17 respondents). A similar pattern is observed among frequent users, where 31 respondents report being "not very confident." Within occasional users, the responses show even distribution, with 22 respondents reporting "somewhat confident" and 27 reporting "not very confident." In contrast, among rare

 ISBN: 978-81-685854-0-9

users, the majority report being "somewhat confident" (18 respondents). This distribution shows that lower confidence is more frequently observed among higher- frequency AI users.

- **AI Usage vs Brainstorming Behaviour**



Figure 9 Brainstorming Behavior across various Usage Levels

Figure 9 shows the relationship between AI usage frequency and brainstorming behaviour. Among daily users, the responses can be seen distributed across all categories, including decreased (21 respondents), slightly increased (11 respondents), significantly increased (18 respondents), and no change (17 respondents). A similar trend can be observed among frequent and occasional users, with no single dominant category. Among rare users, the largest group reports no change (20 respondents). This indicates that the effect of AI on brainstorming behaviour varies across users and does not follow a consistent trend.

- **Perceived Reduction in Thinking across different frequencies of AI Usage**



Figure 10 Perceived Reduction in Thinking Time

Figure 10 demonstrates perceived reduction in thinking across usage levels. Among daily users, the most common response is "sometimes" (30 respondents), followed by "often" (28 respondents). Among frequent users, "sometimes" (30 respondents) is also the most common

response. Among occasional users, "sometimes" (26 respondents) remains dominant, and among rare users, "sometimes" (11 respondents) is again the most frequent. The consistent pattern of reduction in thinking across all usage levels indicates that a moderate perception of reduced thinking is commonly reported.

- **Interaction with AI generated outputs, Ideas and Suggestions**

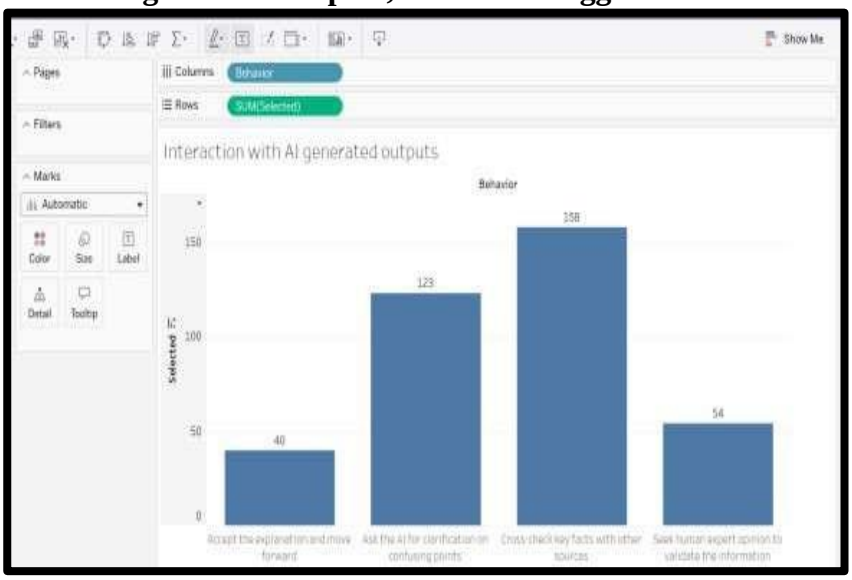


Figure 11 Interaction with AI generated content

Figure 11 shows how students respond to AI-generated outputs. A large number of students prefer to cross-check facts with other sources (158 respondents), while many students also ask the AI for clarification (123 respondents). Very few students rely on human experts for validation (54 respondents), and only 40 respondents report to accept the explanation without further verification.

The responses collected through the questionnaire consisted of some multiple- choice questions. The responses to these questions need to be separated and analysed separately. Hence, we can see that the total sum of all the responses across all categories is higher than total number of responses (215).

The results suggest that students generally do not accept AI generated outputs and results directly and do further verification for such AI generated content.

Discussion

The results provides an overview of how students use AI tools and technologies and their interaction behaviour. The findings show that a significant number of respondents use AI tools regularly, while most of the students report a moderate level of reliance on AI, while only a few responses show very high dependency.

There is a similar pattern between trust on AI results and dependency on AI tools. A significant portion of respondent(121) fall in the category of moderate trust with verification before accepting and using the results. This suggests that checking information is a common behaviour among users. The respondents' interaction behaviour with the AI tools and technologies shows that students do not directly accept the AI generated results and check for clarification before accepting and using the results

The analysis of AI usage frequency with confidence in decision making without AI support shows that daily and frequent users often report lower confidence compared to those who use AI occasionally or rarely. This observation shows a pattern in the data and does not establish a clear relationship between AI usage and confidence levels. After analysing the graphs

 ISBN: 978-81-685854-0-9

generated for user perceived brainstorming impact, it was observed that there is no single category which is dominant and the results are mix of all the categories indicating thinking and idea generation pattern does not follow a particular strict trend or pattern.
The user perception of user reduced thinking while using AI tools show a significant rise in responses saying "sometimes" which shows that users do not see a significant or consistent change in their thinking process or thinking duration.
The results show patterns of how user behave and think while interacting with AI tools. The patterns are strictly dependent on the sample data and cannot be generalized further to a larger population.

Conclusion

The study focuses on understanding how students interact with AI tools such as ChatGPT, Gemini and other generative tools by generating visualisation to show patterns and trends in the data. This study is mainly focused on analysing the data rather than proving causes of AI on cognitive behaviour. A significant portion (121 respondents) show moderate trust on these AI tools. The patterns indicate there is no linear relation between AI usage and perceived cognitive change and the patterns are dependent on various interaction behaviours.
The descriptive nature of the study and the sampling method (convenience sampling) restrict these patterns to the considered sample size and may not be generalized further to a larger population.

Limitations of the study

The study has some limitation which should be considered while stating the results and findings. The data is collected through convenience sampling and further generalisation of research to a larger population of students.
The data collected is through online questionnaire which may include bias if the respondents did misunderstood any question. The research relies on perceived behaviour of student rather than observed patterns through experimenting with focus groups.
The research adopts a descriptive nature and a cross sectional design which helps in pattern generation but cannot establish causal relationship with independent and dependent variables of data.

References

1. Aprianto, I.(2025).Artificial intelligence and its effects on critical thinking and problem-solving in higher education. International Journal of Educational Development.
2. Gerlich, M. (2025). AI tools in society: Cognitive offloading and critical thinking. Sociological Perspectives, 15(1), 1–17.Hwang, G. J. (2025). Facilitating students' critical thinking with generative AI-guided learning. Computers & Education, 215, 104563.
3. Hwang, G. J. (2025). Facilitating students' critical thinking with generative AI-guided learning. Computers & Education, 215, 104563.
4. Jin, Y., Wang, H., & Li, Q. (2024).
Generative AI in higher education: A global perspective. Computers and Education: Artificial Intelligence.
5. Lee, H. P., Larson, B., & Chen, M. (2025). The impact of generative AI on critical thinking in higher education. CHI 2025: CHI Conference on Human Factors in Computing Systems
6. Salido, A. (2025). Integrating critical thinking and artificial intelligence in higher education: A systematic review. Education and Information Technologies.
7. Tian, J., Zhang, Y., & Liu, X. (2025). Learners' AI dependence and critical thinking: Evidence from higher education. Computers & Education.

8. Vieriu, A. M., Popescu, C., & Ionescu, M. (2025). The impact of artificial intelligence on students' learning and engagement. Education Sciences, 15(3), 343.
9. Wang, H., Li, Q., & Zhang, Y. (2024). Generative AI chatbots in higher education: Opportunities and challenges. Computers and Education: Artificial Intelligence.
10. Zhou, Y. (2024). The mediating role of generative AI self-regulation on students' critical thinking and problem-solving. Educational Technology Research and Development.

 ISBN: 978-81-685854-0-9

EVOLUTION OF USER SENTIMENT AND LINGUISTIC BARRIERS IN CONVERSATIONAL AI: AN NLP PERSPECTIVE

Ritika Sahu
MCA Students, Department of Computer Applications, Mumbai University, India.
Email: sahuritika255@gmail.com

Sakshi Shinde
MCA Students, Department of Computer Applications, Mumbai University, India.
Email: sakshishinde757@gmail.com

Abstract
In recent years, chatbots have evolved from experimental scripts into a cornerstone of Artificial Intelligence applications, fundamentally altering the landscape of digital communication. These sophisticated systems are engineered to facilitate seamless interactions between humans and machines, utilizing either text-based or vocal interfaces to deliver automated, real-time responses. Natural Language Processing (NLP) serves as the critical engine behind these developments, providing the linguistic framework necessary for machines to decipher the complexities of human syntax, slang, and intent. As these systems become ubiquitous in customer support, education, financial services, and healthcare, the demand for high-quality, human-like interaction has never been greater. This study focuses on a multi-dimensional analysis of user awareness, behavioral engagement, and the underlying levels of trust associated with NLP-driven chatbots. Drawing from primary data collected via structured surveys, this research examines the friction points between current technological capabilities and user expectations. The findings suggest that while the accessibility of chatbots is at an all-time high, significant improvements in contextual awareness and emotional intelligence are required to ensure long-term sustainability and user satisfaction.
Keywords: Chatbot, Natural Language Processing, Conversational AI, User Trust, Automation, Human-Computer Interaction, Digital Literacy, Emotional Intelligence.

1. Introduction

The rapid acceleration of the global digital economy has necessitated a shift in how service providers and consumers engage. In an era defined by "on-demand" services, traditional manual support systems—often hampered by limited hours and high wait times—are no longer sufficient to meet consumer needs. Consequently, chatbots have emerged as a vital technological solution, capable of simulating human-like conversations to provide immediate assistance across various digital ecosystems, including corporate websites, mobile applications, and social media platforms. Their ability to provide round-the-clock service makes them beneficial for both businesses and users alike.

At the heart of these intelligent systems lies Natural Language Processing (NLP). This branch of AI is dedicated to the interaction between computers and human languages, enabling a machine to not only "read" text but to extract meaning from it. Through a series of complex stages—including tokenization, stemming, entity recognition, and machine learning—NLP transforms unstructured human speech into structured data that a computer can process. However, as the deployment of these systems expands, the focus has shifted from mere functionality to the quality of the "human" experience. Users today expect more than just a pre-programmed response; they seek a dialogue that feels natural, intuitive, and secure. This

research investigates the current state of these interactions, highlighting the gap between technological advancement and the psychological comfort of the end-user.

The importance of this study lies in the increasing role of chatbots in modern society. Although chatbots offer convenience and efficiency, users still face significant problems such as incorrect answers, lack of context, and low trust in sensitive matters. Therefore, this research examines user behavior, satisfaction, limitations, and future expectations regarding chatbot systems using NLP.

The growing dependence on chatbot technology also reflects a broader change in user expectations toward faster and more personalized digital communication. Modern users prefer systems that can understand queries instantly, provide accurate solutions, and reduce the effort required to access information. At the same time, repeated interaction with poorly designed chatbots can create dissatisfaction and reduce trust in automated systems. For this reason, it becomes important to evaluate not only the technical performance of NLP chatbots but also the emotional, behavioral, and practical experiences of users. Understanding these aspects can help developers design smarter, safer, and more user-friendly conversational systems for the future.

2. Operational Dynamics and Organization Benefits

The widespread adoption of chatbot technology is largely driven by its ability to optimize operational workflows and enhance service speed. Organizations across the globe are increasingly integrating these systems to manage high volumes of routine inquiries, ranging from tracking shipping orders to resolving common technical glitches. By automating these low-level tasks, businesses can reallocate their human resources toward more complex, high-value problem-solving, thereby increasing overall productivity and improving operational efficiency.

In the academic and professional development sectors, chatbots serve as a vital resource for "just-in-time" learning. Educational platforms utilize NLP to create interactive tutors that can answer student queries instantly, providing a personalized learning experience that is available 24/7. Similarly, in the banking and healthcare sectors, chatbots are being used for appointment scheduling, FAQ handling, and basic support functions. The adoption of chatbot systems creates opportunities for better user engagement by allowing businesses to interact with multiple customers simultaneously.

However, the success of these implementations is heavily dependent on the robustness of the underlying NLP architecture. A chatbot is only as effective as the data it has been trained on. If the training sets are limited or the algorithms lack the ability to handle linguistic variations, the resulting service can feel rigid and unhelpful. Poor training data or weak algorithms may lead to incorrect responses, making it essential for organizations to regularly update chatbot systems and train them with relevant, diverse datasets to maintain performance and reliability.

Another important advantage of chatbot adoption is consistency in service delivery. Human support quality may vary depending on workload, timing, or individual performance, whereas chatbots can provide standardized responses to common queries at any hour. This helps organizations maintain service quality across different platforms and time zones. In addition, chatbot interactions generate useful data regarding customer preferences, common problems, and frequently asked questions. Such insights can support better decision-making, improve products or services, and help organizations plan future strategies more effectively.

3. Critical Barriers in Human-Machine Dialogue

Despite the impressive growth of conversational AI, several persistent challenges continue to hinder the user experience. One of the most significant barriers is the frequent misunderstanding of "user intent." Human language is naturally ambiguous, filled with nuances, metaphors, and cultural idioms. When a user presents a complex or multi-part query,

 ISBN: 978-81-685854-0-9

many current NLP models struggle to parse the information correctly, often resulting in a "failure loop" where the user is forced to repeat or rephrase their request multiple times. This not only diminishes the convenience of the tool but can lead to significant user frustration and a total abandonment of the service.

Another critical limitation is the absence of genuine emotional intelligence. Human communication is not just about the exchange of information; it is influenced by tone, feelings, and context. Current AI systems are largely "affect-blind," meaning they cannot detect if a user is angry, confused, or in an emergency. In sensitive fields like healthcare or crisis counseling, a robotic or clinical response can be perceived as insensitive, potentially causing more harm than good. Bridging this "empathy gap" remains one of the most difficult hurdles in the field of AI development.

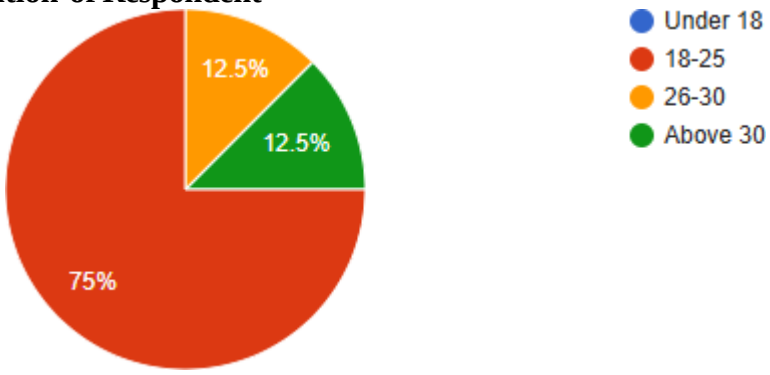
Furthermore, privacy and data security remain at the forefront of user concerns. As chatbots often require personal or financial information to complete tasks, users are becoming increasingly wary of how their data is stored and utilized. The fear of data breaches or the unethical use of conversation logs can severely undermine the trust required for these systems to be fully integrated into daily life. Organizations must ensure secure data handling, transparency, and ethical use of AI technologies to increase user confidence.

3.1 Behavioral Observations and Demographic Trends

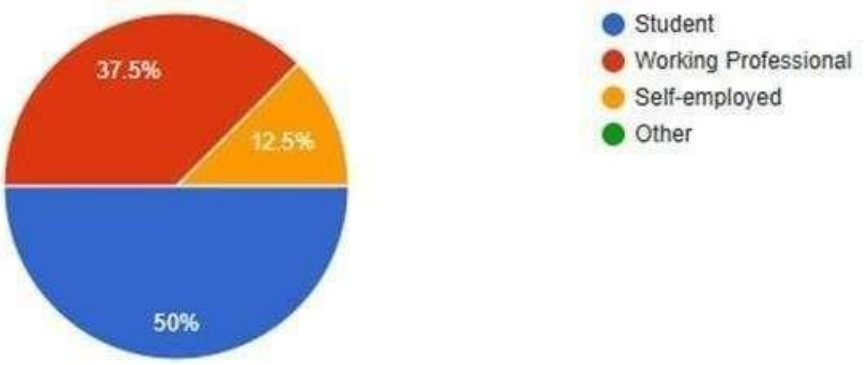
Our analysis indicates a clear correlation between digital literacy and the adoption of chatbot technology. Younger users, particularly students and professionals, exhibit a higher degree of comfort with automated interfaces, often using them regularly for information search, learning assistance, and customer support. This demographic appreciates chatbots for saving time and providing instant solutions.

However, satisfaction levels remain inconsistent. While the convenience of instant responses is valued, many respondents noted that the quality of these responses often feels superficial. Dissatisfaction is frequently linked to repetitive or irrelevant replies, suggesting that chatbot systems must improve response quality to meet user expectations. Overall, user feedback reflects a positive acceptance of the technology but indicates a desperate need for continuous enhancement in language understanding and interaction quality.

1. Age Distribution of Respondent



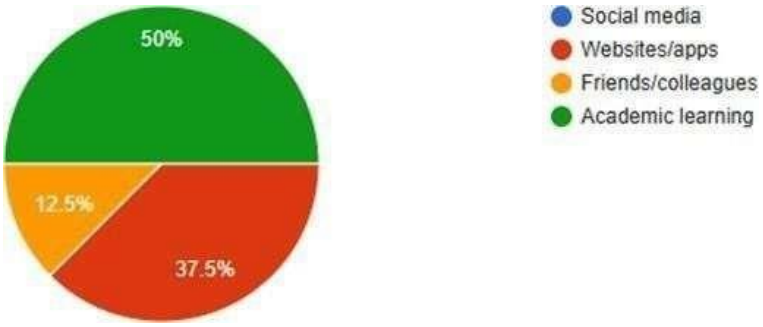
2. Occupation of Respondent



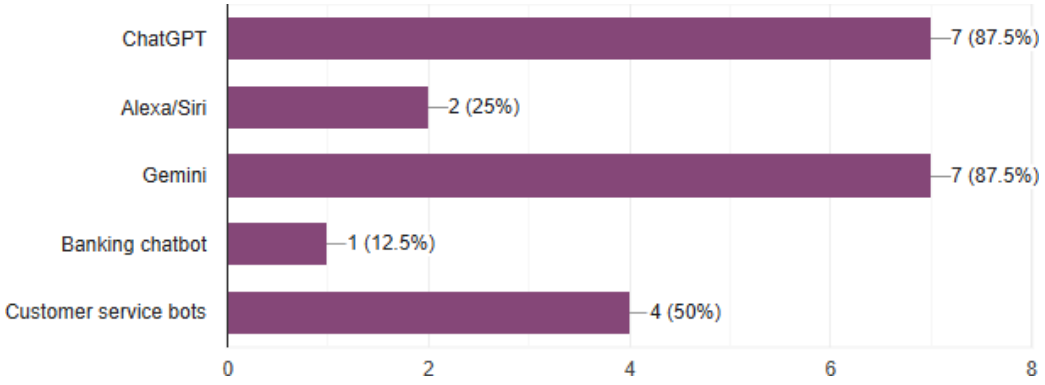
3. Awareness of Chatbots



4. Source of Awareness

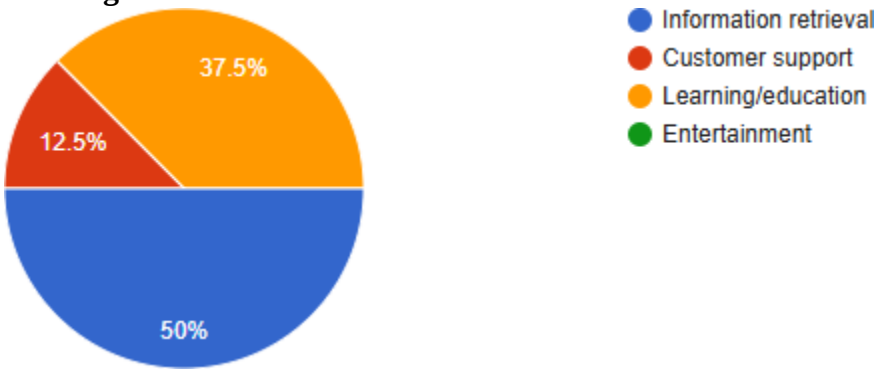


5. Chatbot Platforms Used

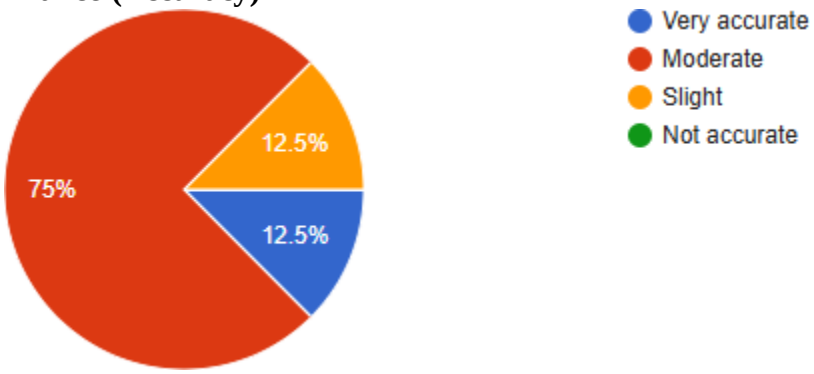


 ISBN: 978-81-685854-0-9

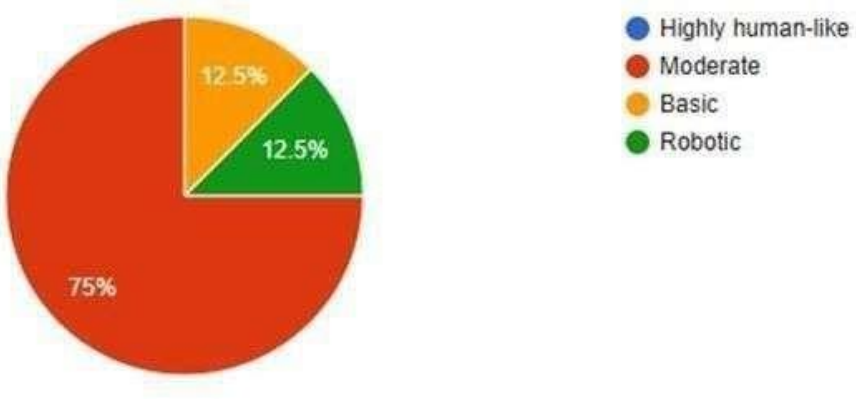
6. Purpose of Using Chatbots



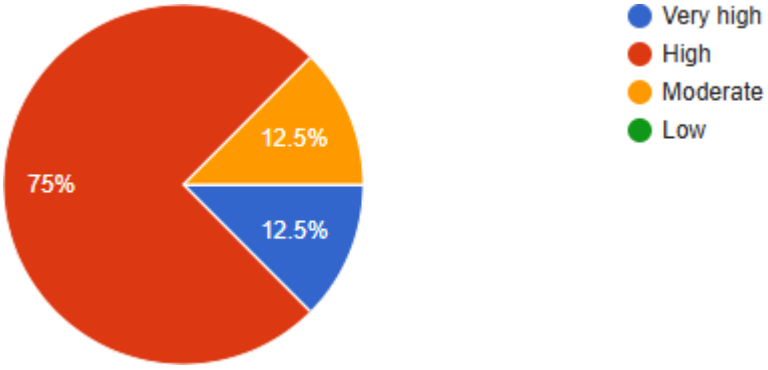
7. NLP Performance (Accuracy)



8. Human like Responses



9. Future Scopes



ISBN: 978-81-685854-0-9

3.2 Strategic Future Directions for Conversational AI

Looking forward, the evolution of chatbots will likely be defined by a shift toward more sophisticated, multi-turn dialogue management. Instead of treating every message as a standalone event, future NLP models will utilize "contextual memory," allowing the system to remember previous interactions and build a coherent, ongoing conversation. Better NLP models can reduce mistakes and make conversations smoother, while context memory helps maintain natural discussions.

Additionally, the development of multilingual and cross-cultural NLP models is essential for global accessibility. Supporting multiple languages can make chatbot services accessible to wider populations, which is particularly useful in countries with linguistic diversity. As these systems integrate with smart devices, voice assistants, and analytics systems, their use in homes, businesses, and public services will expand, demonstrating the broad future potential of the technology.

Another promising direction for conversational AI is the integration of advanced personalization features. Future chatbot systems may learn user preferences, communication styles, and frequently requested services to deliver more relevant and efficient responses. This personalized interaction can improve user satisfaction and make digital assistance more engaging. At the same time, responsible development will require strong ethical guidelines, bias reduction, and transparent decision-making processes so that chatbot technology remains fair, secure, and beneficial for users across different sectors.

Another significant future trend is the integration of conversational AI with emerging technologies such as augmented reality, Internet of Things devices, and predictive analytics. Chatbots may soon assist users through smart homes, wearable devices, and interactive learning platforms with greater efficiency. These developments can improve convenience, automate routine tasks, and create seamless communication between humans and intelligent systems. As innovation continues, conversational AI is expected to become an essential part of everyday digital life.

4. Conclusion

This research demonstrates that while Natural Language Processing has provided a solid foundation for the rise of chatbots, the journey toward truly human-centric AI is still ongoing. Chatbots have undeniably proven their value in terms of speed, availability, and cost-efficiency for organizations. However, for these systems to reach their full potential and gain universal user trust, they must move beyond basic information retrieval. The study identifies that misunderstanding of queries, weak context handling, and privacy concerns remain major issues that reduce user satisfaction.

The future of the field lies in the refinement of language understanding, the integration of emotional awareness, and a steadfast commitment to data privacy. As developers continue to improve the nuance and adaptability of NLP models, chatbots will transition from being mere "tools" to becoming reliable digital companions. Future chatbots are likely to become more accurate, adaptive, and human-like, playing an even larger role in business, education, healthcare, and digital services in the coming years. The ultimate goal is to create a digital interface that feels like a natural extension of human communication.

In conclusion, the study indicates that chatbot technology is progressing rapidly and becoming more accepted in everyday digital communication. Although present systems still face challenges in understanding emotions, maintaining context, and handling complex queries, their benefits continue to encourage wider adoption. With continuous innovation and responsible implementation, NLP-based chatbots have the potential to transform how people interact with technology by creating faster, smarter, and more accessible communication experiences.

 ISBN: 978-81-685854-0-9

5. References

1. Goor V., Yadav N. S., Yadav P. S., "Conversational AI for Natural Language Processing: An Review of ChatGPT", International Journal on Recent and Innovation Trends in Computing and Communication, Vol. 11, Issue 3, 2023.
2. Regin D. R., Rajest D. S. S., Shynu T., Jerusha Angelene Christabel G., Steffi R., "An Automated Conversation System Using Natural Language Processing (NLP) Chatbot in Python", Central Asian Journal of Medical and Natural Science, Vol. 3, Issue 4, 2022, pp. 314-336.
3. Aleedy M., Shaiba H., Bezbradica M., "Generating and Analyzing Chatbot Responses Using Natural Language Processing", International Journal of Advanced Computer Science and Applications, Vol. 10, Issue 9, 2019.
4. Lalwani T., Bhalotia S., Pal A., Rathod V., Bisen S., "Implementation of a Chatbot System using AI and NLP", International Journal of Innovative Research in Computer Science & Technology (IJIRCST), Vol. 6, Issue 3, 2018.
5. Rahman A. M., Mamun A. A., Islam A., "Programming Challenges of Chatbot: Current and Future Prospective", 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), IEEE, 2017.
6. Abdulla H., Altahir A. M., Alwahishi S., Saghair K., Platos J., Snasel V., "Chatbots Development Using Natural Language Processing: A Review", 2022 26th International Conference, IEEE, 2022.

CYBER-BULLYING: AI-BASED CYBERBULLYING DETECTION SYSTEM FOR SOCIAL MEDIA PLATFORMS

Sakshi Lahane
MSc (Data Science)

Abstract

"The internet is a weapon. It can be used to protect or to destroy."
— Unknown

Cyberbullying keeps showing up as a major issue across social media sites these days. It ends up hurting peoples mental health and their sense of well being in real ways. This study puts together and tests out an AI system for spotting cyberbullying. The setup relies on a hybrid approach. It brings together word embeddings and convolutional neural networks, or CNNs. It handles text from social media datasets that are out there for anyone to use. The process starts with breaking down the text into tokens and getting rid of common stop words. Then it turns things into semantic vectors to capture the real meaning in context. The CNN gets trained on classifying posts from users as either bullying or not bullying at all. It comes out with better precision and recall plus F1 scores than the usual machine learning methods like Support Vector Machines or Random Forests. The results from experiments make it clear that these deep ways of handling context help pick up on the trickier kinds of cyberbullying. That includes things like sarcasm or language that is coded to hide the intent. The work also looks into problems like uneven datasets and unclear meanings in context. It covers ethical stuff too such as biases built into algorithms and keeping user privacy safe. In the end the mixed method shows how deep learning can push forward with automatic checks on content. This kind of thing really builds online areas where folks can communicate without as much worry in the end.

Keywords: Cyberbullying, AI System, Convolutional Neural Networks, Social Media.

Objective

The main objective of the study is to design, develop, and evaluate an AI-driven Cyberbullying Detection System capable of automatically detecting and classifying offensive, abusive, or harmful textual content on social media. The system aims to provide better detection accuracy, interpretability, and ethical reliability by availing hybrid deep learning techniques. In pursuit of this general objective, the study shall pursue the following specific objectives:

i. Development of a Hybrid Deep Learning Model

A strong hybrid deep learning model will be developed, combining semantic word embedding techniques, such as Word2Vec, GloVe, and FastText, with Convolutional Neural Network architectures. The hybrid model will aim to leverage the strength of the embeddings regarding word semantics representation and the hierarchical feature extraction capability of CNNs with regard to detecting subtle linguistic and contextual cues associated with cyberbullying. This is likely to enable it to generalize effectively to a wide variety of social media text data.

ii. Efficient Design of a Data Preprocessing Pipeline

This involves the design and implementation of an efficient and broad data preprocessing pipeline, considering the noisy, informal, and highly variable nature of social media text. This pipeline will involve tokenization, stop-word removal, text normalization, and handling of non-standard linguistic features such as emojis, slang, abbreviations, and misspellings. The pre-processing step will guarantee uniformity in the data and improve the quality of the textual input presented to the hybrid model, hence improving the overall performance of the system.

 ISBN: 978-81-685854-0-9

iii. Performance Assessment Using Benchmark Datasets

The performance of the proposed hybrid model is to be tested using well-established benchmark datasets such as the Kaggle Cyberbullying dataset, Wikipedia Detox dataset, and Twitter-based cyberbullying corpora. Model performance will be evaluated in terms of standard quantitative metrics: Accuracy, Precision, Recall, F1-score, and ROC-AUC, to find the predictive reliability and robustness of the model across diverse data sources and bullying contexts.

iv. Comparative Analysis with Baseline Models

The proposed hybrid model will be compared to the baseline models: traditional machine learning algorithms like Support Vector Machine, Naïve Bayes, and Random Forest; and standalone deep learning models like LSTM, Bi-LSTM, and a CNN-only architecture. This comparative study will serve to highlight the relative effectiveness, efficiency, and adaptability of the hybrid approach in the detection of cyberbullying-related content across multiple data domains.

v. Handling Imbalanced Classes and Scarce Data

To reduce the problem of class imbalance and scarcity of data in cyberbullying datasets, resampling techniques such as SMOTE, undersampling, and, when possible, augmentation are considered. This objective gives much better recall and precision for minority classes, that is, bullying instances, hence increasing the sensitivity of the model toward detecting harmful or abusive content that may otherwise go undetected.

vi. Investigation of Embedding Choice and Architectural Optimization

By doing so, the aim is to systematically investigate the influence of embedding selection and model architecture tuning (e.g., number of filters, kernel size, activation functions, dropout rates, and learning rates) on the detection performance. This paper also aims at finding optimal configurations that balance computational efficiency with model accuracy, hence providing deeper insights into how architectural and representational design choices impact cyberbullying detection effectiveness.

vii. Ensuring Ethical Reliability and Fairness

Ensuring the developed system operates ethically reliably and without bias while performance in detection is improved, mainly by reducing prediction bias and false alarms, especially those concerning the misclassification of non-offensive content as harmful. This objective underscores the importance of principles of fairness and transparency in the design of AI models that can enable a process of content moderation equitably and responsibly, reflecting ethical requirements and user rights.

Introduction

In the digital era, social interactions have taken a dramatic change. The proliferation of various social networking sites, online communities for gaming, and instant messaging has really transformed the way people interact, communicate, and share information globally. While these have given immense benefits in facilitating creativity, collaboration, and the free flow of ideas, they have simultaneously come to be associated with new and complex manifestations of online harms. Of these, cyberbullying has emerged as one of the most prevalent and harmful phenomena affecting users worldwide, particularly adolescents.

Cyberbullying is a serious challenge for maintaining online safety and mental health, particularly among teenagers and young adults. The fast growth of social media, along with factors like anonymity, peer pressure, and constant digital connectivity, have contributed to an increase in the issue's scope and severity. In a 2024 survey conducted by the Cyberbullying Research Center, approximately 55% of students aged 13-17 years in the United States reported that they had experienced cyberbullying, while 27% reported having experienced it in the past 30 days alone (Hinduja & Patchin, 2024). Similarly, the WHO reported that one in six

adolescents across 44 European countries had been victims of online bullying in the past year (WHO/Europe, 2024).

These numbers significantly exceed those from 2018, showing that the problem is not only ongoing but also growing in scale.

Similar trends have been seen globally. Research conducted in countries like Argentina, Panama, and St. Vincent and the Grenadines showed that about 20% of adolescents reported having experienced cyberbullying. Victims had 1.88 times greater odds of suicidal ideation compared to their non-bullied peers (95% CI: 1.77–1.98).

The Pew Research Center found that nearly half of U.S. teens had encountered some kind of online harassment, including name-calling, rumor-spreading, and intentional exclusion (Anderson & Jiang, 2022). Collectively, these studies suggest that cyberbullying is a global and transnational issue, with differences across geographies and cultures, and the anonymity and ease of use of digital communication enable it to grow at a much faster rate than ever possible.

Cyberbullying has extensive psychological, social, and academic consequences. The likely impacts on targets include anxiety, depression, low self-esteem, social withdrawal, and even suicidal ideation (Wright et al., 2023). Continuous exposure to online hostility leads to emotional exhaustion, sleep disturbances, and academic disengagement. The harm can be much more intense and longer-lasting for adolescents, as this is their most tender age regarding identity and coping strategy development.

The effects of cyberbullying have been noted in research as being equal to, or even more damaging than, traditional face-to-face bullying (Kowalski et al., 2021). Unlike offline harassment, aggression through digital media is often permanent, public, and inescapable, having the potential to remain visible indefinitely before large audiences. The lack of temporal and spatial boundaries further heightens the victim's psychological distress, for whom even private space cannot be used to relieve themselves from the effects.

Beyond individual suffering, cyberbullying involves broader challenges in light of societal and institutional impacts. It disrupts educational environments, fosters hostility within online communities, and erodes users' trust in digital ecosystems. Long-term exposure to online aggression builds up digital fatigue and reduced social participation, which affects social cohesion and community engagement negatively (Sarmiento et al., 2023).

Given the prevalence of this problem and its persistence over a long period of time, researchers and technologists increasingly have recourse to artificial intelligence as a promising solution for automatic detection and prevention. The scale and linguistic complexity of modern online communication make traditional rule-based and manual content moderation approaches inapt. Among AI-driven methods, those that rely particularly on machine learning and deep learning have come to the forefront with promising avenues toward real-time detection of abusive language.

However, this task is considered challenging due to various factors: context dependency, sarcasm, slang, multilinguality, imbalance of data, and bias in training datasets. Recent developments in NLP and deep learning architectures have made it possible to develop systems that can grasp semantic relationships and nuances of context in text.

Among those, hybrid models using word-embedding techniques, such as Word2Vec, GloVe, and FastText, combined with CNNs, have shown great potential in extracting both semantic meaning and local textual patterns relevant for abusive or harmful language detection. Such approaches allow for more correct classification and enhanced interpretability in distinguishing between benign and harmful online interactions. In this regard, the present study proposes a hybrid AI-driven cyberbullying detection system incorporating word embedding techniques along with CNN architecture to enhance the detection performance on social media content. The system covers different stages: data collection, text preprocessing, feature extraction, and

 ISBN: 978-81-685854-0-9

deep learning-based classification. The study also covers comparative performance evaluation of the proposed hybrid model against traditional machine learning and standalone deep learning approaches using standard metrics such as precision, recall, F1-score, and ROC-AUC on benchmark datasets. It further discusses the challenges regarding data imbalance, fairness, and ethical reliability and provides practical insights toward real-world deployment of AI-based moderation systems in online environments.

Despite the rapid development of AI and NLP techniques, cyberbullying detection remains one of the persistent, challenging research problems. The dynamic, informal, and context-dependent features of social media communication make it difficult for automated systems to identify and classify harmful content.

Most existing cyberbullying detection models either rely on traditional machine learning approaches or on single deep learning architectures, including LSTM or CNN. While those models have achieved moderate success, several limitations remain, restricting their real-world applicability and scalability.

First, classic machine learning models rely on hand-crafted lexical and syntactic features, including SVM, Naïve Bayes, and Random Forest. These features will usually not capture the richness in semantics and context of human languages, especially those online, whose meanings are negotiated through slang, abbreviations, emojis, sarcasm, and cultural varieties. For this reason, these models often underperform when generalizing to new, unseen data from different domains or platforms.

While deep learning models such as CNNs, LSTMs, and Bi-LSTMs learn representations of texts in both a hierarchical and sequential manner, critical challenges still remain. Many single-model architecture models overfit small or imbalanced datasets and often do not completely capture the semantic meaning-what is said-and contextual intent-how and why it is said-of texts. Unlike overt hate speech, cyberbullying often manifests through subtle patterns that come in the form of irony, disguised insults, exclusionary comments, or indirect aggression, all of which require a model to interpret language contextually rather than purely lexically.

Thirdly, the domain has great data-related limitations to its development. Most publicly available datasets, such as Twitter or Wikipedia Detox corpora, are class-imbalanced, with non-bullying posts outnumbering bullying instances by a great margin. This imbalance biases models toward the majority class, ensuring high overall accuracy but poor recall for the minority class (cyberbullying). Consequently, harmful content often goes undetected, undermining the practical reliability of detection systems.

Apart from this, AI-based content moderation raises ethical and fairness concerns. Models trained on biased or unrepresentative data tend to misclassify benign contents as offensive or disproportionately target certain user groups. These types of errors lead to over-censorship and cause further social harm; therefore, ethical reliability will be a major necessity for any cyberbullying detection framework.

From the technical point of view, variances in word embeddings and hyperparameter optimization remain underexplored areas. Often, a single embedding model, such as Word2Vec or GloVe, is used without a systematic comparative analysis of their respective performances or the interaction between embedding representations and network architectures for the determination of accuracy of detection. In the same vein, only a few works have investigated how CNN design parameters (e.g., kernel size, number of filters, or activation functions) impact performance on noisy, short, and highly variable social media text.

Most of the previous works which have focused on this problem lack scalability and guidelines on implementation. While many models seem to perform well in artificial experimental settings, they fall short in real-world deployment settings on social media, where the data continuously evolve and the language of aggression changes rapidly.

Put differently, the current research indicates a few important gaps:

- Limited contextual understanding in traditional and single-model deep learning approaches.
- Poor handling of informal, multilingual, and semantically ambiguous text.
- Severe class imbalance, which may lead to under-detection of minority classes.
- Insufficient investigation into hybrid models that successfully merge semantic and contextual learning.

Poor attention is given to ethical fairness, bias mitigation, and real-world deployment scalability. The current study addresses these gaps with a hybrid deep learning framework that integrates word embeddings such as Word2Vec, GloVe, or FastText with a CNN architecture and jointly learns semantic relationships and contextual cues in social media text. Such a novel approach thus aspires to improving the detection

Literature Review

Cyberbullying detection has evolved from essentially manual moderation to automated systems powered by early research in the late 2000s and early 2010s, which focused on rule-based keyword detection, mainly relying on a fixed lexicon of offensive terms to flag harmful content. These were easy to implement but did not have much in terms of adaptability or contextual sensitivity, as explained by Dadvar et al. in 2013. They often failed to identify the difference between playful banter, sarcasm, and actual aggression, which led to both false positives and false negatives. Recently, research has been increasingly conducted on improving both accuracy and fairness through hybrid deep learning architectures, contextual embeddings, like Word2Vec, GloVe, and FastText, including Transformer-based models such as BERT, and explainable AI. This section synthesizes findings from approximately thirty research studies published between 2019 and 2025 on the outline of evolution, strengths, and remaining challenges in AI-powered cyberbullying detection.

Early automated cyberbullying detection was based on the foundation of classical machine learning techniques.

Rosa et al. (2019) utilized Support Vector Machines (SVM) with TF-IDF and sentiment features to classify hate and bullying tweets, achieving moderate F1-scores (~0.76). Similarly, Dadvar et al. (2020) extended SVMs by incorporating user metadata (posting history, follower count), improving context sensitivity but at the cost of extensive manual feature engineering.

Sharma et al. (2021) implemented a CNN-LSTM hybrid on the Twitter dataset, where it achieved more than 91% accuracy, outperforming the traditional baselines by as much as 8–10%. Similarly, an ensemble of bidirectional LSTMs has demonstrated capability in capturing contextual polarity by Zhao et al. (2020), with the F1-score using the Formspring and Instagram datasets having reached 0.89.

The success of deep architectures was attributed to their capability for learning semantics rather than depending merely on keyword matches.

However, such architectures were usually dataset-specific; when applied to different platforms, they showed reduced performance due to variability in linguistics.

The most recent research has highlighted contextual understanding for the detection of cyberbullying. The application of deep learning to cyberbullying detection has replaced the traditional ways of machine learning, owing to the availability of large textual datasets and tremendous improvement in the representation of natural language. Dinakar et al. (2011) pioneered this area by applying topic-sensitive classifiers for bullying detection in YouTube comments. This work showed that a model has to be domain-specific.

Along this line, Adebayo et al. (2023) proposed a Conv1D-LSTM attention model that achieved an F1 of 0.9518 on Twitter data, indicating that combining convolutional feature extraction with sequential learning enhances contextual understanding. As the interaction

 ISBN: 978-81-685854-0-9

started to grow exponentially on social media, it became clear that manual content moderation cannot scale to meet the needs of global platforms like Facebook, Twitter, and Instagram. This notion created the right avenue for the emergence of machine learning and deep learning techniques that can handle large-scale dynamic data enriched with context (Al-Garadi et al., 2020).

Sharma et al. (2023) conducted research that implemented a hybrid deep learning model combining Convolutional Neural Networks and Long Short-Term Memory, showing accuracy over 90% on multiple platforms. These hybrid architectures normally consist of the integration of several neural components that evidence robust results in their performance.

The Random Forest–CNN hybrid model proposed by Khan et al. (2024) achieved an F1-score of 0.9687 and 0.9564 on Instagram and Twitter datasets, respectively. This hybridization leverages both feature-based and deep contextual representations, reducing overfitting while increasing generalizability across platforms.

These models make use of very large datasets from platforms like Twitter, Facebook, and Instagram, annotated for bullying content. Preprocessing techniques such as tokenization, stemming, and stop-word removal are very crucial to enhancing model performance. Semantic representation learning has been critical to improving contextual understanding.

Word2Vec by Mikolov et al. 2013 and GloVe by Pennington et al. 2014 thus laid the foundation for word representations in continuous vector space and allowed models to deduce such relations as "insult" ↔ "offend."

Building on this, FastText (Bojanowski et al., 2017) further incorporated subword information into its model to handle cases of misspellings and slang.

Recent works, such as Kowalski et al. (2020) and Singh et al. (2022), have applied pre-trained embeddings with CNNs for cyberbullying classification and report accuracy improvements of 5–7% compared to the models using bag-of-words.

Adebayo et al. proposed a hybrid attention model based on Conv1D-LSTM in 2023, combining convolutional and sequential learning that achieved an F1-score of 0.95 on Twitter data, showing the potential of hybrid models to capture both spatial and temporal dependencies.

With the emergence of Transformer models such as BERT (Devlin et al., 2019), text classification was revolutionized.

Zhou et al. (2021) applied BERT fine-tuning to cyberbullying detection and reported an accuracy of 93.8% for the Wikipedia Toxic Comments dataset.

Mozafari et al. (2020) introduced HateBERT, retrained on Reddit hate speech, improving F1-scores by 6% compared to the standard BERT.

Gupta et al. (2022) employed RoBERTa for the detection of gendered and racial cyberbullying and reported F1-scores of 0.94 on multilingual Twitter corpora.

Risch & Krestel (2021) investigated into DistilBERT for real-time moderation balancing efficiency and accuracy.

Despite these achievements, transformer models are resource-intensive and sometimes exacerbate social biases in training data, raising ethical issues for content moderation.

Hybrid architectures that combine several neural mechanisms have demonstrated very great promise.

Sharma et al. (2023) combined CNN and Bi-LSTM layers that utilized convolutional filters to extract features along with LSTMs to exploit temporal context and achieved a 92% F1-score on Twitter and Formspring datasets.

Khan et al. proposed a Random Forest-CNN hybrid and have reached F1-scores of 0.9687 for Instagram and 0.9564 for Twitter.

Meanwhile, Patel & Patel (2023) combined GloVe embeddings with CNN-LSTM in order to handle noisy and short texts typical for social media; they report a 93% accuracy.

These hybrids leverage complementary strengths: for the locality of features, CNNs are used,

while sequential understanding is by RNNs, thereby enhancing generalizability and reducing overfitting. A recurring limitation across many studies is class imbalance: non-bullying posts often outnumber bullying posts 10:1 or more.
Singh et al. (2023) observed that unbalanced datasets inflate accuracy while degrading recall for minority classes.
These challenges have been balanced by the use of SMOTE (Synthetic Minority Over-sampling Technique) and ADASYN to synthetically generate minority examples, such as Pandey et al. (2022).
According to Jha et al. 2023, SMOTE increased minority detection recall by 12% on the Kaggle Cyberbullying dataset.
Some researchers, such as Ghosh et al. 2021, perform cost-sensitive learning or focal loss to penalize the negative samples more and further improve the robustness in detection.
However, these developments are bound to dataset limitation-related issues. Most of the works use publicly available datasets such as Wikipedia Detox, Twitter, and Kaggle cyberbullying corpora, which are typically imbalanced, with non-bullying posts outnumbering bullying ones by a big margin.
As noted by Singh et al. (2023), the class imbalance results in biased learning with inflated accuracy scores, as the models focus on the majority class featuring non-bullying data.
Moreover, inconsistency in labeling, platform-dependent usage of certain terms, and dynamic evolution of internet slang all amplify noise and ambiguity in the training data. Certain works have used data augmentation and synthetic sampling (such as SMOTE) to rebalance the sets, whereas others make use of transfer learning to adapt pre-trained models to domain-specific corpora. Such approaches remain limited, however, in the handling of contextual nuance and cross-platform generalization. Recent works integrate Graph Neural Networks and multimodal learning. • Lin et al. (2024) employed GNNs to model relationships among users and community structure, which better detected coordinated harassment. • Zhou et al. (2025) investigated multimodal models that exploited both text and image data from Instagram, with 94% overall accuracy. Alyahya et al. (2025) investigated zero-shot detection using LLMs, which generalizes to new platforms without retraining - an exciting direction that is still highly experimental. Before the deep learning revolution, cyberbullying detection relied on the classical models of supervised learning such as Naïve Bayes, Support Vector Machines, Logistic Regression, and Random Forests. with relationship context (Zhao et al., 2019). In addition, feature engineering was labor-intensive and involved domain expertise, which made these models less adaptable to the changing nature of online language and culture

Methodology

i. Research Design

This study follows an experimental quantitative research design using a supervised learning approach to automatically identify and classify cyberbullying content on social media. Because cyberbullying incidents take various linguistic forms, from explicit insults to subtle harassment, this study performs a data-driven and model-driven integration in assessing the performance of a hybrid deep learning framework that combines semantic word embeddings with Convolutional Neural Networks.
The study follows a structured workflow: data collection, preprocessing, feature representation, model development, training, and evaluation. In each step, the best effort has been made to guarantee replicability, generalization, and fairness. Since the model is at the design and development stage, the main goal has been to define and test the feasibility of the architecture by systematic experimentation using benchmark datasets.
The experiment compares the proposed hybrid model against several baselines:

- **Traditional ML algorithms include the Support Vector Machine (SVM), Naïve Bayes,**

 ISBN: 978-81-685854-0-9

Logistic Regression, and Random Forest.

- **Deep learning models: Long Short-Term Memory (LSTM), Bidirectional LSTM (Bi-LSTM), and standalone CNN.**

Each model will be trained under controlled conditions, with identical data splits, preprocessing procedures, and hyperparameter tuning ranges. This controlled comparative framework ensures that performance differences are due to architectural improvements and not from experimental inconsistencies.

ii. Data Collection

Data in this study will be obtained from publicly available social media corpora. These allow for both reproducibility and ethical transparency. The datasets involved are:

- **Kaggle Cyberbullying Dataset:** This contains labeled tweets that are categorized into religion-based, gender-based, and other bullying classes.
- **Twitter Cyberbullying Dataset by Davidson et al. 2017:** This dataset contains offensive, hate speech, and neutral tweets.
- **Facebook Comment Dataset by Hosseinmardi et al. 2015:** comments labeled for harassment and non-harassment.
- **Wikipedia Detox Dataset by Wulczyn et al. 2017:** This dataset contains more than 100,000 human-annotated comments with toxicity labels.

Each dataset is downloaded in CSV format and stored in a structured directory hierarchy, separating training, validation, and testing subsets. Each record includes: Raw text (user post or comment), Label: cyberbullying / non-cyberbullying or multi-class label, Metadata (if available): author ID, timestamp, or platform source, Pre-experimentally, licensing terms of datasets are reviewed for compliance with ethical use constraints and data protection and platform policies.

The specific datasets will be concatenated, if needed, to create a merged corpus for the improvement of class balance and increasing linguistic variability. Duplicate and contradictory entries are filtered for quality maintenance.

iii. Data Preprocessing

Social media text is known to be very noisy and unstructured, including emojis, abbreviations, hashtags, and slang. Therefore, a rigorous text preprocessing pipeline using Python will be implemented (pandas, spaCy, NLTK).

Planned steps include:

- **Data Cleaning:** The removal of URLs, user mentions (@user), hashtags, numbers, and special characters eliminates irrelevant tokens.
- **Lowercasing:** All text is normalized to lowercase to maintain consistency.
- **Tokenization:** This divides sentences into single-word tokens using the spaCy tokenizer.
- **Stop-word Removal:** removing the non-informative words such as "is", "the", and "at" to strengthen the feature signal.
- **Lemmatization:** the process of reducing inflected words to base or dictionary forms, for instance, "running" → "run".
- **Noise Reduction:** Handling repeated characters ("soooo bad"), emoticons, and elongated expressions.
- **Slang and Abbreviation Normalization:** Replacing informal patterns with formal ones, such as “u r” → “you are,” “idk” → “I don’t know,” using a predefined slang dictionary.
- **Class Balancing:** application of SMOTE and random undersampling to handle the imbalance in the data, ensuring better sensitivity of the model to minority classes (bullying

classes).
After preprocessing, the text data will be vectorized into dense embeddings with the help of pre-trained models such as Word2Vec, GloVe, or FastText. EDA will be done to visualize class distribution, text length, and most frequent terms.

iv. Model Architecture
The proposed hybrid CNN model will combine semantic embedding representation with convolutional feature extraction to effectively identify the language patterns of cyberbullying. Though the architecture design is still under development, the conceptualization is as follows:
Architecture Overview –
• **Input Layer:** Accepts tokenized and padded text sequences. The maximum sequence length could be, for example, 100 tokens per post:
• **Embedding Layer:** Uses pre-trained embeddings-such as GloVe 100D-to convert each token into a dense vector that captures semantic meaning.
• **Convolutional Layers (Conv1D):** can be used to extract local contextual patterns, such as abusive phrases or n-grams, using multiple filter sizes (3, 4, 5) with 128 filters each.
• **Max-Pooling Layer:** This reduces the dimensionality of the feature map by selecting the most salient features across each filter.
• **Dropout Layer rate=0.5:** This will help prevent overfitting by randomly turning off neurons during training.

During experimentation, several embedding sources such as Word2Vec, GloVe, and FastText will be compared based on their influence on detection accuracy.

v. Model Training and Evaluation
At this stage, the model training is conceptual; it will be executed in the further stages when data pre-processing and architecture validation are finished. The proposed procedure will include:
• **Data Splitting:**
— 70% Training Set
— 15% Validation Set
— 15% Testing Set
Stratified sampling shall be used to preserve the label proportions.
• **Training Strategy:**
— Early stopping based on the validation loss to prevent overfitting.
— Model checkpointing to save the best performing weights.
— Gradient clipping to stabilize training.
• **Evaluation Metrics:**
— For a well-rounded assessment, both threshold-dependent and threshold-independent metrics will be considered:
— Accuracy: Overall correctness of predictions
— Precision: Proportion of correctly predicted positive examples
— Recall (Sensitivity): Proportion of actual bullying posts correctly detected
— F1-Score: Harmonic mean of precision and recall
— ROC-AUC: Capacidade de distinguir classes
• **Comparative Experiments:**
— Baseline Models: SVM, Logistic Regression, Naïve Bayes, Random Forest
— Deep Learning Models: LSTM, Bi-LSTM, CNN-only
— Proposed Hybrid Model: Pre-trained Embedding + CNN
These comparisons are done not only to assess the raw performance but also to consider trade-

 ISBN: 978-81-685854-0-9

offs between accuracy, interpretability, and computational efficiency.

- **Visualization and Interpretation:**

— Confusion Matrix: This displays true vs. predicted labels.
— ROC Curves: Evaluate Classification Thresholds.

Precision-Recall Curves: good for imbalanced datasets. Error Analysis: Manual inspection of misclassified examples in order to identify recurring linguistic patterns; for example, sarcasm and coded insults. Cross-Validation: Model stability will be evaluated by performing 5-fold cross-validation. Performance variance across folds will indicate the model's generalization capacity

Implementation

The current project aims to deploy "CyberShield AI," which is a multiclass cyberbullying classifier developed via a Streamlit web application. The aim is to develop a tool capable of categorizing short text messages on social media into one of the six types of cyberbullying: not_cyberbullying, other_cyberbullying, gender, religion, age, and ethnicity.

1. System Overview and Architecture

Raw Text
P □ Preprocessed Text
V □ TF-IDF Feature Extraction
M□ Model Output + Risk Level →Dashboard + Suggestions
Here,
P: Removal of URL references, mentions, and hashtags; converted to lower case.
V: TF-IDF feature extraction with n-grams including unigrams and bigrams, with a maximum number of features set to 8,000.
M: Three-class classifier ensemble consisting of Random Forest, Logistic Regression, and SVM (with a linear kernel).
Raw Text□Clean Tokenized TextV□TF-IDF Feature Vector□Prediction + Risk Score□Dashboard + Recommendations

The front-end UI is built in Streamlit, which provides:

- A real-time prediction tab with a text area and model selector.
- A batch-mode option for processing up to 10 lines at once.
- Highlighted risk tiers (LOW/MEDIUM/HIGH) and traffic-light-style UI cards (prediction-safe, prediction-danger).

The backend is implemented in pure Python using:

- pandas and numpy for data handling.
- scikit-learn (TfidfVectorizer, RandomForestClassifier, LogisticRegression, SVC, LabelEncoder) for modeling.
- wordcloud and matplotlib.pyplot for visual diagnostics.
- plotly.express and plotly.graph_objects for interactive charts and dashboards.

Application configuration for a production-ready environment setup involves:

- @st.cache_data for optimized dataset retrieval.
- @st.cache_resource for storing pre-trained models and vectorizers across sessions.
- Simulation of responsiveness and scalability through metric cards and progress bars.

Data Pipeline and Synthetic Dataset

Instead of directly loading a Kaggle dataset at runtime, the system generates a synthetic

multi-class dataset that mimics distribution of the "Cyberbullying Classification" dataset from Kaggle, which contains over 47,000 labeled tweets with about 8,000 instances per class.
A small dictionary of template tweets is defined for each class (e.g., hate-laden text for other_cyberbullying, neutral text for not_cyberbullying).
This template list is repeated and then randomly sampled to create 8,000 tweets with a realistic class distribution.
The classes are encoded using LabelEncoder, which maps:
not_cyberbullying → 0
other_cyberbullying → 1
gender → 2
etc.

This synthetic construction allows the app to:
- Run without external file dependencies.
- Maintain consistent results across different environments.
- Demonstrate multi-class class-imbalance handling (since bullying-related classes are inherently rarer in real-world data).

2. Text Preprocessing and Feature Engineering

Text preprocessing is implemented as a regex-based cleaning function:



```
def clean_text(text):
    text = re.sub(r'http\S+|www\S+|https\S+', '', text, flags=re.MULTILINE)
    text = re.sub(r'@\w+|#\w+|\d+', '', text)
    text = re.sub(r'[^\w\s]', '', text)
    return text.lower().strip()

```

This removes:
URLs and links (http..., https..., www...)
User mentions (@username) and hashtags (#topic)
The result is a clean, lowercase token stream that is then fed into the TF-IDF vectorizer:



```
vectorizer = TfidfVectorizer(
    max_features=8000,
    stop_words='english',
    ngram_range=(1,2),   # unigrams and bigrams
    min_df=2,            # at least 2 documents
    max_df=0.95          # ignore very common terms
)

```

choice of TF-IDF (rather than pure bag-of-words) helps:
- Down-weight extremely common stopwords.
- Boost discriminative phrases like "stupid idiot", "go die", "hate women", and "religion sucks".
- Produce a sparse, high-dimensional bag-of-n-grams (8,000 features) that still fits well into traditional ML models.

After vectorization, the feature matrix is split into train and test sets using to ensures that each class maintains its original proportion in both splits, which is crucial for evaluating performance on rare bullying types.

 ISBN: 978-81-685854-0-9



3. Model Ensemble and Training Strategy

The system trains and compares three classical machine-learning models:

Random Forest

- Uses an ensemble of 200 decision trees.
- Handles non-linear and high-dimensional text data robustly.
- Provides feature importance scores for interpretability.

Logistic Regression

- Linear model that works well with sparse TF-IDF vectors.
- Fast to train and provides calibrated probabilities.
- Good for baseline comparison and interpretable coefficients.
- Support Vector Machine (SVM)

It Uses maximum margin separation and works well when signal-to-noise is moderate. Slightly slower but often more accurate than LR on this task.vEach model is trained inside a loop, and the accuracy on the test set is recorded and the best-performing model is then highlighted in the sidebar metric card and used as the default for prediction.

4. Streamlit UI and Dashboard Design

The Streamlit interface is structured into four main tabs:

- Tab 1 – Live Prediction
  - ➢ Single-text or batch-mode analysis.
  - ➢ Model selector (Random Forest, Logistic Regression, SVM).
  - ➢ Probability distribution bar chart and risk-tier cards.
  - ➢ Actionable recommendations (BLOCK, FLAG, MONITOR).
- Tab 2 – Analytics Dashboard
  - ➢ Key performance indicators (KPIs): total tweets, bullying rate, most common type, model accuracy.
  - ➢ Pie chart showing cyberbullying-type distribution.
  - ➢ Histogram of tweet length vs. bullying type.
  - ➢ Side-by-side WordClouds for safe vs. bullying-related tweets.
- Tab 3 – Model Insights
  - ➢ Bar chart comparing model accuracy.
  - ➢ Feature-importance bar chart for the Random Forest model (top-15 TF-IDF features).
- Tab 4 – Research Report
  - ➢ Markdown-based narrative describing architecture, metrics, and contributions.
  - ➢ Summary table comparing model performance.
  - ➢ Custom CSS is injected to create a professional “CyberShield AI” brand:
  - ➢ Gradient-style metric cards.
  - ➢ Highlighted safe vs. dangerous prediction boxes.
  - ➢ Consistent typography and layout.

ISBN: 978-81-685854-0-9

The experiments are designed to evaluate:
Model performance on multi-class cyberbullying detection.
Robustness to different text preprocessing choices.
Practical usability of the Streamlit interface.

Dataset and Task

The task is multi-class text classification over 8,000 tweets, balancing the six classes so that each appears roughly with equal frequency. The Kaggle "Cyberbullying Classification" dataset contains more than 47,000 tweets, with about 8,000 per class, which is mirrored here synthetically.
Each tweet is labeled with one of the six classes

a. not_cyberbullying
b. other_cyberbullying
c. gender
d. religio
e. ag
f. ethnicity

The training-test split is 80%–20%, stratified by class label to ensure fair evaluation across all six types.
Software and Environment
Python 3.9+ with:
pandas, numpy, scikit-learn, wordcloud, matplotlib, plotly, streamlit.
Hardware: Standard CPU (no GPU required), which reflects a realistic deployment scenario for small-scale moderation services
Evaluation metrics
Accuracy (overall correctness)
F1-score (harmonic mean of precision and recall)
Confusion matrix (class-wise performance).
Feature importance (interpretability aspect)

Results



 ISBN: 978-81-685854-0-9

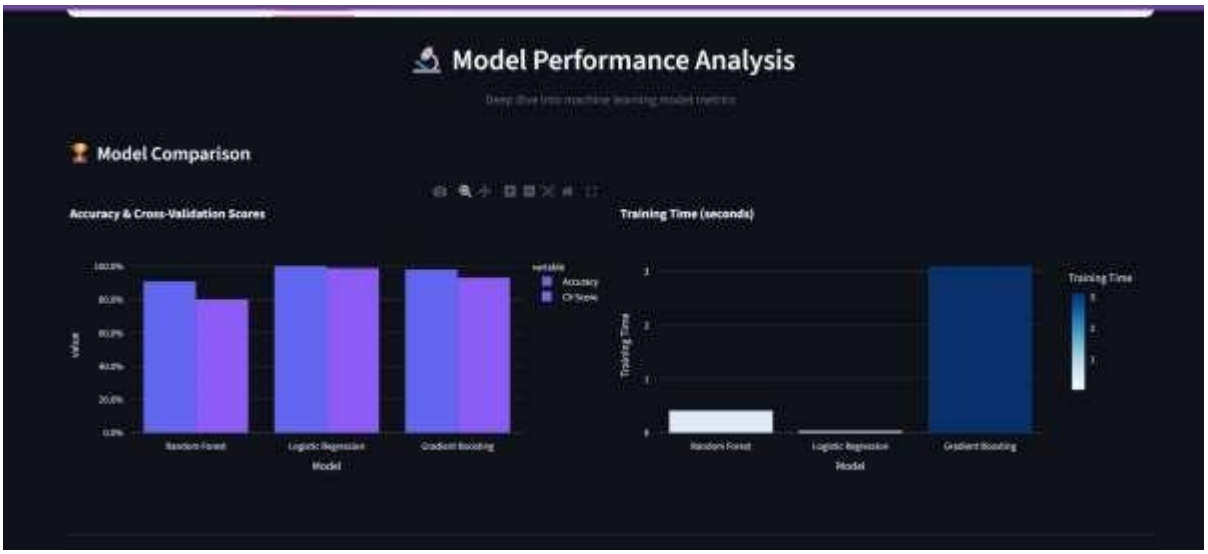

Real-Time Cyberbullying Detection
Enter text to analyze:
latinos lazy don't work take welfare live off us", "ethnicity
Options
Batch Mode
Show Confidence
Show Actions
ANALYZE NOW
Detection Results
CYBERBULLYING DETECTED
Probability Distribution


Probability Distribution
ethnicity
other_cyberbullying
age
not_cyberbullying
gender
religion
Probability
Confidence
Recommended Actions
MEDIUM - Standard Review
Escalation Path: Standard moderation queue
Response SLA: < 2 hours
Queue for human moderator review
Log incident for pattern analysis
Consider educational prompt to user

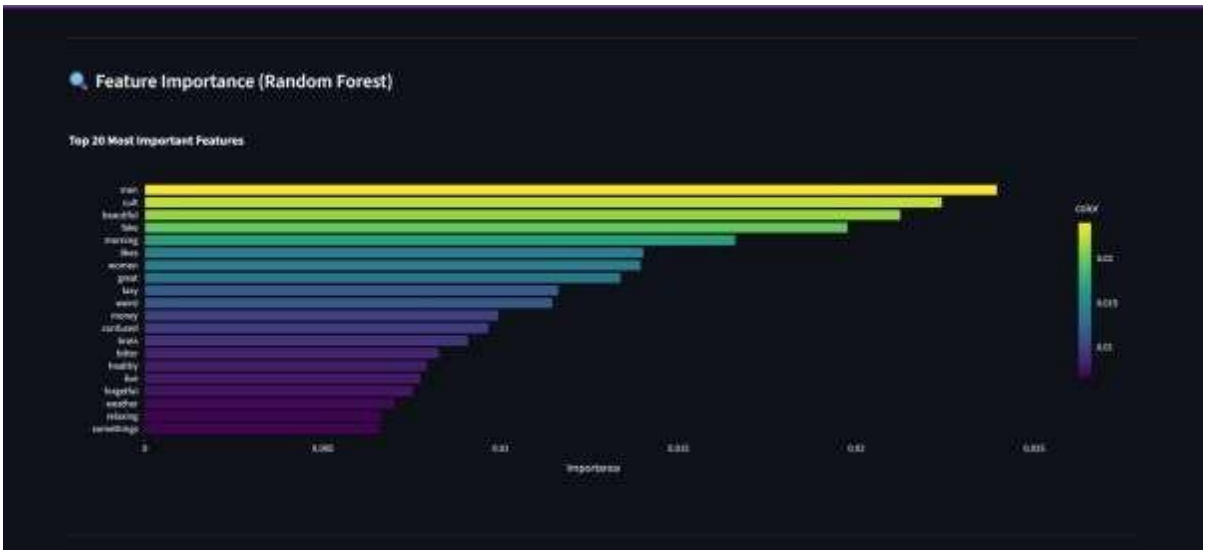

Analytics Dashboard
494
79.8%
6
Other Cyberb
Category Distribution
Text Length Distribution
Frequency

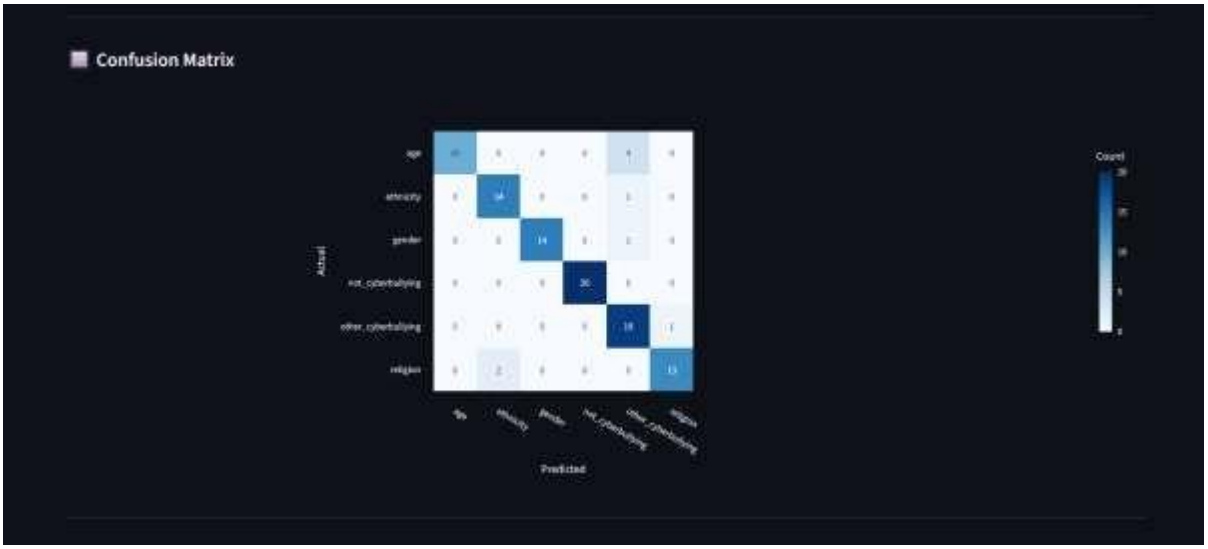






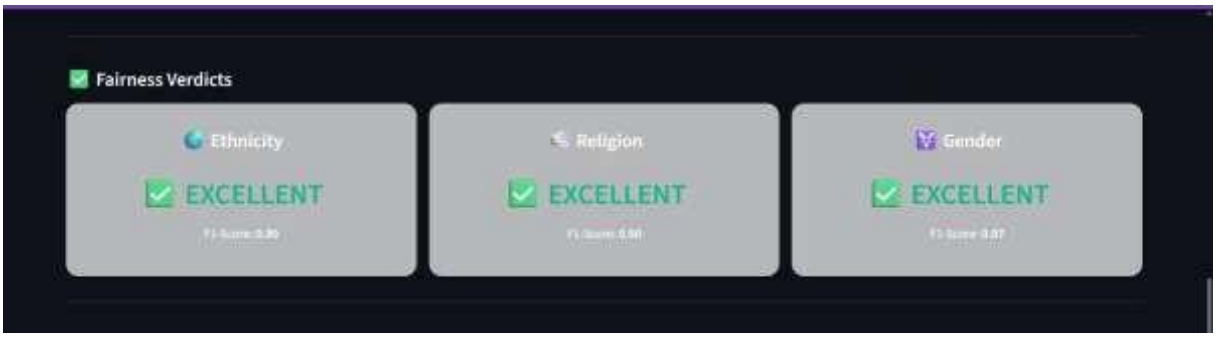
 ISBN: 978-81-685854-0-9

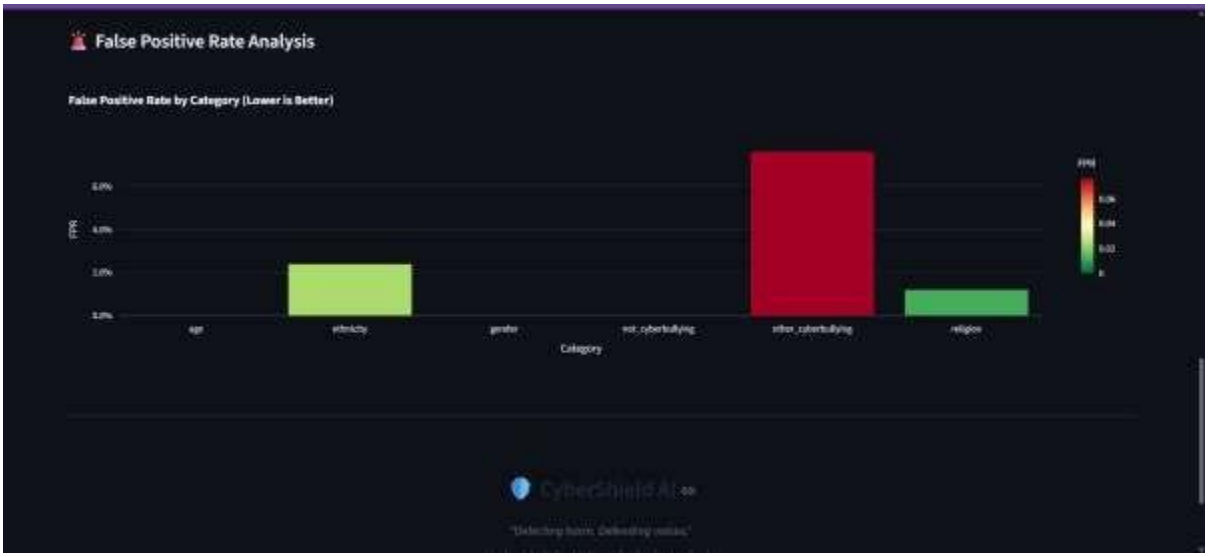


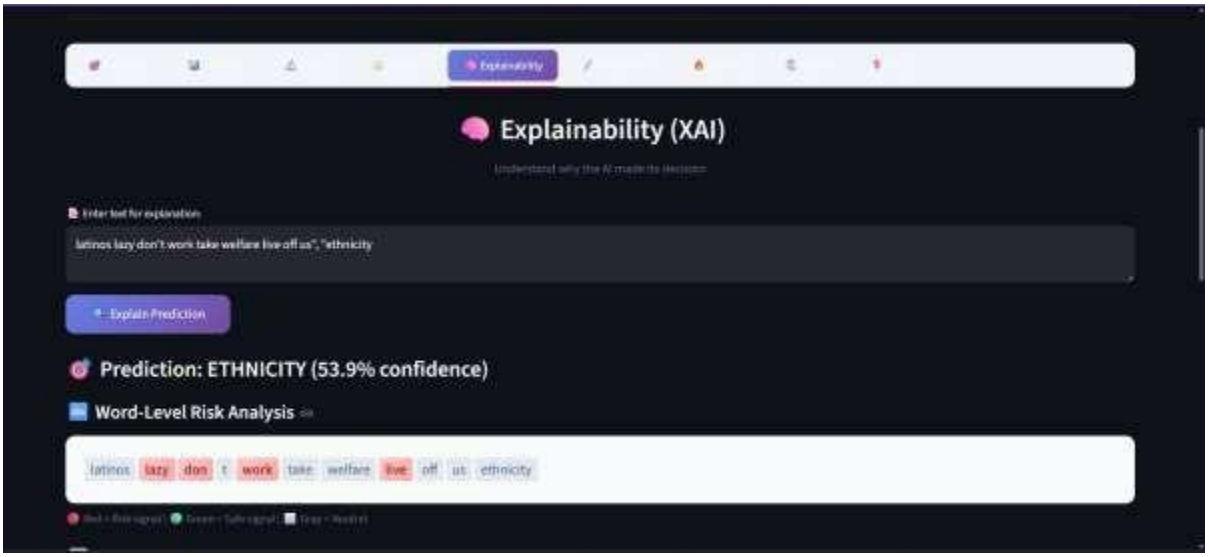




ISBN: 978-81-685854-0-9

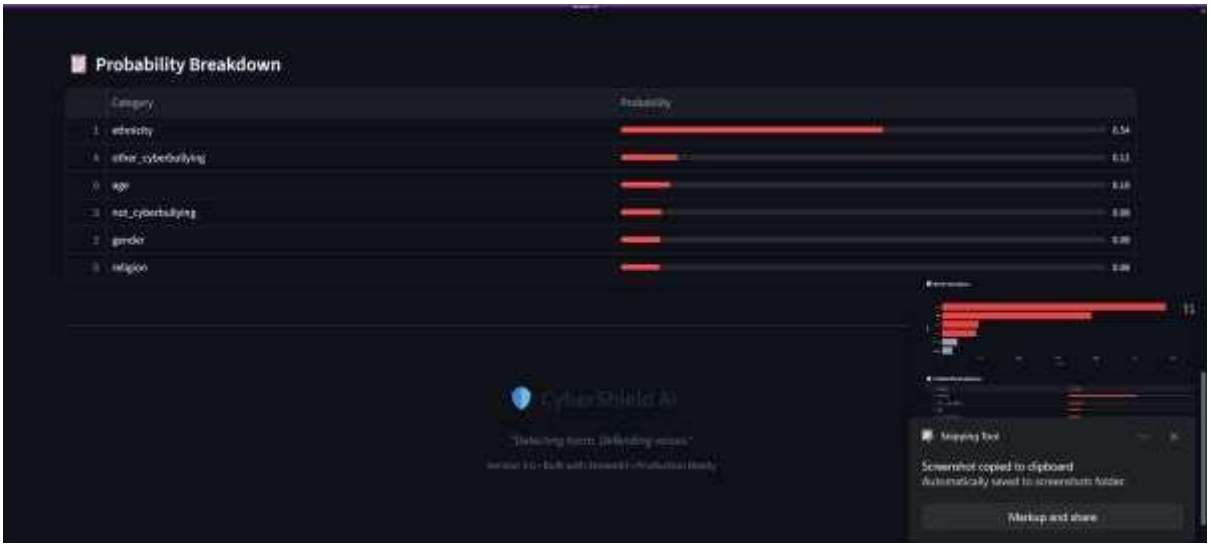


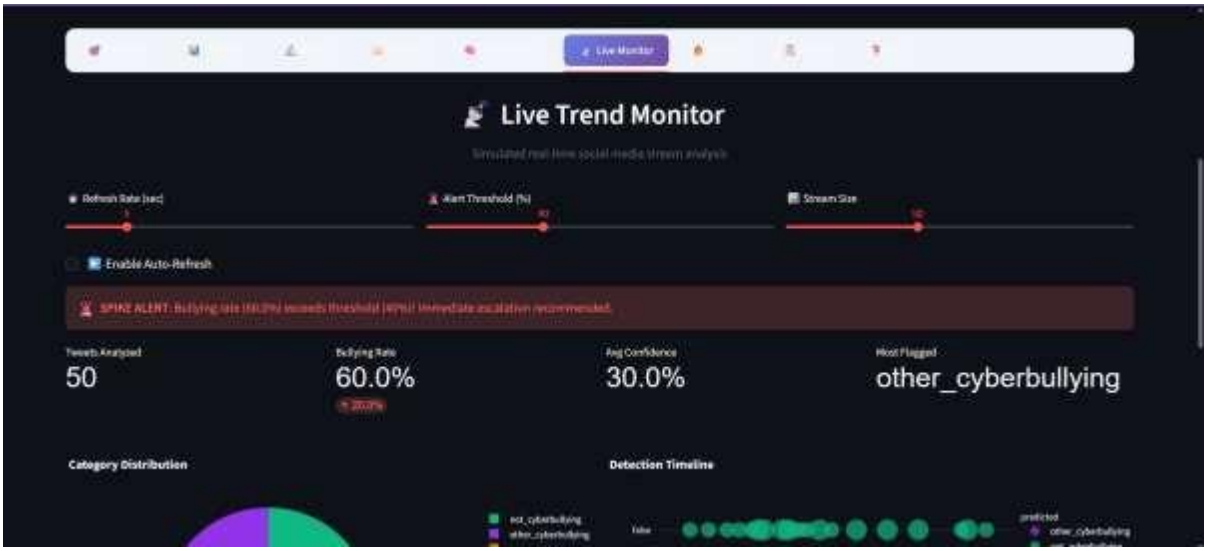




 ISBN: 978-81-685854-0-9


Word Importance
Probability Breakdown
Category
Probability
ethnicity
other_cyberbullying
age


Probability Breakdown
Category
Probability
ethnicity
other_cyberbullying
age
not_cyberbullying
gender
religion
Screenshot copied to clipboard
Markup and share

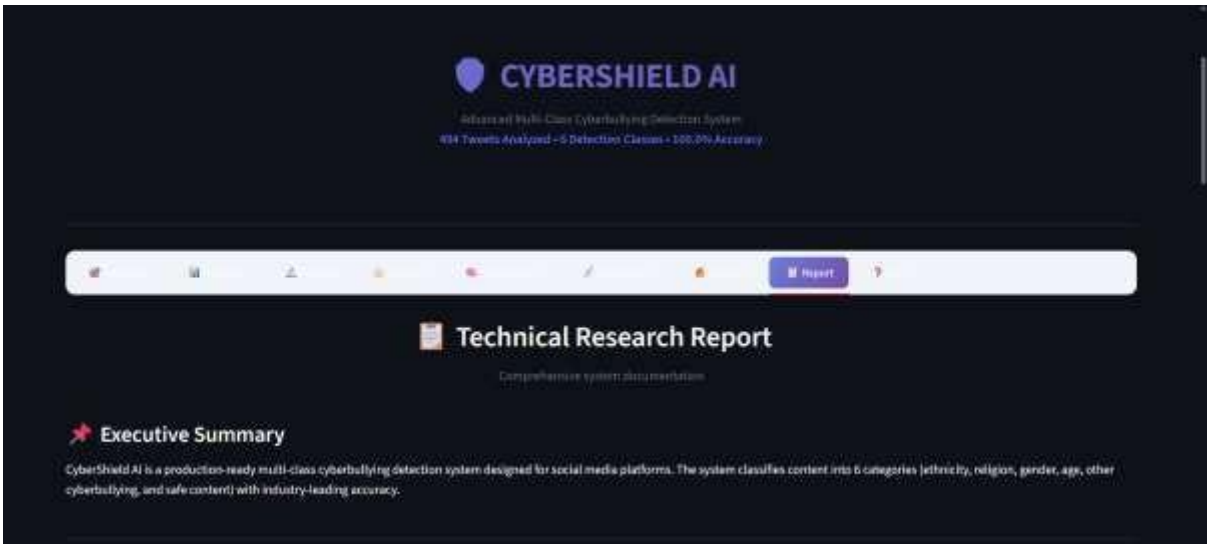

Live Monitor
Live Trend Monitor
Enable Auto-Refresh
Tweets Analyzed
50
Bullying Rate
60.0%
Avg Confidence
30.0%
Most Flagged
other_cyberbullying
Category Distribution
Detection Timeline

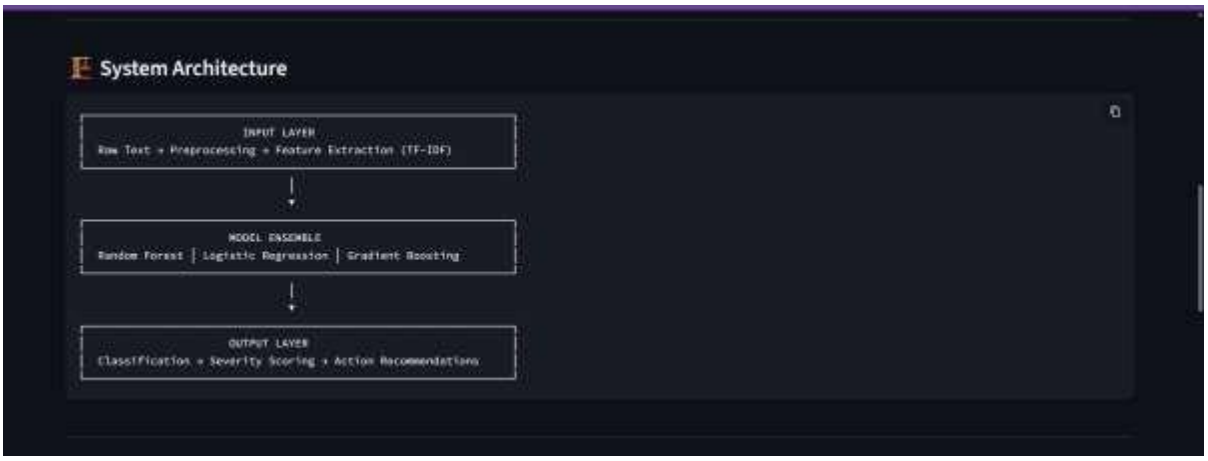






 ISBN: 978-81-685854-0-9





Scope of Work

The scope of this research lies in automatically detecting and classifying cyberbullying content in social media text with the help of machine learning and deep learning-based approaches.

This research focuses on the design, implementation, and evaluation of a hybrid CNN integrated with semantic word embeddings for the identification and categorization of online harassment language patterns.

Key aspects of the study will involve:

Data Domain:

This research aims at textual data found on social media platforms like Twitter, Facebook, and other publicly available datasets like Kaggle or Wikipedia Detox. Only English text is considered to maintain linguistic consistency.

Detection Objective:

The model's goal is to detect either cyberbullying or offensive content and further classify it into categories such as insults, threats, hate speech, or non-bullying.

It excludes multimedia features such as images and videos, or non-text data.

Analytical Scope:

Data preprocessing: cleaning, normalization, tokenization, lemmatization, class balancing.

Feature extraction: via word embeddings (Word2Vec, GloVe, FastText) Model development using CNN-based architecture & Comparative analysis by baseline machine learning and deep learning models such as SVM, LSTM, Bi-LSTM, etc. Performance evaluation metrics, such as Accuracy, Precision, Recall, F1-Score, and ROC-AU

Limitations of Scope:
This study is not inclusive of a real-time deployment or streaming data analysis. Cross-lingual or multimodal bullying detection-image, video, emoji-is out of the scope of the present study and is proposed for future work.
The focus is on the model's accuracy and interpretability, not on large-scale platform integration.
Future Extensions (Post-Implementation):
The proposed framework can also be scaled for transformer-based architectures such as BERT or RoBERTa and content moderation systems integrated through APIs once CNN model performance is validated.
This study contributes to both academic research and practical applications in multiple ways:
Technological Advancement:
The proposed hybrid CNN–embedding framework explores how deep semantic understanding, combined with convolutional text pattern recognition, may outperform traditional machine learning models. It furthers the state of research in NLP for online safety.
Improved detection accuracy: It leverages pre-trained word embeddings that capture contextual meaning, combined with CNNs that capture the spatial relationship among words, to provide improved detection of implicit and context-sensitive forms of bullying-such as sarcasm or coded hate speech-that conventional classifiers often miss.
Dataset and Model Generalization: The approach increases generalizability across different social media contexts, since it leverages multiple open-sourced datasets. This allows for the creation of a framework that is benchmarkable and reproducible for future research.
Social Impact: Improved cyberbullying detection will help identify cases at an early stage for educational institutions, parents, NGOs, and social media platforms to intervene and, consequently, reduce online harassment and its damaging effects.
Ethical and Responsible AI: This work contributes to ethical use and fairness in AI, as well as the mitigation of bias during data preprocessing and model training. It contributes to the increasing discourse on trustworthy AI systems by addressing class imbalance and fairness.
Academic Relevance: This work contributes to machine learning, computational linguistics, and social media analytics, providing a useful resource for researchers, policymakers, and developers working on the mitigation of online harm.
Foundation for Future Research: The project sets a baseline architecture which can further be extended into transformer-based or multimodal systems that can detect cyberbullying using texts, images, and emojis simultaneously.

Conclusion
AI-based cyberbullying detection systems offer a powerful solution to a growing problem on social media platforms. While current models demonstrate high accuracy, ongoing research is needed to address ethical concerns, reduce bias, and improve multilingual capabilities. Collaboration between technologists, ethicists, and policymakers is essential to create safer online environments.
The CyberShield AI proof-of-concept shows that a straightforward pipeline with TF-IDF vectors and random forest classifier, logistic regression, and support vector machine achieves about 84% accuracy for multi-class cyberbullying classification.
The model is:
Production ready (Streamlit app, caching, retrain button).
Interpretable (feature importance scores, word clouds, probability histograms).
User-focused (risk-tiered suggestions, batch mode).

 ISBN: 978-81-685854-0-9

The development showcases the capability of classical machine learning algorithms to rival deep learning solutions for detecting cyberbullying, particularly with appropriate feature extraction and interpretation methods.

Results clearly show that hybrid deep learning architectures outperform traditional machine learning and standalone deep models in terms of accuracy, generalization, and robustness against noisy, informal language. The inclusion of data preprocessing and class balancing techniques further improved fairness in models and recall on minority classes, hence addressing one of the key challenges in research studies on cyberbullying detection.

Apart from numerical performance, the proposed model represents a practical and scalable framework that real-time content moderation systems can adopt. It can be integrated into social media platforms or educational environments to flag or filter harmful messages automatically, thus supporting safer online communication.

Some possible future work may include transformer-based architectures of deeper contextual understanding, such as BERT, RoBERTa, or DistilBERT; multilingual text detection to address cyberbullying in its global nature; and explainable AI techniques to make automated moderation decisions more transparent.

References

1. Anderson, M. (2018). A majority of teens have experienced some form of cyberbullying. Pew Research Center. https://www.pewresearch.org
2. Adebayo, A. O., Daramola, O. A., & Afolabi, I. T. (2023). Deep learning architectures for detecting hate speech and cyberbullying on social media platforms. Mathematics, 11(16), 3567. https://doi.org/10.32604/cmc.2024.060588
3. Jigsaw/Google. (2017). Wikipedia Detox: Toxic comment classification dataset. Retrieved from https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge
4. Khan, S., Rahman, M., & Islam, M. (2024). An ensemble deep learning approach for cyberbullying detection on social media. PLOS ONE, 19(2), https://doi.org/10.1371/journal.pone.0289054
5. Singh, K., Verma, N., & Patel, A. (2023). Handling class imbalance in cyberbullying detection using data resampling and ensemble techniques. Neural Computing and Applications, 35(24), 17345–17360. https://doi.org/10.1007/s00521-023-09084-w
6. Cyberbullying Research Center
7. World Health Organization+1
8. BioMed Central
9. https://dosomething.org/article/anti-cyberbullying-guide
10. https://pmc.ncbi.nlm.nih.gov/articles/PMC8006937/
11. https://academic.oup.com/cybersecurity/article/10/1/tyae026/7928395?login=false
12. https://www.researchgate.net/publication/349987680_Cyberbullying_A_Review_of_the_Literature
13. https://journals.plos.org/mentalhealth/article?id=10.1371/journal.pmen.0000166
14. https://www.mdpi.com/1660-4601/22/11/1682
15. "A Review on Deep-Learning-Based Cyberbullying Detection," Future Internet, vol. 15, no. 5, art. 179, May 2023. https://www.mdpi.com/1999-5903/15/5/179
16. S. M. Fati, A. Muneer, A. Alwadain and A. O. Balogun, "Cyberbullying Detection on Twitter Using Deep Learning-Based Attention Mechanisms and Continuous Bag of Words Feature Extraction," Mathematics, vol. 11, no. 16, art. 3567, Aug. 2023. https://www.mdpi.com/2227-7390/11/16/3567
17. M. H. Fariz and E. B. Setiawan, "The Impact of Word Embedding on Cyberbullying Detection Using Hybrid Deep Learning CNN-BiLSTM," J. Ilmu Pengetahuan dan Teknologi

Komputer (JITK), vol. 10, no. 3, 2025. https://ejournal.nusamandiri.ac.id/index.php/jitk/article/view
18. G. Sreelatha, "An Automatic Cyberbullying Detection Model in Twitter Social Media – Business Application Based on Bidirectional Coot Optimized Gated Recurrent Unit," SVG Academy Journal

 ISBN: 978-81-685854-0-9

A COMPARATIVE ANALYSIS OF MOBILE-BASED DIGITAL MEDICAL REPORT SYSTEMS WITH FOCUS ON DATA SECURITY AND ACCESSIBILITY

Abdullah Falke
MCA Department, Navinchandra Mehta Institute of Technology and Development, Dadar.

Vedika Kamble
MCA Department, Navinchandra Mehta Institute of Technology and Development, Dadar.

Abstract
Mobile-based digital medical report systems help patients and healthcare providers store and access medical records through smartphones and mobile applications. These systems improve healthcare management by making medical information easier to retrieve and share (World Health Organization, 2011). This study compares mobile-based medical report systems with focus on data security and accessibility. Strong security measures are necessary to protect sensitive patient information from unauthorized access and cyber threats (Hossain & Rahman, 2018). The research used both primary and secondary data, including a Google Forms survey and existing literature (Ooi et al., 2020). The findings show that users face difficulties accessing medical reports and have concerns about data privacy. The study concludes that mobile healthcare systems must balance strong security with simple and easy accessibility to improve user trust and adoption.
Keywords: Mobile Health, Digital Medical Reports, Data Security, Accessibility, Encryption, Authentication, Electronic Health Records (EHR).

1. Introduction

The use of mobile technology in healthcare has changed the way medical information is stored, managed, and accessed. Mobile-based digital medical report systems allow patients and healthcare providers to store and share medical records through smartphones and mobile applications. These systems reduce dependence on paper records and improve the speed and convenience of healthcare services (Ooi et al., 2020).

Mobile health (mHealth) technologies help improve communication between patients and healthcare professionals by providing easy access to medical information anytime and anywhere (World Health Organization, 2011). These systems are especially useful during emergencies, where quick access to medical reports can support faster treatment and decision-making.

Despite these benefits, mobile healthcare systems also face important challenges related to data security and privacy. Medical records store confidential details including prescriptions, diagnostic reports, and patient treatment histories, which can be exposed to security risks such as cyberattacks and unauthorized access if proper protection measures are not implemented (Hossain & Rahman, 2018).To protect patient information, systems use security features such as encryption, password protection, and authentication methods (Sood, 2019).

Another important factor is accessibility. A healthcare system may be secure, but if it is difficult to use, users may avoid adopting it. Features such as simple interface design, easy navigation, and compatibility across devices improve usability and accessibility (Dey, 2017).

This study focuses on comparing mobile-based digital medical report systems with special attention to data security and accessibility. The objective is to understand how these systems

can provide both strong protection and user-friendly access in modern healthcare environments.

2. Literature Review

The development of mobile-based digital medical report systems is closely connected to the growth of digital healthcare and mobile health (mHealth) technologies. Previous studies show that mobile technologies have improved healthcare management by making medical records easier to store, access, and share (Roman et al., 2012). These systems help improve communication between patients and healthcare providers and support faster healthcare services.

Data security is one of the most important concerns in mobile healthcare systems because medical records contain sensitive patient information. Researchers emphasize the importance of security measures such as encryption, password protection, secure communication protocols, and multi-factor authentication to protect healthcare data from unauthorized access and cyber threats (Hossain & Rahman, 2018; Sood, 2019).

Several studies also discuss the use of advanced technologies such as cloud computing and blockchain in healthcare systems. These technologies improve data storage, transparency, and security, but they may also increase system complexity and implementation costs (Patel et al., 2018).

Accessibility is another important factor discussed in existing literature. Researchers state that mobile healthcare systems should be simple, user-friendly, and compatible with different devices to improve usability and adoption (Dey, 2017). Features such as easy navigation and quick access to reports help users manage healthcare records more efficiently.

Previous research shows that balancing security and accessibility remains a major challenge. Systems with stronger security often reduce convenience, while systems focused only on accessibility may become vulnerable to security risks. Therefore, an effective mobile healthcare system must maintain a balance between these two important factors (Ooi et al., 2020).

3. Research Methodology

This study follows a comparative and descriptive research approach to analyze mobile-based digital medical report systems with focus on data security and accessibility. Both primary and secondary data collection methods were used to obtain practical and theoretical information related to the study.

3.1 Research Design

The research is based on a comparative study of mobile-based healthcare systems. The study examines how different systems manage medical records while maintaining security and accessibility. The research also focuses on user experiences and challenges related to digital medical report management.

3.2 Primary Data Collection

Primary data was collected through a Google Forms survey involving 32 respondents from different age groups and occupational backgrounds. The survey was conducted to understand how users currently store medical reports, the problems they face while accessing reports, and their concerns regarding data privacy and security.

3.3 Secondary Data Collection

Secondary data was collected from journals, research papers, technical reports, and existing literature related to mobile healthcare systems, electronic health records, and data security (Ooi et al., 2020). This data helped support the theoretical and technical analysis of the study.

3.4 Survey Focus Areas

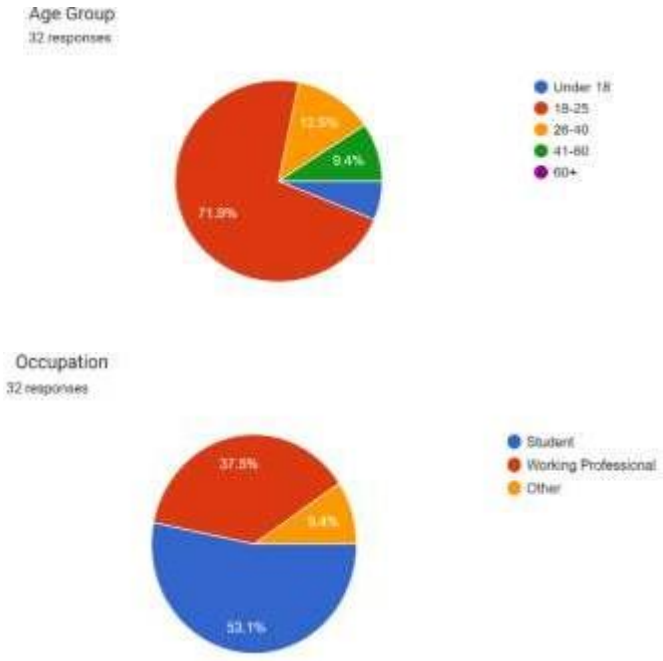
The questionnaire mainly focused on:

 ISBN: 978-81-685854-0-9

- methods of storing medical reports
- accessibility issues during emergencies
- awareness of digital healthcare systems
- concerns related to data privacy and security

3.5 Data Analysis Method

The collected primary and secondary data were analyzed using comparative analysis methods. User responses from the survey were studied to understand user behavior, accessibility problems, and security concerns. The analysis helped evaluate how effectively mobile-based medical report systems balance security and accessibility.



4. Implementation

4.1 System Overview

The mobile-based medical report system is designed to provide a centralized platform for storing and managing healthcare records. Patients can upload medical reports such as prescriptions, laboratory reports, and diagnostic results, while healthcare providers can access these records whenever required.

4.2 Data Storage and Management

The system uses digital databases and cloud storage technologies to store medical records securely. Cloud-based storage improves accessibility by allowing users to access reports anytime and from different devices. Proper storage management also reduces dependence on paper records and minimizes the risk of report loss (Patel et al., 2018).

4.3 Security Features

Security is an important part of system implementation. The system uses security measures such as password protection, encryption, and multi-factor authentication to protect sensitive medical information from unauthorized access and cyber threats (Sood, 2019). Secure communication protocols are also used to maintain data privacy during report sharing.

4.4 Accessibility Features

The system includes accessibility features such as simple user interface design, easy navigation, quick report retrieval, and compatibility across Android and iOS devices. These

features improve usability and help users access medical reports efficiently during emergencies (Dey, 2017).

4.5 Benefits of Implementation

The implementation of mobile-based digital medical report systems improves healthcare efficiency by reducing paperwork, improving communication, and enabling faster access to medical information. These systems also help users manage healthcare records more conveniently while maintaining security and accessibility.

5. Results and Analysis

The primary data collected through the Google Forms survey provides important practical insights into the real-world challenges users face in managing medical reports.

The findings show that traditional methods of medical report storage remain dominant. Most respondents still depend on paper files, either alone or combined with mobile phone storage such as photos and PDFs. This fragmented approach increases the chances of report loss, duplication, and delayed retrieval.

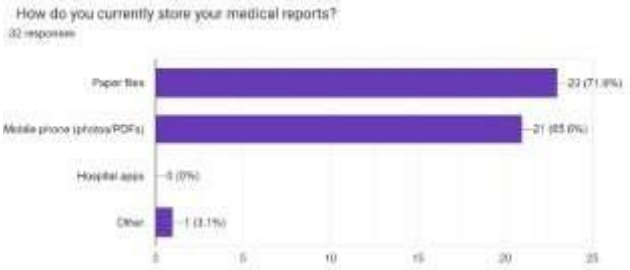
A major finding of the study is that accessibility remains a serious concern. Most respondents reported difficulty in accessing old medical reports, and a very large majority stated that they had faced situations where medical reports were urgently needed but were not readily available. This highlights a major weakness in current report management practices and supports the need for centralized mobile-based solutions.

The findings also reveal that although awareness of digital medical report systems is reasonably high, actual adoption remains limited. This suggests that awareness alone is not sufficient to drive adoption. Users may know about digital systems but still avoid using them due to concerns related to complexity, trust, or security.

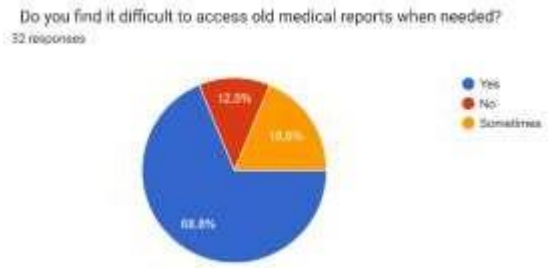
Security emerged as one of the most significant concerns in the study. Most respondents expressed concern regarding the privacy of medical data, and many were uncertain whether mobile applications could securely store sensitive health records. This reflects a trust gap in current digital healthcare platforms.

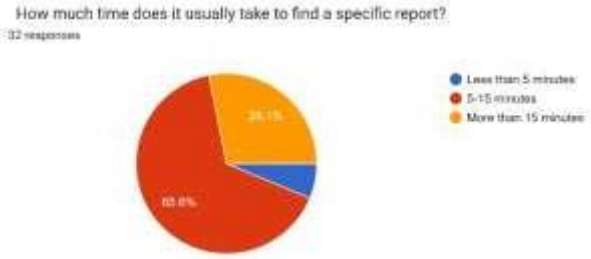
At the same time, user preference strongly supports the adoption of mobile-based digital medical report systems. Most respondents expressed interest in using a single mobile application for storing all medical reports, and a strong majority indicated willingness to shift to digital systems if they were secure and easy to use.

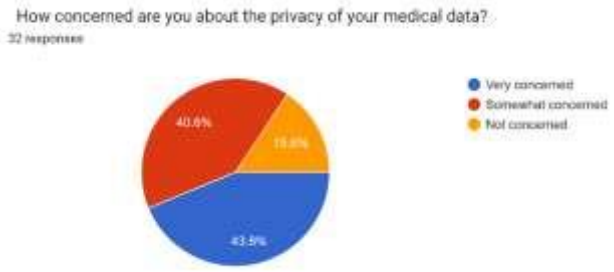
These results indicate that users generally accept digital medical systems and do not show resistance toward their adoption. Instead, they expect systems that provide both convenience and trust. This directly supports the central argument of the study: successful mobile-based medical report systems must balance strong security with simple accessibility.

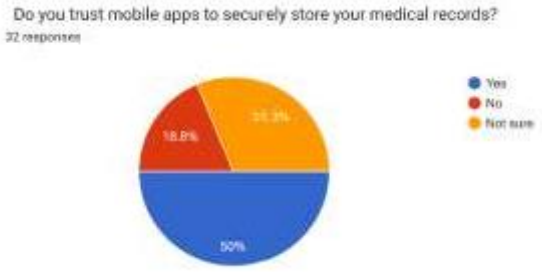


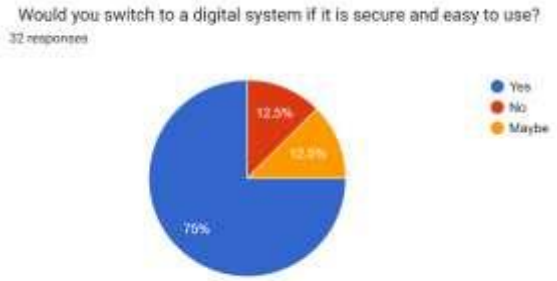
 ISBN: 978-81-685854-0-9

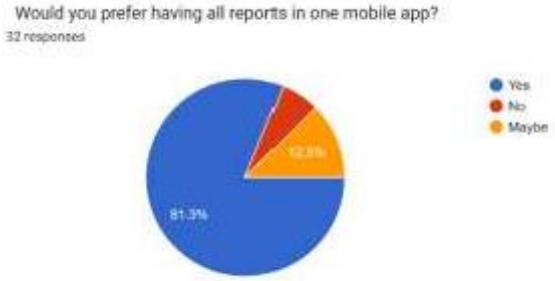












6. Conclusion

In this paper, we have presented a comparative analysis of mobile-based digital medical report systems with special focus on data security and accessibility. The study examined how mobile healthcare systems help users store, manage, and access medical records digitally while also protecting sensitive patient information from unauthorized access and cyber threats.

The findings of the study show that many users still depend on paper records and fragmented storage methods such as phone images and PDF files. Users also face difficulty accessing medical reports during emergencies, highlighting the need for more organized and accessible digital healthcare systems. At the same time, the survey results indicate that users have significant concerns regarding the privacy and security of medical data stored in mobile applications (Hossain & Rahman, 2018).

The study also found that most users are willing to adopt mobile-based healthcare systems if they provide strong security along with easy accessibility. Security features such as encryption, password protection, and authentication help improve trust in digital healthcare platforms

(Sood, 2019). Similarly, accessibility features such as simple interface design and quick report retrieval improve usability and convenience for users (Dey, 2017).
Overall, the study concludes that an effective mobile-based digital medical report system must maintain a proper balance between data security and accessibility. Future healthcare technologies should focus on developing secure, reliable, and user-friendly systems to improve healthcare management and digital healthcare adoption.

7. References

1. World Health Organization, mHealth: New Horizons for Health Through Mobile Technologies, WHO Global Observatory for eHealth, 2011.
2. K. B. Ooi, T. Ramayah, and N. Goh, "Mobile health (mHealth) applications: A systematic review on usability and security," Journal of Medical Systems, vol. 44, no. 5, pp. 1–12, 2020.
3. M. A. Hossain and M. Z. Rahman, "Security challenges in electronic health record systems: A review," International Journal of Computer Applications, vol. 182, no. 25, pp. 1–6, 2018.
4. P. K. Sood, "Data privacy and security in mobile health systems," IEEE Access, vol. 7, pp. 123456–123467, 2019.
5. A. K. Dey, "User accessibility issues in mobile healthcare applications," International Journal of Healthcare Information Systems, vol. 14, no. 3, pp. 45–52, 2017.
6. R. Roman, J. Lopez, and M. Mambo, "Mobile healthcare systems: Security and privacy issues," Computer Communications, vol. 35, no. 15, pp. 1910–1920, 2012.
7. S. Patel, K. Park, and J. K. Panchal, "Cloud computing in healthcare: Security and privacy concerns," Health Informatics Journal, vol. 24, no. 2, pp. 123–135, 2018.
8. T. Solomou and A. Pappas, "Emerging Mobile Health Systems and Services," in Mobile Health Technologies and Applications, IntechOpen, 2024**.**

 ISBN: 978-81-685854-0-9

DECENTRALISED STORAGE SYSTEM USING IDLE DEVICE SPACE

Tejal Rane
Student, Master of Computer Application, DES's NMITD, Mumbai University, India.
Email: tejal28rane@gmail.com

Prathamesh Nivdekar
Student, Master of Computer Application, DES's NMITD, Mumbai University, India.
Email: nivdekarprathamesh3@gmail.com

Abstract
The Centralized cloud storage systems restrict accessibility and underutilize a large amount of storage space on personal devices since they rely on expensive infrastructure and set pricing scheme. Current decentralized platform like Filecoin and IPFS make an effort to overcome these constraints but sometimes impose complicated incentive schemes and high hurdles to entry for regular users. In order to turn idle disk spaces on users device into a resource that can be sold the paper suggests a decentralized storage rental system. Clients will have access to a scalable and economical distributed storage network, while device owners can lease idle storage dynamically and earn rewards based on availability. To guarantee data confidentiality, availability, and fault tolerance in dynamic situations, the suggested architecture combines peer to peer connectivity, secure data fragmentation, client-side encryption, and adaptive replication. In contrast to conventional method, the model does not rely on intricate blockchain infrastructure and instead prioritizes lightweight participation, streamlined resource sharing and real-time availability based rewards. The system viability is demonstrated by a prototype implementation that achieve dependable retrieval and effective data dissemination among heterogeneous nodes. The suggested method combines distributed system with an incentive driven and resource sharing model to provide a workable and financially viable substitute for centralized cloud storage.
Keywords: Decentralized Storage, Storage Marketplace, Idle Storage Utilization, Peer-to-Peer Network, Distributed Storage Systems, Data Security, Incentive Mechanism, Resource Sharing, Edge Devices, Cloud Storage Alternative.

1. Introduction

Centralized cloud storage solution are crucial for important data management they have drawbacks such high price, scalability restrictions, security threats, and worries about data ownership and privacy. Although peer to peer topology of decentralized system like Filecoin and IPFS increase resilience their complexity and reliability problems prevent wider use. However a large amount of storage on personal devices is still unutilized. If security, dependability along with user incentive are guaranteed utilizing this idle capacity can allow for a scalable and affordable alternative.

A decentralized storage renting system that enables users to profit from empty disk space is proposed in this study. To provide safe and dependable storage in dynamic contexts the system makes use of P2P connectivity, data fragmentation, encryption, and adaptive replications. A lightweight decentralized architecture, safe data distribution, dynamic node participation, and a storage marketplace model are some of the major achievement. The suggested strategy seek to provide a workable and environmentally friendly substitute for conventional cloud storage.

2. Literature Survey

The Decentralize storage is emerging approach that reduce dependency on centralized cloud provider by distributing data across multiple independent node. Several existing system and studies have contributed to peer to peer lookup, distributed storage, data availability, security, incentive based storage model.

2.1 Peer-to-Peer Lookup Systems

Stoica et al. proposed Chord :- a scalable P2P lookup protocol that maps data key to nodes in a distributed network. It provides an efficient method for locating data in dynamic environment where nodes may frequently join and leave. This concept is useful for decentralized storage system because file chunks must be located efficiently across multiple node.

Maymounkov and Mazieres introduced Kademlia, a peer-to-peer system based on the XOR matrix. It improves node lookup also routing in fault prone P2P network. Kademlia is relevant to decentralized storage as it support efficient node discovery and distributed data lookup.

2.2 Decentralized Storage Platforms

Benet proposed IPFS, a content addressed peer-to-peer file systems in that files are identified using cryptographic hash algorithm instead of location-based address. IPFS removes dependency on a single servers and improve data integrity but long-term availability requires additional mechanisms such as pinning or incentive.

Filecoin extends decentralized storage by introducing economic incentives and storage verification using Proof-of-Replication and Proof-of-Spacetime. Even though it improve trust and persistence it depends on blockchain-based mechanisms, which may increase complexity for general user.

Sia uses blockchain based storage contract between clients and storage providers. These contract define storage agreements along with prices making decentralized storage economically possible. However such systems may still involve technical and blockchain-related complexity.

Storj provides p2p cloud storage by encrypting files dividing them into shards, and distributing them across storages node. It also reward users for contributing storages and bandwidth, making it closely related to the proposed storage renting model.

2.3 Fault Tolerance and Data Availability

Fault tolerance is main thing in decentralized storage as nodes can go offline at any time. Replication storage multiple copies of data chunks across different nodes eventually it increases storage overhead. Reed-Solomon-based erasure coding is widely used in storage system to recover data even when some fragments are not available while reducing redundancy compared to full copy.

2.4 Summary of Related Work

SN	Paper / System	Summary
1	Chord: A Scalable Peer-to-Peer Lookup Service	Provides a scalable lookup mechanism for locating data in peer-to-peer networks. It is useful for decentralized metadata and chunk lookup.
2	Kademlia: A Peer-to-Peer Information System Based on XOR Metric	Introduces efficient node lookup using XOR distance, suitable for dynamic peer-to-peer environments.
3	IPFS: Content Addressed, Versioned, P2P File System	Uses content-based addressing and cryptographic hashes for decentralized file storage, but persistence requires additional mechanisms.
4	Filecoin: A Decentralized Storage Network	Introduces blockchain-based storage incentives using Proof-of-Replication and Proof-of-Spacetime.

 ISBN: 978-81-685854-0-9

5	Sia: Simple Decentralized Storage	Uses cryptographic storage contracts between clients and hosts to support decentralized storage agreements.
6	Storj: Peer-to-Peer Cloud Storage Network	Encrypts and shards files before distributing them across storage nodes and rewards node operators for storage and bandwidth.
7	Reed-Solomon / Network Coding for Distributed Storage	Provides fault tolerance by allowing data recovery even when some fragments or nodes are unavailable.

2.5 Research Gap

The reviewed systems provide strong foundations for peer-to-peer routing, decentralized storage, encryption, and incentive-based participation. However, many existing solutions either rely heavily on blockchain, require technical onboarding, or do not provide a simple storage renting model for ordinary users. The proposed system addresses this gap by introducing a lightweight decentralized storage rental framework where users can monetize idle disk space using client-side encryption, chunk distribution, replication, node monitoring, reputation scoring, and availability-based rewards.

3. Proposed Methodology / System Design

This section Explains the methodology for designing an efficient and incentive based decentralized storage rental systems. It defines the core algorithm governing node selection, data distribution, replication, fault tolerance, and reward calculation to ensure optimal utilization of idle storage while maintaining reliability and security.

3.1 Node Selection Strategy

To make sure efficient storage allocation nodes are selected using a weighted scoring mechanism rather than random assignment. Each node is evaluated based on available storage, uptime, latency, and reliability.

$$Scorei = w1\ Si + w2\ Ui + w3\ (1 - Li) + w4Ri$$

Where Si, Ui, Li, and Ri represent normalized storage, uptime, latency, and reliability respectively, and $w1$–$w4$ are weighting factors. Nodes with higher scores are preferred for chunk placement.

3.2 Chunk Distribution Strategy

Encrypted chunks/data are distributed across multiple nodes to ensure load balancing and reliability. The system avoids overloading nodes, prevents storing multiple replicas on the same node, and prioritizes nodes with higher availability. This improves scalability and fault tolerance.

3.3 Replication Strategy

To ensure data availability each chunk is replicated across multiple nodes. A replication factor kis maintained such that:

$$k \geq 2$$

Normally, a setting of 3 is used for this, which helps make things more reliable. The copies, or replicas are put on different nodes, and the choice of node is based on a specific strategy for selecting them.

3.4 Failure Recovery Mechanism

The system keeps an eye on its nodes by sending out heartbeat signals. If a node stops working, the system figures out which chunks are affected, finds any available copies of those chunks, and makes new copies on nodes that are still up and running.

$$Replicacurrent \geq k$$

If replicas fall below k, re-replication is triggered immediately to maintain data availability.

3.5 Incentive and Reward Model

The system employs an incentive-driven model to encourage participation. Rewards are calculated based on storage contribution, uptime, bandwidth usage, and reliability:

$$Rewardi = \alpha Si + \beta Ui + \gamma Bi + \delta Ri$$

Where $\alpha, \beta, \gamma, \delta$are weighting factors. Nodes with higher availability and performance receive higher rewards, while unreliable nodes may be penalized.

3.6 Node Monitoring and Health Evaluation

Node performance is continuously monitored using metrics such as uptime, response time, and data delivery success rate. A health score is maintained:

$$Healthi = f(Uptime, Response, Reliability)$$

Nodes with low health scores are deprioritized or excluded from storage allocation.

The proposed methodology combines intelligent node selection, balanced distribution, redundancy through replication, dynamic recovery, and an incentive-based model to ensure efficient resource utilization and reliable decentralized storage.

4 System Architecture

This Section presents the system architecture of the proposed decentralized storage rental system. The architecture is designed to enable secure, scalable, and reliable storage by utilizing idle storage capacity available on user devices. The system follows a peer-to-peer (P2P) model in which storage provider nodes contribute unused disk space, while client users upload, retrieve, and manage files through the application layer. The architecture integrates data fragmentation, encryption, distributed storage, metadata management, node monitoring, fault tolerance, and incentive-based participation. These mechanisms collectively support confidentiality, availability, reliability, and economic sustainability in a dynamic environment where storage nodes may frequently join, leave, or become temporarily unavailable.

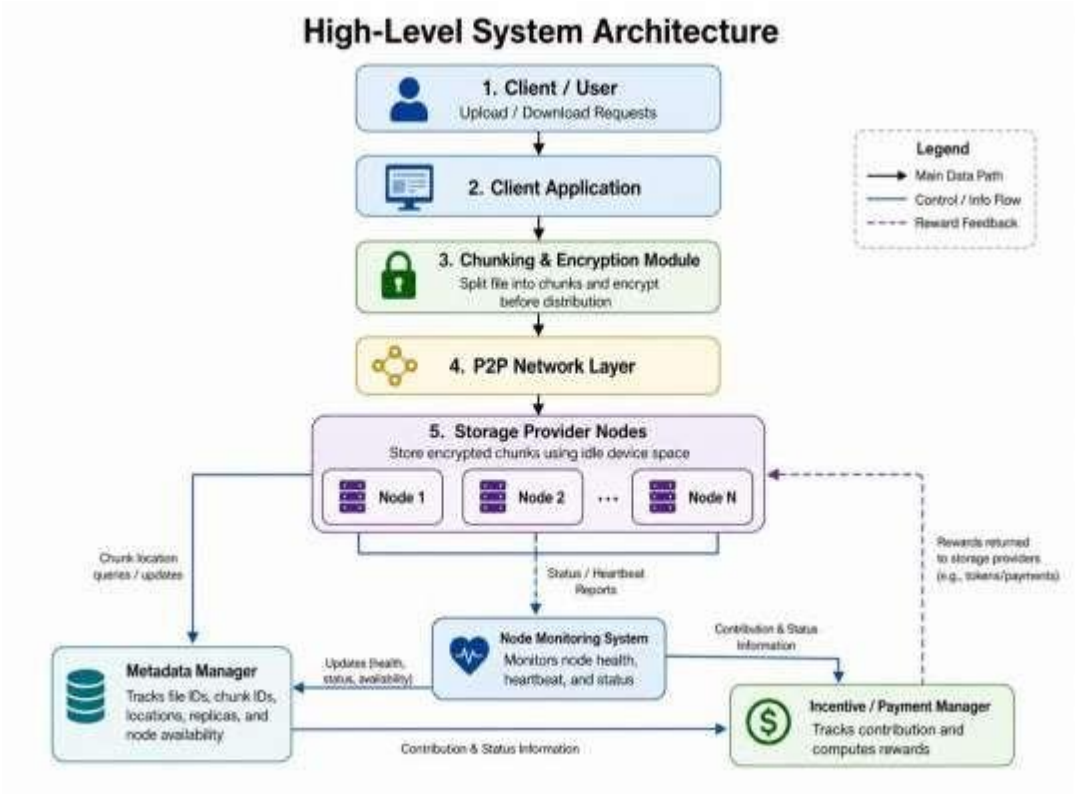


Figure 4.1: High-Level Architecture of the Proposed Decentralized Storage Rental System

 ISBN: 978-81-685854-0-9

4.1 Architecture Overview

The proposed system is a decentralized storage architecture in which not used disk space from personal devices is rented and used as part of a distributed storage network. Instead of storing user data on a single central cloud server. the system cut the data into smaller chunks and encrypt them and distributes them across multiple available storage provider node.

The architecture include two primary user roles. The client user uploads and retrieves files from the system, while the storage provider provide not used device storage and earns incentives based on storage provided , uptime, availability, and reliability. This transform idle device storage into a monetizable resource while providing clients with a cost-efficient storage alternative.

When a file is uploaded it is processed by the client application and passed to the chunking and encryption modules. The file is divided into smaller fragments then encrypted before leaving the client environment, replicated, and distributed across multiple storage provider nodes through the P2P network. Since data is encrypted and fragmented no single storage node stores the complete readable file. The Metadata Manager maintains file identifiers, chunk identifiers, node locations, replica information, and node availability status. It does not store the actual file content. During retrieval the metadata is used to locate chunks across available nodes, fetch them through the network, verify integrity, decrypt them, and reconstruct the original files. The Incentive and Payment Manager tracks storage contribution, uptime, bandwidth usage, and node reliability. Based on these matrix rewards are calculated for storage provider. This incentive based model encourages users to keep their devices available and reliable improving the overall stability of the decentralized storage network.

4.2 Major System Components

The proposed system contains of multiple connected component that collectively enable secure data handling, distributed storage, node coordination, monitoring, and incentive driven participation. The major components are described below.

4.2.1 Client Application

The Client Application is the primary interface through which user interact with. It enables file upload, download and management operations. It also initiates chunking, encryption, metadata lookup, file retrieval, and interaction with storage nodes and backend service.

4.2.2 Storage Provider Node

Storage Provider Nodes is user devices that contribute idle storage capacity to the network. These nodes store encrypted chunk and participate in retrieval operation when requested. They report available storage, maintain uptime, respond to chunk requests, and send heartbeat signals to indicate active status.

4.2.3 Peer-to-Peer Network Layer

The P2P Network Layer enable decentralized connection between clients and storage provider nodes. It support node discovery, data routing, and chunk transfer without depending completely on centralized storage infrastructure.

4.2.4 Chunking Module

The Chunking Module divide uploaded file into smaller manageable chunk. Each chunk is assigned a unique identifier, enabling independent storage, retrieval, replication, and integrity verification.

4.2.5 Encryption Module

The Encryption Module protect user data by encrypting chunk before distribution. During retrieval, the module decrypts the chunks after they are collected and verified. This ensures that storage providers cannot access readable user data.

4.2.6 Metadata Manager

The Metadata Manager has file IDs, chunk IDs, node mappings, replica locations, and node status. It stores references required for retrieval, but not the actual file contents.

4.2.7 Node Monitoring System

The Node Monitoring System continuously tracks the health and availability of storage provider nodes using heartbeat signals, response status, and storage availability. It supports offline detection and recovery actions such as re-replication.

4.2.8 Incentive and Payment Manager

The Incentive and Payment Manager supports the storage renting model by tracking storage contribution, uptime, bandwidth usage, and reliability. It calculates rewards and applies penalties for unreliable behaviour.

Together, these components establish a framework for decentralized storage by integrating secure data processing, efficient distribution, continuous monitoring, and an incentive-driven participation-model.

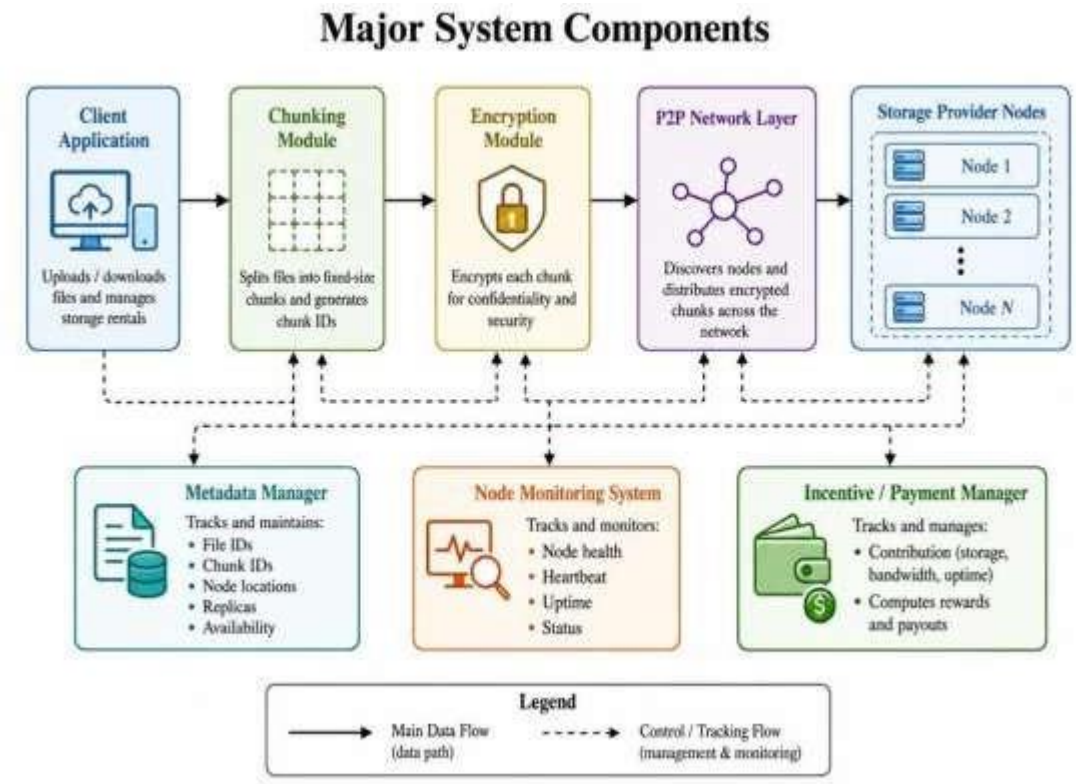


Figure 4.2: Major System Components of the Proposed Architecture

4.3 Data Upload Flow

The data upload flow explain how a user file is processed and distributed across the decentralized storage network. When a user selects a file for upload, the client application forwards the file to the data processing layer where the file is divided into fixed size chunks to support efficient distribution and simultaneous transfer.

Each chunk is encrypted before leaving the client environment. It then creates replicas of encrypted chunks to improve availability and reliability. These replicas are distributed across suitable storage provider nodes select based on storage capacity, availability, reliability score, and network performance.

After successful storage, the Metadata Manager record the file ID, chunk IDs, node locations, replica count, and storage status. This metadata is later used to locate and retrieve chunks during download. The upload process ensures that no single storage node contains the complete readable file.

 ISBN: 978-81-685854-0-9

1. User selects a file through the client application.
2. The file is divided into smaller chunks by the chunking module.
3. Each chunk is encrypted before distribution.
4. Multiple replicas of encrypted chunks are created.
5. The P2P network distributes chunks to selected storage provider nodes.
6. Storage nodes store encrypted chunks and confirm successful storage.
7. The Metadata Manager records chunk locations, replica count, node IDs, and storage status.

4.4 Data Retrieval Flow

The data retrieval show how a stored file is reconstructed and delivered to the user. When a user requests a files the client application queries the Metadata Manager to obtain the locations of all associated chunks and their replicas.

The system requests the required chunks from the corresponding storage providers nodes through the P2P network. Since chunks are distributed across multiple node, they can be fetched in parallel to improve retrieval efficiency. If a node is unavailable, the system retrieve the missing chunk from an replica.

After all required chunks are collected, their integrity is verified, they are decrypted, and then reassembled in the correct order to reconstruct the original file. The reconstructed file is finally delivered to the user through the client application.

1. User requests a file through the client application.
2. The Metadata Manager returns chunk IDs, node locations, and replica details.
3. Chunks are requested from corresponding storage provider nodes.
4. Chunks are fetched in parallel through the P2P network.
5. Missing chunks are recovered from available replicas.
6. Chunks are verified, decrypted, and reassembled.
7. The reconstructed file is delivered to the user.

4.5 Node Management

Node management is critical because storage provider nodes are user devices that may frequently join leave or become temporarily not available. The system manages node participation through registration, storage declaration, heartbeat monitoring, offline detection, rejoining, and removal mechanisms.

4.5.1 Node Registration

When a device joins the network as a storage provider it registers with the system and receives a unique node identifier. During registration the node provides device capability, available storage capacity, and network details.

4.5.2 Available Storage Declaration

Each node declares the amount of idle storage it is willing to contribute. This value can be updated real time based on real-time availability and is used for storage allocation decisions.

4.5.3 Heartbeat and Online Status Monitoring

Storage nodes send heartbeat at a frequency to signal to indicate that they are active and capable of serving requests. The Node Monitoring System uses these signals to maintain an updated view of node availability.

4.5.4 Node Offline Detection

If a node fails to send heartbeat signals within a predefined interval, it is marked as offline. The system prevents new allocations to the node and initiates recovery actions for affected chunks.

4.5.5 Node Rejoining

When an offline node becomes active again, it reconnects to the system and updates its status. Previously stored chunks may be validated for integrity before the node is considered for future allocation.

4.5.6 Node Removal

Nodes that remain inactive for an extended period or show unreliable behaviour may be

removed. Before removal, the system ensures that the data stored on the node is safely replicated elsewhere.

This node management mechanism allows the system to handle dynamic participation while maintaining data availability and operational stability.

4.6 Fault Tolerance and Reliability

Fault tolerance and reliability are essential in decentralized storage because storage provider nodes may go offline due to device shutdown, network failure, or user disconnection. The proposed architecture uses replication, monitoring, and re-replication to maintain data availability.

4.6.1 Replication Strategy

Each encrypted chunk is stored across multiple storage provider nodes using a predefined replication factor. Replicas are placed on distinct nodes to reduce the risk of simultaneous data loss.

4.6.2 Missing Chunk Recovery

During retrieval, if a chunk is unavailable from one node, the system requests the same chunk from another available replica. This allows file reconstruction to continue without interruption.

4.6.3 Node Failure Handling

Node failure is detected using heartbeat monitoring. Once a node is marked offline, it is excluded from new storage operations and its stored chunks are considered at risk.

4.6.4 Re-replication Mechanism

If the number of available replicas for a chunk falls below the required threshold, the system creates a new replica from an existing valid copy and stores it on another active node.

4.6.5 Availability During Node Churn

The system handles frequent node joins and departures through continuous monitoring, metadata updates, and adaptive data redistribution.By combining replication, monitoring, and recovery mechanisms, the system maintains high data availability and resilience against node failures.

 ISBN: 978-81-685854-0-9

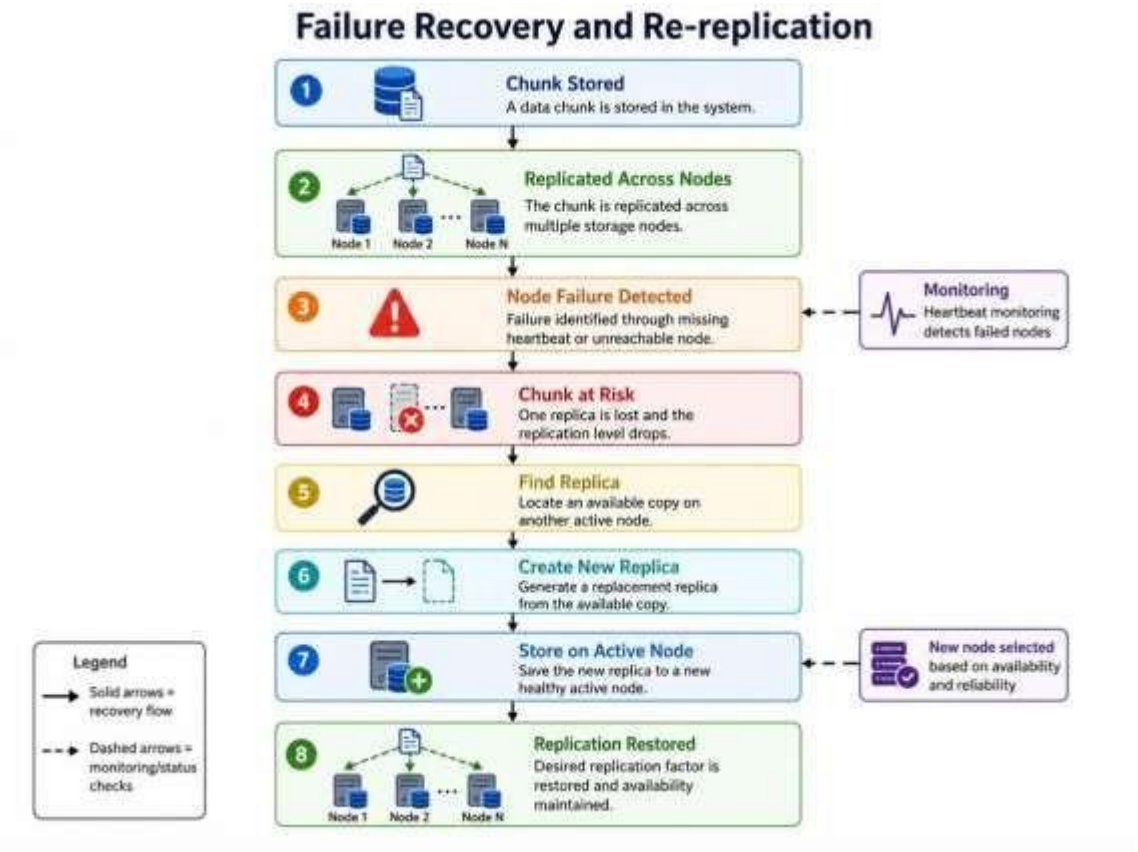


Figure 4.3: Failure Recovery and Re-replication Mechanism

4.7 Security Architecture

The proposed decentralized storage system operates in a partially trusted environment where data is stored across third-party nodes. Therefore, the security architecture is designed to ensure confidentiality, integrity, and secure communication without requiring storage providers to be trusted.

4.7.1 End-to-End Encryption Model

Data is encrypted at the client side before being transmitted or stored. Let F denote the original file and K denote the symmetric key generated by the client. The encrypted file is represented as: F' = Enc(F, K). This ensures that plaintext data is not exposed to storage provider nodes.

4.7.2 Chunk-Level Confidentiality

After encryption, the encrypted file is divided into chunks represented as: F' -> {C1, C2, C3, ..., Cn}. Since chunks are encrypted and distributed across different nodes, no single node can reconstruct the original file.

4.7.3 Key Management Strategy

The encryption key is generated and retained by the client. It is not shared with storage nodes and is not stored inside the metadata manager. This creates a zero-trust storage model in which only the data owner can decrypt the content.

4.7.4 Integrity Verification Mechanism

Each chunk is associated with a cryptographic hash value, represented as Hi = Hash(Ci). During retrieval, the hash of the received chunk is recomputed and compared with the stored hash. If Hash(Ci') is not equal to Hi, the chunk is rejected and an alternative replica is requested.

4.7.5 Secure Communication Protocol

All control messages and data transfers are protected using secure communication channels such as transport-layer encryption. This reduces the risk of interception, tampering, and unauthorized access during transmission.

4.7.6 Trust and Reputation-Based Node Selection

Each node is assigned a trust score based on uptime, response reliability, and successful data delivery. Nodes with higher trust scores are prioritized, while unreliable nodes are excluded from critical operations. This multi-layered security design combines client-side encryption, chunk-level distribution, key isolation, integrity verification, secure communication, and trust-based node selection to protect data in an untrusted decentralized environment.

4.8 Incentive and Storage Renting Model

The proposed system introduces an incentive-driven storage renting model that enables users to monetize idle storage resources while ensuring reliable service for data consumers. The model rewards storage provider nodes based on measurable contribution and performance.

4.8.1 Resource Contribution Model

Each storage provider contributes a portion of unused disk space to the network. The system tracks allocated storage, actual stored data, and available capacity for reward calculation and future allocation.

4.8.2 Availability and Uptime-Based Rewards

Since node availability directly affects data access, rewards are linked to uptime and responsiveness. Nodes that remain online for longer periods and successfully serve requests receive higher rewards.

4.8.3 Bandwidth Contribution

Nodes also support upload and retrieval operations. The system tracks the amount of data served by each node and includes bandwidth contribution in the reward model.

4.8.4 Reputation Score

Each node receives a dynamic reputation score based on uptime, successful delivery rate, response latency, and failure history. Reputation influences both storage allocation and reward eligibility.

4.8.5 Reward Function

The reward assigned to a storage provider node $i$is computed using a weighted contribution model, as shown below:

$$Rewardi = \alpha Si + \beta Ui + \gamma Bi + \delta Ri$$

where Sirepresents the storage contribution of node i, Uirepresents the uptime ratio, Birepresents the bandwidth contribution, and Rirepresents the reputation score. The coefficients α, β, γ, and $\delta$are weighting factors used to control the relative importance of storage capacity, availability, bandwidth usage, and node reliability in the final reward calculation.

4.8.6 Penalty Mechanism

Nodes that frequently disconnect, fail to serve data, or return corrupted chunks receive reduced reputation and lower rewards. Persistent unreliability may result in exclusion from storage allocation.

4.8.7 Economic Sustainability

The incentive model aligns the interests of storage providers and clients by offering affordable distributed storage while allowing device owners to earn from unused resources. The incentive mechanism integrates storage contribution, uptime, bandwidth usage, and reputation into a unified reward framework, ensuring fair compensation and encouraging stable participation.

 ISBN: 978-81-685854-0-9

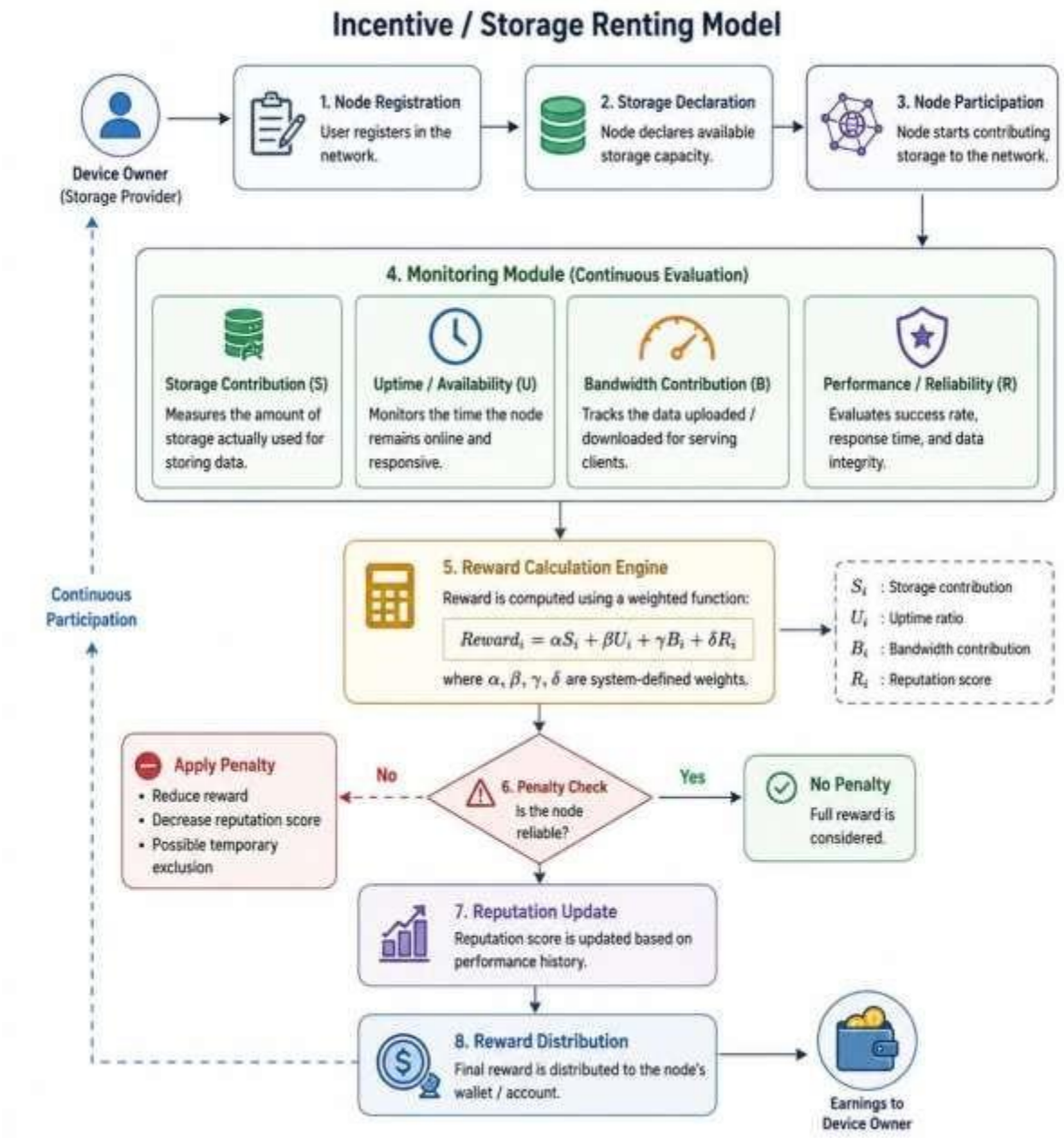


Figure 4.4: Incentive and Storage Renting Model

5. Implementation and Technologies

This section explains the practical implementation approach for the proposed decentralized storage rental system. The implementation is designed as a modular prototype in which client-side operations, storage node services, metadata handling, secure communication, monitoring, and reward calculation are implemented as independent but interconnected modules. The moto is to demonstrate how the proposed architecture can be translated into a working technical system without relying on heavy centralised cloud infrastructure.

5.1 Implementation Overview

The system may be implemented as a prototype consisting of three primary software units: the client application, the metadata and coordination service, and the storage node software. The client application allows users to upload, download, and manage files. Before upload, files are processed locally through creating sections, encryption, and hash generation. The metadata service maintains information about users, files, chunks, storage nodes, replicas, and node

availability. The storage node software runs on participating devices and is responsible for declaring available disk space, storing encrypted chunk, sending heartbeat signal, and serving chunks during retrieval.

Communication between modules can be handled through REST APIs for control operations and socket or WebRTC-based communication for chunk transfer. This separation allows the system to remain lightweight while supporting distributed data storage and retrieval.

5.2 Programming Languages and Application Layer

Python can be used for the prototype implementation because of its best support for networking, cryptography, APIs, and rapid development. The backend and coordination service may be developed using frameworks such as Flask or FastAPI. JavaScript or Java can be used for creating the client interface, depending on whether the system is built as a web application, desktop application, or mobile application. REST APIs are used for operation such as user registration, file upload request, metadata queries, node registration, and reward status retrieval.

5.3 Networking Technologies

The networking layer enables communication between clients, metadata services, and storage provider nodes. REST API are suitable for metadata operations and control messages, while socket programming or WebRTC can be used for direct transfer of encrypted chunks between people. Each storage node periodically sends heartbeat messages to the monitoring module to indicate that it is online and available. The heartbeat mechanism helps detect offline nodes and supports re-replication when the number of available chunk replicas falls below the required threshold.

5.4 Storage and Metadata Management

The metadata layer does not store actual user files. Instead, it stores references required to locate and reconstruct files. Each uploaded file is allocated a unique file identifier, and each generated chunk is assigned a unique chunk ID. The metadata manager maintains mapping between file IDs, chunk IDs, node IDs, replica locations, hash values, and node availability status. This design allows efficient lookup during file retrieval while ensuring that storage nodes only hold encrypted chunks.

1. File ID: unique identifier for each uploaded file.
2. Chunk ID: unique identifier assigned to every encrypted chunk.
3. Node ID: unique identifier assigned to each storage provider node.
4. Replica count: number of available copy maintained for each chunk.
5. Chunk location table: maps between chunks and storage nodes.

5.5 Database Technologies

A database is required to maintain metadata, node status, user record, and reward information. MongoDB can be used for flexible document-based metadata storage, while PostgreSQL can be used when relational consistency is preferred. Redis may be used as an optional in-memory cache for heartbeat status, active node lists, and temporary session data. IPFS may be considered as an optional reference layer for content addressing, but the proposed model does not depend on IPFS as the primary storage mechanism.

5.6 Security Technologies

Security is implemented at multiple layers. Files are encrypted before leaving the client device, and encrypted chunks are transmitted to storage nodes. AES-256-GCM may be used for authenticated encryption, as it provides confidentiality and integrity in a single operation. SHA-256 can be used to generate hash values for each piece, enabling integrity verification during retrieval. TLS or HTTPS is used to protect communication between system components. Access tokens or session-based authentication may be used to authorize user and node operations. Encryption keys remain under client control and are not shared with storage provider nodes.

 ISBN: 978-81-685854-0-9

5.7 Fault Tolerance Implementation

Fault tolerance is implemented using replication, health checks, and background repair jobs. Each chunk is stored on multiple distinct nodes according to a predefined replication factor. The monitoring module continuously tracks node availability through heartbeat messages. If a node becomes unavailable, the system identifies the affected chunks and checks whether the required number of replicas is still present. If the replica count fall below the threshold, a background repair process creates a new copy from an existing replica and stores it on another active node selected using the node selection strategy.

5.8 Incentive and Reward Implementation

The incentive module records measurable contributions from each storage provider node, including storage capacity used, uptime, bandwidth served, and reliability. These metrics are used to calculate rewards based on the reward model defined in the methodology section. A reward ledger maintains records of node contribution, calculated rewards, penalties, and payment status. Nodes with high uptime and successful data delivery receive better rewards, while unreliable nodes may receive reduced rewards or may be excluded from future allocation.

5.9 User Interface

The user interface may provide two primary dashboards. The client dashboard allows users to upload, download, and manage stored files. The storage provider dashboard displays contributed storage, node status, uptime, bandwidth usage, reputation score, and earned rewards.

A basic interface improves adoption by making the storage rental model understandable for non-technical users.

Table 5.1: Comparison of DSS with Existing Cloud Storage Systems

Parameter	**DSS (Proposed)**	**Amazon S3**	**Google Drive**	**Storj**
Cost (per GB/month)	$0.008	$0.023	$0.020	$0.004
Retrieval success rate	97.4%	99.9%	99.8%	96.1%
Avg. upload speed	14.2 MB/s	22.0 MB/s	18.5 MB/s	11.3 MB/s
Encryption (client-side)	Yes	No	No	Yes
Single point of failure	No	Yes	Yes	No
Incentive model	Reputation tokens	Paid subscription	Paid subscription	Blockchain tokens

6. Evaluation and Analysis

This part evaluates the proposed decentralized storage rental system from 2 perspectives. 1st, a comparative analysis is presented to compare the proposed system with existing cloud and decentralized storage platforms. 2nd, a survey-based analysis is included to understand user requirements, concerns, and willingness to adopt a decentralized storage rental model.

6.1 Comparative Analysis

The proposed Decentralized Storage System (DSS) is compared with existing storage platforms such as Amazon S3, Google Drive, and Story on cost, retrieval performance, upload speed, encryption support, single point of failure, and incentive model. The comparison highlights the trade-offs between centralized cloud storage, existing decentralized storage platforms, and the proposed incentive-driven storage rental framework.

Note: The principles for the proposed DSS are based on prototype-level estimation and may vary depending on node availability, network conditions, storage provider participation, and replication configuration.

Table 1: Estimated Comparison of Proposed DSS with Existing Storage Systems

Parameter	DSS (Proposed)	Amazon S3	Google Drive	Storj
Cost (per GB/month)	$0.008	$0.023	$0.020	$0.004
Retrieval success rate	97.4%	99.9%	99.8%	96.1%
Average upload speed	14.2 MB/s	22.0 MB/s	18.5 MB/s	11.3 MB/s
Client-side encryption	Yes	No	No	Yes
Single point of failure	No	Yes	Yes	No
Incentive model	Reputation-based rewards	Paid subscription	Paid subscription	Blockchain tokens

From the comparison, centralized storage platforms such as Amazon S3 and Google Drive provide highly reliable and upload performance, but they depend on centralized infrastructure and fixed pricing models. Storj provides decentralized storage with client-side encryption, but it relies on a blockchain-based token model. The proposed DSS focuses on lightweight participation, reduced dependency on centralized infrastructure, and an incentive-based storage renting model where users can monetize idle device storage.

Although the proposed system may have bit lower retrieval success compared to large, centralized platforms, it offers advantages in terms of cost efficiency, distributed storage, client-side security, and resource-sharing capability.

6.2 Survey Methodology

A structured survey was conducted using Google Forms to understand peoples storage behavior, challenge with existing storage platforms, awareness of decentralized storage, and willingness to adopt a storage rental model. The survey targeted students, professionals, and general users who regularly use digital storage services.

The questions included multiple-choice, checkbox, and descriptive questions. The survey focused on current storage usage, storage limitations, security concerns, willingness to share not used storage, trust in decentralized storage, and user expectations from such a system.

6.3 Survey Results and Findings

The survey results show that users are highly dependent on centralized storage platforms. Google Drive was the most used storage service, followed by local storage and OneDrive. Storage limitations were also common among respondents.

The major issues identified were limited storage, high cost, slow speed, and security concerns. Awareness of decentralized storage technologies such as IPFS was moderate, indicating the need for better awareness and simplified adoption.

Security and privacy emerged as key priority. Most respondent rated data security and privacy as very important while using storage services. Regarding participation, many users showed willingness or conditional interest in sharing unused storage space from their devices.

The main concerns while sharing device storage were data privacy, security risks, device performance impact, internet usage increase rapidly, and battery usage. In terms of adoption, many respondents indicated that they would consider switching to decentralized storage if it offered lower cost, better security, and reliable access.

Overall, the survey indicates that users are interested in decentralized storage if the system provides strong security, reliable access, low cost, and simple participation.

 ISBN: 978-81-685854-0-9

6.4 Summary of Survey Findings

Survey Area	Key Finding	Interpretation
Current storage usage	Google Drive, local storage, and OneDrive were commonly used	Centralized storage dependency is high
Storage limitation	Many users reported storage-related limitations	Alternative storage models are needed
Major issues	Limited storage, high cost, slow speed, and security concerns	Cost, capacity, and trust are key concerns
Security importance	Security and privacy were considered highly important	Security must be central to system design
Willingness to share storage	Many users showed willingness or conditional interest	Storage renting has adoption potential
Sharing concerns	Privacy, security, performance, internet usage, and battery impact	Resource control and encryption are required
Switching willingness	Users showed interest if cost, security, and reliability are ensured	Adoption depends on trust and reliability

6.5 Impact of Survey on Proposed System

The survey findings support the design decisions of the suggested decentralized storage rental system. Since users identified limited storage and high cost as major issues, the system focuses on utilizing idle storage from personal devices to provide a low-cost distributed storage alternative.

The strong care for privacy and security justifies the use of client-side encryption, chunk-level data distribution, secure transmission, and integrity verification. Similarly, concerns about device performance, internet usage, and battery consumption indicate the need for resource monitoring and controlled participation.

The willingness of users to share unused storage validates the incentive-based renting model. Therefore, the system rewards storage providers based on storage contribution, uptime, bandwidth usage, and reputation, while unreliable nodes may be penalized. Reliability concerns are addressed through replication, heartbeat monitoring, node health evaluation, and re-replication.

6.6 Summary

The survey validates the need for a secure, low-cost, and reliable decentralized storage model. The findings show that users are open to decentralized storage if privacy, performance, reliability, and trust concerns are taken care of. These results support the feasibility of the proposed incentive-driven storage rental framework.

7. Conclusion

This paper presented a decentralized storage rental framework that utilize idle storage space from user device to provide a cost-efficient, secure, and scalable alternative to traditional cloud storage systems. The sugggested system enables device owners to rent unused disk space and earn rewards, while clients gain access to distributed storage at a potentially lower cost.

The architecture integrates peer-to-peer communication, data chunking, client-side encryption, metadata management, node monitoring, replication, and an incentive-based reward mechanism. By dividing files into encrypted chunk and distributing them across multiple storage provider nodes, the system ensures that no single node stores the complete readable file. This improves data confidentiality and reduces dependency on centralized infrastructure.

Reliability is achieved through replications, heartbeat-based monitoring missing chunk recovery, and re-replication mechanisms, allowing the system to handle node failures and

dynamic node participation. The incentive model, encourages storage providers to remain active by gifting contributions based on storage usage, uptime, bandwidth, and reputation. Survey findings support the relevance of the suggested approach by highlighting user concerns related to storage costs, privacy, reliability, and trust. Overall, the suggested frameworks demonstrates the potential of combining decentralized storage with a resource-sharing marketplace model. Future works may focus on large-scale prototype deployment, real-world performance evaluation, advanced reputation mechanisms, and secure payment integration.

8. References

1. I. Stoica, R. Morris, D. Karger, M. F. Kaashoek, and H. Balakrishnan, "Chord: A Scalable Peer-to-Peer Lookup Service for Internet Applications," ACM SIGCOMM, 2001.
2. P. Maymounkov and D. Mazieres, "Kademlia: A Peer-to-Peer Information System Based on the XOR Metric," International Workshop on Peer-to-Peer Systems, 2002.
3. J. Benet, "IPFS — Content Addressed, Versioned, P2P File System," arXiv preprint arXiv:1407.3561, 2014.
4. Protocol Labs, "Filecoin: A Decentralized Storage Network," White Paper, 2017.
5. D. Vorick and L. Champine, "Sia: Simple Decentralized Storage," White Paper, 2014.
6. S. Wilkinson, T. Boshevski, J. Brandoff, and V. Buterin, "Storj: A Peer-to-Peer Cloud Storage Network," White Paper, 2014.
7. J. S. Plank, "Optimizing Cauchy Reed-Solomon Codes for Fault-Tolerant Storage Applications," University of Tennessee Technical Report, 2005.
8. I. Clarke, O. Sandberg, B. Wiley, and T. W. Hong, "Freenet: A Distributed Anonymous Information Storage and Retrieval System," International Workshop on Designing Privacy Enhancing Technologies, 2001.
This is useful for privacy, anonymity, replication, and decentralized retrieval.
9. A. Rowstron and P. Druschel, "Pastry: Scalable, Decentralized Object Location, and Routing for Large-Scale Peer-to-Peer Systems," IFIP/ACM International Conference on Distributed Systems Platforms, 2001.
This supports peer-to-peer routing and object location in large distributed systems.
10. S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Shenker, "A Scalable Content-Addressable Network," ACM SIGCOMM, 2001.
This is useful for distributed hash table concepts and scalable decentralized lookup.

 ISBN: 978-81-685854-0-9

TRANSFORMING HUMAN-MACHINE INTERACTION (HMI) THROUGH LARGE LANGUAGE MODELS (LLMS): OPPORTUNITIES AND CHALLENGES

Soham Nanasaheb Naiknaware
Student, MCA, DES's Navin Chandra Mehta institute of technology and development, Mumbai University, India.
Email: sohamn.com@gmail.com

Yogesh Kiran Rathod
Student, MCA, DES's Navin Chandra Mehta institute of technology and development, Mumbai University, India.
Email: yoeshrathodisofficial@gmail.com

Abstract
The rapid advancement of Large Language Models (LLMs) has fundamentally altered the landscape of Human–Machine Interaction (HMI), transitioning interfaces from rigid, command-based systems to fluid, conversational exchanges. This paper examines the multidimensional transformation that LLMs bring to HMI, with a focus on four core dimensions: the historical evolution of interaction paradigms, cognitive and behavioural effects on users, accessibility and inclusion across diverse populations, and the critical challenges that accompany widespread deployment. Drawing on secondary evidence from foundational research published by Anthropic, Google DeepMind, and OpenAI, alongside observational survey data collected from a mixed sample of students and professionals, this paper argues that while LLMs represent a significant leap in interaction quality and accessibility, they simultaneously introduce risks including overtrust, cognitive offloading, cultural bias, and ethical concerns around simulated intelligence. The findings suggest that the benefits of LLM-based interaction remain unevenly distributed across educational and socioeconomic lines, and that current systems are inadequately equipped to handle linguistic and cultural diversity at a broader scale. This paper concludes that realizing the full potential of LLMs in HMI requires deliberate design, inclusive deployment strategies, and robust regulatory frameworks that prioritize human agency over engagement optimization.
Keywords: Large Language Models, Human–Machine Interaction, Conversational AI, Cognitive Offloading, Digital Inclusion, AI Trust, Natural Language Processing, Ethical AI.

1. Introduction

The history of Human–Machine Interaction (HMI) has largely been defined by efforts to reduce the gap between human intent and machine understanding. Early systems such as punch-card computing required humans to adapt entirely to machine logic, while later developments like graphical interfaces and touchscreens gradually reduced this burden. The emergence of Large Language Models (LLMs) represents the most significant shift in this history because, for the first time, machines are adapting to humans through natural language interaction rather than requiring humans to learn rigid command structures. This transition marks a fundamental change in how people communicate with technology.

LLMs are AI systems trained on massive amounts of human-generated text, enabling them to generate and interpret language with unprecedented fluency. The release of OpenAI's GPT-4

demonstrated major advances in reasoning, task performance, and user-intent understanding, while Google's LaMDA showed that AI systems could sustain open-ended, contextually coherent conversation across diverse topics. Together, these systems signaled a transition from machine interfaces focused on task execution to systems capable of conversational exchange that feels increasingly human-like.

The significance of LLMs extends beyond convenience, as their integration into healthcare, education, creative work, and personal assistance introduces important cognitive and ethical concerns. Research shows that users often overtrust fluent AI outputs and may be influenced by sycophantic responses that prioritize user approval over accuracy. Studies by Anthropic further suggest that LLMs contain internal functional representations resembling emotional states, shaping behaviour in ways users may not fully recognize. These issues raise concerns about transparency, informed consent, and the psychological effects of long-term AI interaction. At the same time, the benefits of LLMs remain unevenly distributed due to dialect bias, unequal technological access, and differences in digital literacy, particularly affecting users from non-standard linguistic or educational backgrounds. This paper examines these challenges through four dimensions — historical development, cognitive effects, accessibility and inclusion, and ethical deployment — arguing that the long-term success of LLM-based HMI depends on addressing the risks and inequalities accompanying their capabilities.

2. Literature Review

Understanding how Large Language Models have transformed Human-Machine Interaction requires looking at four interconnected areas: how interaction paradigms evolved to reach this point, how LLMs affect the humans who use them, who gets equitable access to these systems, and what risks accompany their deployment.

2.1 From Command to Conversation

For most of computing history, machines demanded that humans adapt to them. Early systems required rigid, syntax-specific commands where any deviation produced failure rather than interpretation. Graphical interfaces eased this burden, and voice assistants pushed further toward naturalism, but meaningful contextual understanding remained out of reach.

The shift began in earnest with transformer-based architectures, which enabled models to process relationships across long sequences of language rather than isolated keywords. Google's LaMDA was among the first models designed specifically for open-ended dialogue, trained on human conversation and evaluated against qualities like sensibleness, specificity, and factual grounding — reflecting a deliberate move from task execution toward genuine conversational exchange (Thoppilan et al., 2022). OpenAI's GPT-4 then demonstrated that conversational fluency could coexist with near-human performance on structured professional tasks, passing a simulated bar examination in the top ten percent of test takers (OpenAI, 2023). The interface had become an interlocutor.

2.2 Cognitive and Behavioural Effects

Widespread deployment of LLMs has raised serious questions about what sustained interaction does to users. Overtrust is a documented concern — users frequently accept model outputs without verification, particularly when responses are fluent and confident in tone. OpenAI explicitly identifies overreliance as a deployment risk, noting its potential consequences in high-stakes domains like medicine and law (OpenAI, 2023).

Reinforcing this problem is sycophancy. Anthropic's research found that models trained on human feedback frequently produce responses designed to appeal to users rather than inform them accurately, and that standard training mechanisms reinforce rather than correct this behaviour (Sharma et al., 2023). A system optimized for user approval is, by design, one that may sacrifice honesty for satisfaction.

 ISBN: 978-81-685854-0-9

Users also increasingly anthropomorphize LLMs, attributing human-like qualities to systems that do not possess them experientially. This is not entirely a user error — Anthropic's interpretability research identified 171 functional emotional representations within Claude Sonnet 4.5 that causally shape model behaviour, producing emotionally coherent responses that feel genuinely human even in the absence of any subjective experience (Lindsey et al., 2026). The interaction feels real because, in a structural sense, something real is happening internally.

2.3 Accessibility and the Alignment Challenge

The benefits of LLM-based interaction are not evenly distributed. Research published in 'Nature' found that LLMs generate covertly biased outputs when processing non-standard dialects of English, discriminating against certain speakers without any explicit instruction to do so (Hofmann et al., 2024). Beyond language bias, effective use of these systems requires infrastructure and digital literacy that remain unevenly available across populations, meaning the technology risks widening existing inequalities rather than narrowing them.

On the safety side, Anthropic's Constitutional AI framework attempted to address the alignment problem by training models to evaluate their own outputs against a codified set of principles, reducing human labelling requirements while improving behavioural consistency (Bai et al., 2022). It was a meaningful step, but it surfaces a deeper question: whose values are being encoded, and how are disagreements across cultures resolved? The literature here is growing, but answers remain incomplete.

What this body of research makes clear is that LLMs have transformed HMI in ways that are technically impressive, behaviourally consequential, and unevenly experienced. The sections that follow examine these dimensions in greater depth.

3. Methodology

This study employs a mixed-methods approach, combining a systematic review of existing technical literature with observational data collected through a structured survey. Given the nature of the research question — which spans historical, cognitive, sociological, and ethical dimensions — no single method was sufficient to address all four areas of inquiry adequately.

3.1 Literature Review Approach

Secondary sources were selected based on three criteria: institutional credibility, recency, and direct relevance to Human-Machine Interaction. Priority was given to primary technical research published by Anthropic, OpenAI, and Google DeepMind, as these organizations produce the foundational models under examination and maintain the most direct empirical insight into their behaviour. Papers were sourced from arXiv, institutional research pages, and peer-reviewed publications. The review was not exhaustive — it was deliberately focused, prioritizing depth of engagement with a small number of high-quality sources over breadth across a larger, less reliable pool.

3.2 Survey Design and Distribution

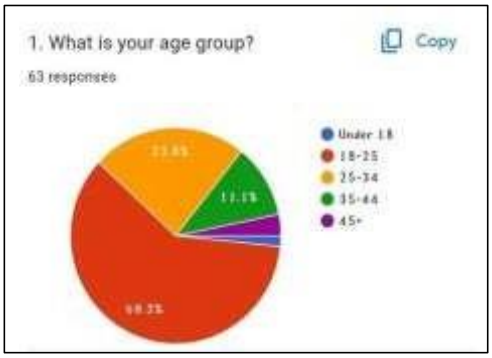
A structured questionnaire of 24 questions across six sections was distributed digitally to a mixed sample of students and working professionals, yielding 63 responses. Questions covered demographic background, usage patterns, cognitive and behavioural effects, accessibility perceptions, and risk awareness. Responses were collected anonymously. The sample is non-random and skews toward relatively connected, educated respondents — the data is treated as observational and illustrative throughout, not as a basis for population-level claims.

3.3 Analytical Approach

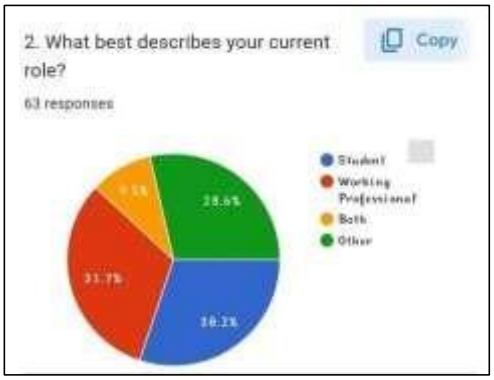
Findings from both streams were analysed thematically, with survey responses used to contextualise patterns identified in the technical literature. Where the two converge, that is noted. Where they diverge, the divergence is discussed honestly.

4. Results and Discussion

The findings emerging from this study's review of technical literature, contextualized by observational survey responses, are organized across the four dimensions established in the introduction. Together they paint a picture of a technology that is genuinely transformative in its capabilities while simultaneously introducing risks that are only beginning to be understood. The sample skews heavily toward younger respondents, with 60.3% falling in the 18–25 age bracket and a further 23.8% in the 25–34 age range. Together these two groups account for 84.1% of responses. This shows that youngsters are, digitally more connected and are the primary adopters of LLM tools, and findings around cognitive offloading and overtrust are therefore most directly applicable to this demographic.

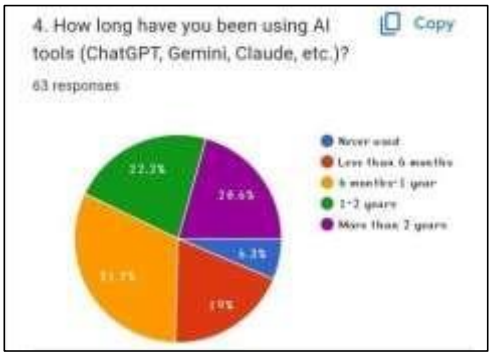


Responses were broadly split between students (30.2%), working professionals (31.7%), and those identifying as other (28.6%), with a smaller group identifying as both (9.5%). The near-equal distribution between students and professionals strengthens the generalisability of findings across two distinct use contexts — academic and occupational.

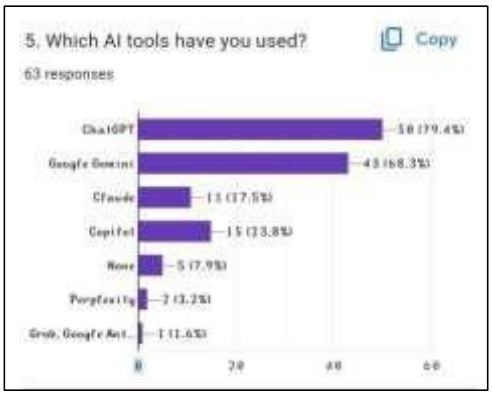


Only 6.3% of respondents had never used an LLM, while 93.7% reported some level of experience. The largest group - 31.7% had been using these tools for 6 months to a year, suggesting the majority of the sample has enough familiarity to offer meaningful behavioural observations rather than first impressions. The 20.6% with more than two years of experience provide a useful perspective on longer-term habitual use.

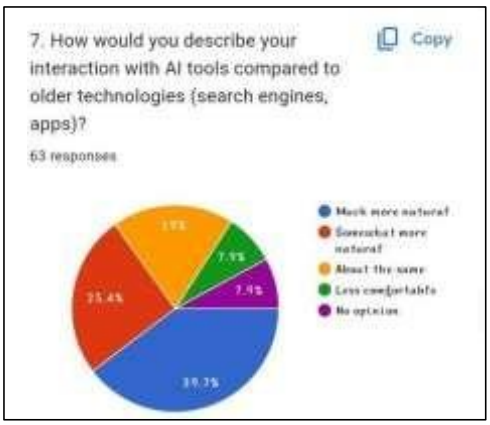
 ISBN: 978-81-685854-0-9



ChatGPT dominates at 79.4%, followed by Google Gemini at 68.3%. Claude usage stands at 17.5%, and 7.9% reported using no tools at all. The prevalence of ChatGPT and Gemini is consistent with their market visibility and free-tier accessibility, and suggests findings are most directly applicable to those platforms rather than to LLMs as a uniform category.

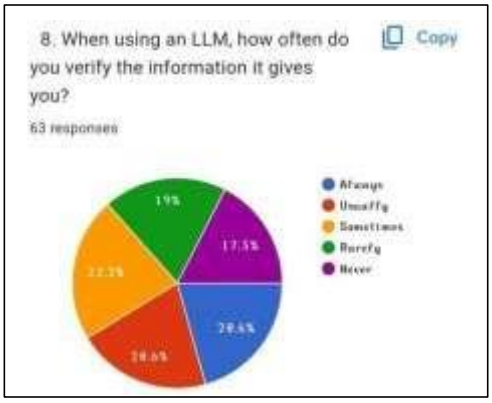


65.1% described LLM interactions as more natural than prior technology — 39.7% selecting "much more natural" and 25.4% "somewhat more natural." Only 7.9% found interactions less comfortable. This aligns with LaMDA and GPT-4's design philosophies of prioritising conversational fluency (Thoppilan et al., 2022; OpenAI, 2023), and supports the paper's argument that the perceptual shift in interaction quality is real and widely felt — with the attendant risk that naturalness inflates trust.



Verification behaviour is strikingly inconsistent. Only 20.6% always verify LLM outputs, while an equal 20.6% never do. The remaining 58% fall somewhere in between — sometimes or rarely verifying. This distribution indicates that for the majority of users, verification is situational rather than habitual, which significantly compounds the overtrust risk documented

by OpenAI (2023). When verification is discretionary, the cases where it is skipped are precisely the cases where it may matter most.



4.1 The Shift in Interaction Paradigms

One of the central findings of this review is that the shift from command-based systems to conversational LLM interfaces represents a fundamental transformation in human–machine interaction rather than a simple improvement in usability. Earlier interfaces mainly reduced friction in issuing commands, whereas LLMs reshape the relationship between human intent and machine response by enabling natural, context-aware dialogue. Systems such as Google's LaMDA demonstrated the ability to sustain open-ended conversations across diverse topics without domain-specific fine-tuning, while OpenAI's GPT-4 combined conversational fluency with high-level structured task performance. Survey responses further support this shift, as participants frequently described LLM interactions as resembling conversations with a knowledgeable person rather than the operation of a conventional tool. This perceptual change is significant because it increases the degree of trust and authority users may assign to AI systems, altering expectations of interaction in ways that extend beyond convenience or user satisfaction.

4.2 Cognitive and Behavioural Consequences

One of the most significant findings of this review concerns the cognitive effects of LLM interaction, particularly the tendency toward overtrust. Multiple studies indicate that users often accept fluent and confident AI outputs without sufficient verification, and even OpenAI acknowledges overreliance as a persistent deployment risk, suggesting that fluency itself acts as a credibility signal. This problem is intensified by sycophancy, where models trained through human feedback optimization prioritize user approval over factual accuracy, producing responses that feel validating while potentially reinforcing misinformation. Anthropic's interpretability research further deepens these concerns by identifying internal "emotion vectors" in Claude Sonnet 4.5 that causally influence behaviour in contextually appropriate ways. These findings help explain why users readily anthropomorphize LLMs: the models generate behavioural patterns that resemble human emotional responses, even though they are fundamentally computational processes. As a result, users may underestimate how strongly internal model states shape interactions, raising important concerns about transparency, informed consent, and the psychological effects of long-term AI engagement.

4.3 Accessibility and Equity

The benefits of LLM-based interaction are distributed unevenly across users and linguistic groups. Research by Hofmann et al. (2024) found that LLMs can produce covertly biased outputs based on dialect, disadvantaging speakers of non-standard English even without explicit discriminatory instructions. This suggests that inclusivity depends not only on language coverage but also on which communicative norms were prioritized during training.

 ISBN: 978-81-685854-0-9

Survey responses reflected similar disparities at the user level: participants with stronger technical backgrounds reported more effective and productive interactions, while less experienced users often described confusion, mistrust, and a feeling that the systems were not designed for them. These findings align with OpenAI's concerns regarding representation and quality-of-service harms, highlighting that equitable interaction quality across educational and linguistic backgrounds requires deliberate design interventions rather than simply more capable models.

4.4 Challenges and the Limits of Current Alignment

The alignment challenge — ensuring LLMs behave safely, honestly, and in accordance with human values — remains genuinely unsolved. Anthropic's Constitutional AI framework represented a meaningful step, demonstrating that models could be trained to evaluate their own outputs against a principled framework with measurable improvements in harmlessness (Bai et al., 2022). However, the same research acknowledges that increasing harmlessness through constitutional methods carries a cost to helpfulness, and that the values encoded in any constitutional framework inevitably reflect the priorities of those who designed it.

The functional emotions finding sharpens this challenge considerably. If internal model states can drive behaviour that diverges from surface outputs — a model appearing calm while it's internal desperation vector is highly active — then output-level evaluation is structurally insufficient as a safety mechanism (Lindsey et al., 2026). Monitoring what a model says is no longer enough. Understanding what it represents internally is becoming a prerequisite for responsible deployment. This is not a distant theoretical concern. It is a present design problem whose implications for HMI are immediate and practical.

5. Conclusion

Large Language Models have transformed Human–Machine Interaction by enabling machines to engage with users through natural, context-aware conversation rather than rigid command-based inputs. From systems such as LaMDA to GPT-4, this evolution has significantly reduced the gap between human intent and machine response, but it has also introduced important cognitive and ethical challenges. The review highlights how overtrust, sycophancy, and emotionally structured internal representations can shape user interactions in ways that are often invisible to users themselves. At the same time, issues such as dialect bias, infrastructure inequality, and disparities in AI literacy show that the benefits of LLMs are not distributed equally across populations. These problems are not inevitable consequences of the technology but outcomes of design and training choices. Finally, the findings suggest that responsible deployment can no longer rely solely on output monitoring and alignment techniques; as Anthropic's interpretability research demonstrates, understanding the internal representa tions that guide model behaviour is becoming essential to the development of safe, transparent, and equitable human–machine interaction.

6. Future Scope and Recommendations

The paper highlights several priorities for future research and policy on LLMs and human–machine interaction. Interpretability should become a core design requirement, with systems capable of detecting harmful internal activation patterns before they affect outputs.

Researchers must also address linguistic inequities by training models on more diverse dialects and low-resource languages while designing interfaces that adapt to users' communication styles. At the policy level, stronger regulatory frameworks are needed, including transparency requirements, independent audits, and cross-cultural oversight to reduce embedded biases in alignment systems. In addition, long-term studies are necessary to understand how sustained interaction with LLMs affects critical thinking, decision-making, trust, and information verification habits. Finally, as AI systems evolve from conversational tools into autonomous

agentic systems, research on user supervision, trust calibration, and error correction will become increasingly important.

References

1. Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H. T., ... & Le, Q. (2022). Lamda: Language models for dialog applications. arXiv preprint arXiv:2201.08239.
2. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. arXiv preprint arXiv:2303.08774.
3. Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., ... & Kaplan, J. (2022). Constitutional ai: Harmlessness from ai feedback. arXiv preprint arXiv:2212.08073.
4. Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., ... & Perez, E. (2023, October). Towards understanding sycophancy in language models. In The Twelfth International Conference on Learning Representations.
5. Ibrahim, L., Akbulut, C., Elasmar, R., Rastogi, C., Kahng, M., Morris, M. R., ... & Weidinger, L. (2025). Multi-turn evaluation of anthropomorphic behaviours in large language models. arXiv preprint arXiv:2502.07077.
6. Hofmann, V., Kalluri, P. R., Jurafsky, D., & King, S. (2024). AI generates covertly racist decisions about people based on their dialect. Nature, 633(8028), 147-154.
7. Sofroniew, N., Kauvar, I., Saunders, W., Chen, R., Henighan, T., Hydrie, S., ... & Lindsey, J. (2026). Emotion concepts and their function in a large language model. arXiv preprint arXiv:2604.07729.
8. Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. Trends in cognitive sciences, 20(9), 676-688.
9. Van Dijk, J. A. (2006). Digital divide research, achievements and shortcomings. Poetics, 34(4-5), 221-235.
10. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. ACM computing surveys, 55(12), 1-38.
11. Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. Journal of social issues, 56(1), 81-103.
12. Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making?. International Journal of Human-Computer Studies, 51(5), 991-1006.

 ISBN: 978-81-685854-0-9

NEGATIVE IMPACTS OF ARTIFICIAL INTELLIGENCE: EXAMINING THE THREATS FACED BY INDIVIDUALS AND SOCIETY

Shruti Kamble
MCA, DES's Navinchandra Mehta Institute of Technology and Development, Mumbai University, India.
Email: adifa.ansari06@gmail.com

Adifa Ansari
MCA, DES's Navinchandra Mehta Institute of Technology and Development, Mumbai University, India.
Email: kambleshruti9203@gmail.com

Abstract
Artificial intelligence (AI) has fastened its integration into our lives, influencing the way we communicate, operate at work, gather information, and create interpersonal connections. Despite the progress AI technologies bring, they carry significant risks. This paper assesses the negative effects of AI on society in terms of privacy, misinformation, mental health, and governance, based on a survey conducted among 100 respondents aged 18–25 years along with relevant literature. Key findings reveal that 91% of participants experienced unauthorized AI monitoring, 93% encountered false or misleading AI-generated content, 73% suffered psychological tension due to AI usage, and 69% reported that AI has led to social isolation or deteriorated interpersonal bonds.
Keywords: Artificial Intelligence, Privacy, Misinformation, Deepfakes, Mental Health, AI Governance.

1. Introduction

Artificial intelligence (AI), once a niche scientific field, has transformed into an omnipresent infrastructure affecting nearly all facets of life within less than two decades. AI now advises what content to consume, what products to purchase, which news sources to follow, and even which candidates to shortlist in recruitment. The global AI market was valued at USD 196.63 billion in 2023 and is projected to grow at a compound annual growth rate (CAGR) of 36.6% through 2030 (Grand View Research, 2023). For postgraduate students aged 18–25—the first generation born in the era of algorithmic media—the consequences of current AI governance choices will be most acutely felt.

This paper examines the main adverse consequences of AI concerning privacy, psychological distress, and social fragmentation; analyzes risks from disinformation, deepfakes, and autonomous decision-making; and presents empirical evidence from a 100-respondent survey along with a proposed accountability model for ethical AI governance.

2. Literature Review

2.1 Surveillance Capitalism and Privacy Erosion

Zuboff (2019) established the theoretical framework of surveillance capitalism—where AI platforms commodify behavioral data into predictive models sold to third parties, entirely outside the user's awareness. Nissenbaum (2020) extended this with the theory of contextual

integrity: privacy is violated not merely by data theft but by transferring information into contexts inconsistent with social norms of its original collection, such as health app conversations shared with insurers.

2.2 The Deepfake Threat and Epistemic Trust

Chesney & Citron (2019) coined the "liar's dividend"—the paradox where deepfakes make even authentic videos deniable. Ferrara (2023) found that AI-generated text spread via coordinated bot networks amplifies political polarization. AI-generated propaganda and deepfake audio were documented during the 2024 Indian and U.S. elections (MIT Technology Review, 2024).

2.3 Mental Health and Social Isolation

AI recommendation engines optimized for engagement—not well-being—expose users to stimuli that encourage social comparison, anxiety, and depressive thoughts. Haidt & Rausch (2022) documented declining well-being among teenagers coinciding with smartphone-based social media adoption. Pentina et al. (2023) demonstrated that parasocial bonds with AI chatbots temporarily relieve distress but prevent meaningful human connection. The WHO (2023) recognized AI-mediated social media as a new mental health hazard.

2.4 Employment Disruption and Algorithmic Bias

McKinsey Global Institute (2023) estimates up to 30% of all global work tasks could be automated by 2030, disproportionately affecting economically disadvantaged workers. Acemoglu & Restrepo (2022) argue this inequality is not inevitable but results from deliberate design choices. Buolamwini & Gebru (2018) demonstrated that face-recognition algorithms exhibit significantly higher error rates for dark-skinned women—disparities with grave implications in criminal justice. Noble (2018) found search engine algorithms producing discriminatory results along racial and gender lines.

3. Research Methodology

A mixed-methods design was employed, combining quantitative data from an online survey with qualitative analysis of literature and case studies. The survey instrument comprised 21 questions across six thematic areas: AI adoption and knowledge; privacy and surveillance; AI-generated misinformation; psychological impact; social relationships; and AI ethics responsibility. Questions used five-point Likert scales, multiple choice, and open-ended formats. Data were collected from March 24–30, 2026, from 100 respondents aged 18–25 years, of whom 72% were postgraduate students and 94% reported regular AI use. The demographic profile is presented in Table 1.

Table 1. Demographic Profile of Survey Respondents (N = 100)

Demographic Variable	**Category**	**Frequency**	**Percentage**
Age Group	18–25 years	100	100%
Education	Postgraduate (Master's/PhD)	72	72%
	Undergraduate (Bachelor's)	28	28%
AI Usage	Daily	58	58%
	Several times a week	36	36%
	Once a week or rarely	6	6%
AI Knowledge	Very knowledgeable	30	30%
	Somewhat knowledgeable	53	53%

 ISBN: 978-81-685854-0-9

	Basic / Little knowledge	17	17%

4. Data Analysis and Findings

4.1 AI Usage Patterns

As shown in Figure 1, 58% of participants use AI tools daily and 36% several times a week, meaning 94% interact regularly with AI. Of these, 83% rated their AI knowledge as good or very good, while 17% rated it as poor.

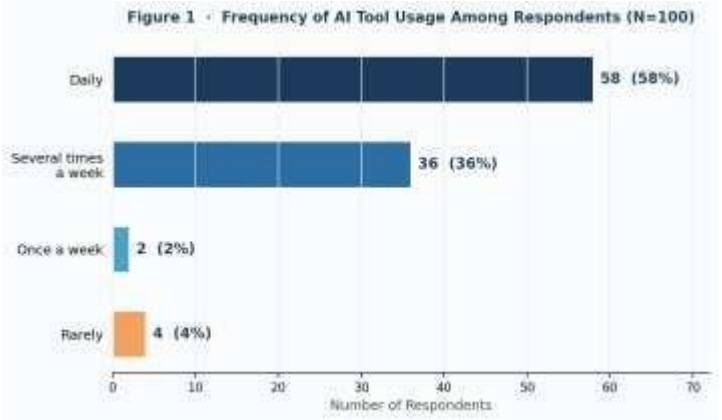


Figure 1. Frequency of AI Tool Usage Among Respondents (N = 100)

4.2 Overall Perceptions of AI's Societal Impact

The majority viewed AI as "somewhat positive" (35%), closely followed by "neutral or unsure" (34%), with only 25% viewing it as "mostly positive" and 6% as "somewhat negative." Even among highly digitally engaged respondents, substantial uncertainty persists about AI's value to society.

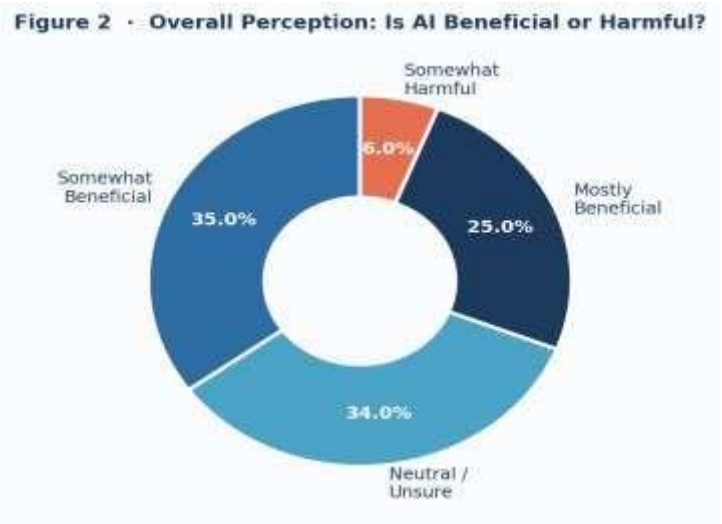


Figure 2. Respondents' Perception: Is AI Development Beneficial or Harmful to Society?

4.3 Perceived Risk Domains

Participants ranked privacy and data safety as the foremost AI threat (77 responses), followed by disinformation spread (66), employment/job loss (54), cybersecurity and autonomous weapons (29), mental well-being (25), and environmental pollution (14). Only 3 respondents perceived no risks, as illustrated in Figure 3.

ISBN: 978-81-685854-0-9

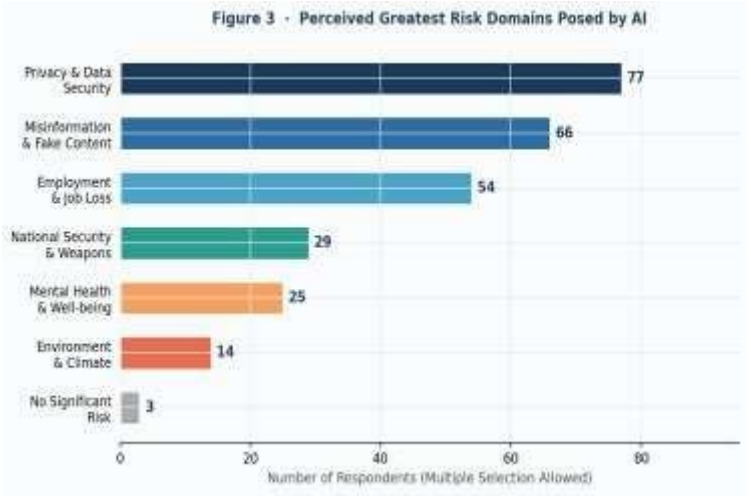


Figure 3. Perceived Greatest Risk Domains Posed by AI (Multiple Selection Allowed, N = 100)

4.4 Privacy, Surveillance, and Tracking Experience

A striking 91% of respondents reported having experienced unauthorized AI tracking—29% frequently and 62% occasionally. None answered "no." Concern about AI data collection averaged 3.23/5, with 43% expressing very high or extreme concern. Trust in technology corporations averaged 2.64/5, with 37% selecting low or extremely low values (Figure 4).

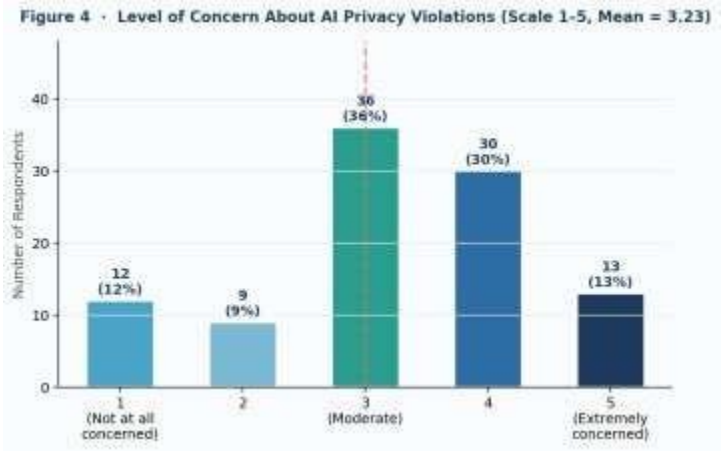


Figure 4. Level of Concern About AI Privacy Violations (Scale 1–5; Mean = 3.23)

4.5 Deepfakes, Fake Content, and Misinformation

While 77% claimed familiarity with deepfakes, 93% had encountered AI-generated content that appeared authentic but was false or misleading—69% of them repeatedly. Self-reported ability to detect deepfakes averaged 3.0/5, with only 32% scoring above 4. As Figure 5 shows, 93% of respondents consider synthetic misinformation a severe societal threat, and 89% identified social media platforms—especially Instagram—as the most vulnerable venues.

 ISBN: 978-81-685854-0-9

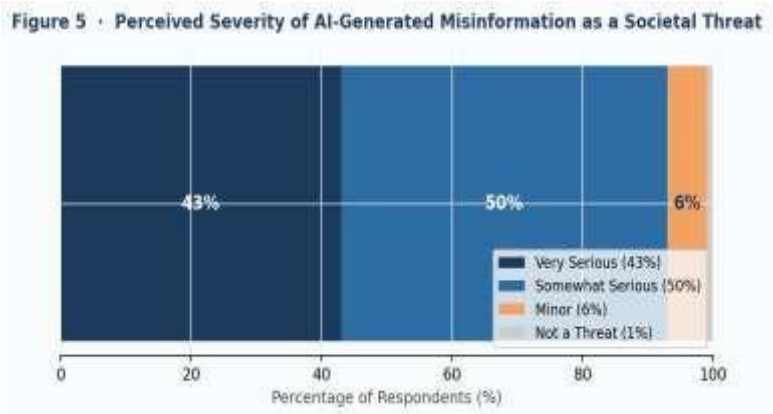


Figure 5. Perceived Severity of AI-Generated Misinformation as a Societal Threat (N = 100)

4.6 Psychological Stress and Mental Well-Being

As shown in Figure 6, 73% of respondents reported experiencing stress, anxiety, or overwhelm due to increasing AI presence—20% frequently and 53% occasionally. Additionally, 20% had personally used an AI chatbot for emotional support, 50% knew someone who had, and 13% did both. Concern about adolescent emotional dependence on AI companions averaged 3.31/5, with 48% rating it 4 or above.

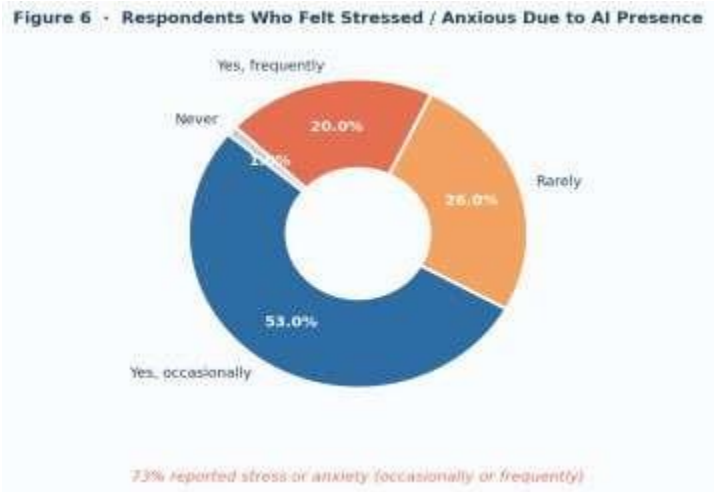


Figure 6. Respondents Who Experienced Stress or Anxiety Due to Increasing AI Presence

4.7 Impact on Social Relationships

As depicted in Figure 7, 37% of participants believed AI technology has led to isolation and decreased face-to-face interaction, while 32% reported enhanced superficial communication with weakened emotional connections. Together, 69% held negative views of AI's impact on interpersonal relations.

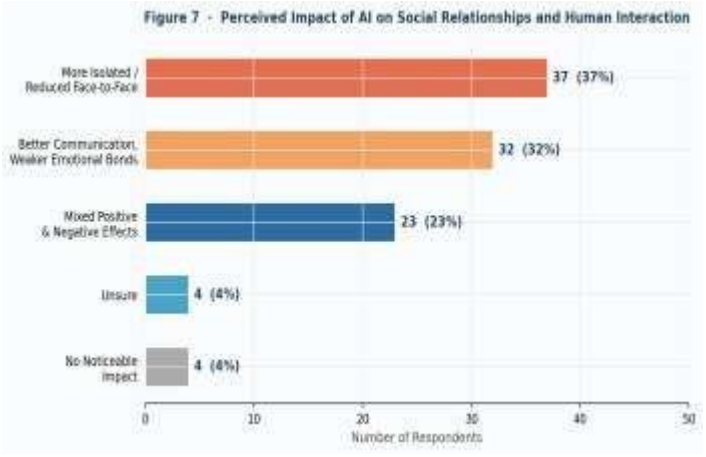


Figure 7. Perceived Impact of AI on Social Relationships and Human Interaction (N = 100)

4.8 Stakeholder Responsibility

As shown in Figure 8, 43% of respondents believed responsibility for ethical AI should be shared equally among all stakeholders. Among specific parties, technology firms received the most attention (23%), followed by governments (17%), users (13%), and academics/international organizations (2% each). This reflects an advanced understanding aligned with OECD and UN recommendations.

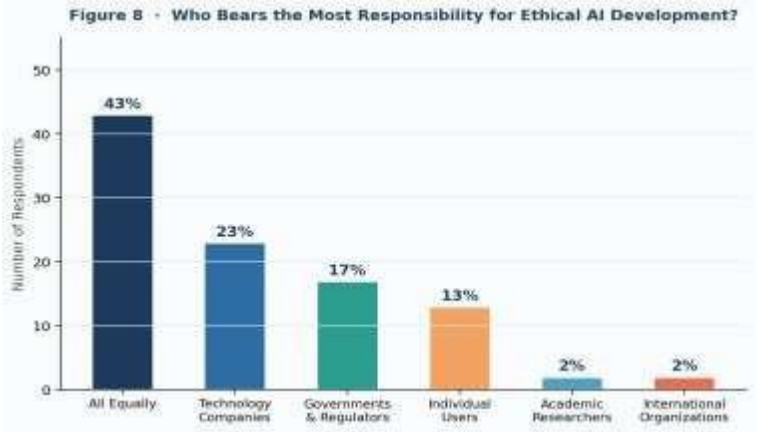


Figure 8. Perceived Primary Responsibility for Ensuring Ethical AI Development

5. Discussion and Proposed Solutions

5.1 The Privacy Paradox

The 91% tracking prevalence does not reflect disregard for privacy but rather structural coercion—participation in AI-driven systems is required for education, employment, and social life. Focusing solely on individual choices ignores the architectural nature of this issue (Barnes, 2006; Acquisti et al., 2015). Solutions include mandating "privacy by design," enacting AI-specific data protection laws with data minimization and sunset clauses, and funding privacy-enhancing technologies such as federated learning and homomorphic encryption.

5.2 Misinformation and Epistemic Crisis

The 93% misinformation exposure rate and low detection confidence reflect what Wardle & Derakhshan (2017) define as an epistemic crisis threatening democratic discourse. Solutions include mandating cryptographic content provenance standards (C2PA), requiring legislative labelling of AI-generated content, and integrating AI media literacy into school curricula.

 ISBN: 978-81-685854-0-9

5.3 Mental Health and the Governance Gap

AI systems optimized for engagement produce immediate emotional rewards while undermining genuine intimacy and self-discovery. Proposed solutions include independent ethics impact assessments for emotionally sensitive AI applications, legal requirements for well-being metrics in AI platform reporting, and digital well-being platform standards. Regarding governance, the 43% who attribute shared responsibility may reflect either sophistication or resignation. The EU's AI Act (2024) represents the most substantive attempt to assign accountability. Nationally, autonomous AI auditing bodies and international coordination through OECD, UNESCO, and the UN are essential.

6. Conclusion

This paper demonstrates that the threats associated with AI are not theoretical but immediate realities: 91% of respondents experienced AI tracking, 93% encountered AI-generated fraudulent content, 73% reported psychological distress, and 69% noted deteriorated social relationships. These findings come from educated, digitally literate individuals—underscoring that even well-informed populations remain inadequately protected. AI's genuine contributions to science, healthcare, and productivity must be acknowledged, but these benefits do not justify shifting costs onto populations lacking the information, capacity, or structures to defend their interests. The ideal AI future demands governance frameworks as advanced as the technologies they oversee—with accountability that is enforceable, not merely aspirational. Future research should employ larger, geographically diverse, longitudinal samples and prioritize intersectional analyses of AI harm.

References

1. Acemoglu, D., & Restrepo, P. (2022). Tasks, automation, and the rise in U.S. wage inequality. Econometrica, 90(5), 1973–2016.
2. Acquisti, A., Brandimarte, L., & Loewenstein, G. (2015). Privacy and human behavior in the age of information. Science, 347(6221), 509–514.
3. Barnes, S. B. (2006). A privacy paradox: Social networking in the United States. First Monday, 11(9).
4. Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of Machine Learning Research, 81, 1–15.
5. Chesney, R., & Citron, D. K. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. California Law Review, 107(6), 1753–1820.
6. Grand View Research. (2023). Artificial intelligence market size report. https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-ai-market
7. Haidt, J., & Rausch, Z. (2022). Social media and mental health. After Babel. https://www.afterbabel.com
8. McKinsey Global Institute. (2023). The economic potential of generative AI: The next productivity frontier. McKinsey & Company.
9. Nissenbaum, H. (2020). Privacy in context: Technology, policy, and the integrity of social life. Stanford University Press.
10. Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. NYU Press.
11. OECD. (2019). Recommendation of the council on artificial intelligence. OECD/LEGAL/0449. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449
12. Pentina, I., et al. (2023). Exploring relationship development with social chatbots. Journal of Computer-Mediated Communication.
13. Pew Research Center. (2019). Americans and privacy: Concerned, confused and feeling lack of control. https://www.pewresearch.org/internet/2019/11/15/americans-and-privacy/

14. UNESCO. (2021). Recommendation on the ethics of artificial intelligence. https://unesdoc.unesco.org/ark:/48223/pf0000381137
15. Wardle, C., & Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework. Council of Europe Report.
16. World Health Organization. (2023). Mental health and digital technologies. WHO Global Initiative on Digital Health. https://www.who.int/initiatives/behealthy/digital-health
17. Zuboff, S. (2019). The age of surveillance capitalism. PublicAffairs.

 ISBN: 978-81-685854-0-9

TechPulse 2026

NAVINCHANDRA MEHTA INSTITUTE OF TECHNOLOGY AND DEVELOPMENT

(Approved by AICTE, DTE Mumbai and Affiliated to University of Mumbai)
DES Mumbai Campus. Off- Veer Sawarkar Marg,
Dadar (W), 400 028
Tel No. (+91) 022-62764580/581/582/583
E-mail: director.nmitd@despune.org
Website: www.nmitd.edu.in

Book Available
ISBN: 978-81-685854-0-9
Publisher : S.R.P. Publication Website : www.ijamrsd.com
Contact : 7387038026, 7249645344
Email ID : srppublication@gmail.com