IITM Journal of Information Technology

Volume 12 (January - December 2026)
ISSN No. 2395 – 5457

IITM
Nurturing Excellence

INSTITUTE OF INNOVATION IN TECHNOLOGY & MANAGEMENT
Affiliated to GGSIPU, NAAC Grade 'A', ISO 14001:2015, 17020:2012,
21001:2018 & 50001:2018 Certified, A Grade by GNCTD, A++ Grade by SFRC

IITM Journal of Information Technology

IIITM Journal of Information Technology is a National Annual Journal of Information Technology intended for professionals and researchers in all fields of Information Technology. Its objective is to disseminate experiences, ideas, case studies of professionals in Information Technology to propagate better understandings. Its focus is on empirical and applied research and reflections that are relevant to professionals of Information Technology with academic standards and rigor within the purview. The views expressed in the Journal are those of the authors. No part of this publication may be reproduced in any form without the written consent of the publisher.The Journal intends to bring together leading researchers and Information Technology practitioners from around the country.

IITM Journal of Information Technology

Annual Journal of Institute of Innovation in Technology & Management

Volume 12 **January - December, 2026**

CONTENTS

Research Papers & Articles

Digital Object Identifier: https://doi.org/*10.48165/iitmjit.2026.12.1.1*

Untangling the Complexity: Comparative Analysis of AI/ML Methodologies in Knot Theory

Mariya Ajaz[1], Priyanka Bhutani[2]
[1,2]University School of Information, Communication and Technology, GGSIPU, New Delhi
[1]mariya.716404825@std.ggsipu.ac.in,[2]priyanka.b@ipu.ac.in

Abstract- The computational challenge of determining whether a tangled loop is topologically equivalent to the unknot is a fundamental benchmark in computational topology. While classical algorithms established this problem in the complexity class NP, the exponential growth of structural data has catalyzed a paradigm shift toward Artificial Intelligence (AI) and Machine Learning (ML) approaches. This paper presents a comparative analysis of contemporary AI/ML techniques applied to the unknotting problem and knot theory at large. We survey foundational and modern studies, mapping the evolution from classical polynomial invariants to modern Deep Reinforcement Learning (DRL) agents and supervised learning models. Furthermore, we conduct a comparative analysis evaluating the inductive biases of different data representations, specifically contrasting sequence-based models with planar graph representations processed via Graph Neural Networks (GNNs). Our analysis extends to real-world applications in molecular biology and robotics, concluding that while sequence models offer computational efficiency, graph-based topological modeling provides superior geometric generalization.

Keywords- Knot Theory, Deep Reinforcement Learning, Graph Neural Networks, Computational Topology, Unknotting

1. Introduction

In mathematical topology, a knot is defined as a smooth embedding of a circle (S^1) into three-dimensional Euclidean space (R^3). The fundamental challenge is the recognition problem: devising an algorithm to determine if two distinct descriptions—typically presented as 2D diagrams—represent the same underlying topological object. A historically significant sub-problem is the unknotting problem, which asks if a given knot can be deformed into the simplest possible "unknot" through an ambient isotopy. Determining unknot equivalence is a benchmark for topological complexity; while decidable, it is known to be NP-hard, and the search space for the correct sequence of moves is vast [1].

1.1 Background

Knot diagrams represent 3D structures on a 2D plane, with breaks at crossings indicating strand orientation. The bridge between continuous deformation and discrete diagrams is formalized by the Reidemeister moves: Type I (Twist), Type II (Poke), and Type III (Slide) [2],[3]. Any function on these diagrams that remains unchanged under these moves is a knot invariant. Classical invariants include the Alexander polynomial [4] and the Jones polynomial [5], which can often distinguish a knot from its mirror image (chirality). Despite their power, computing these invariants for high-crossing knots remains computationally expensive [6].

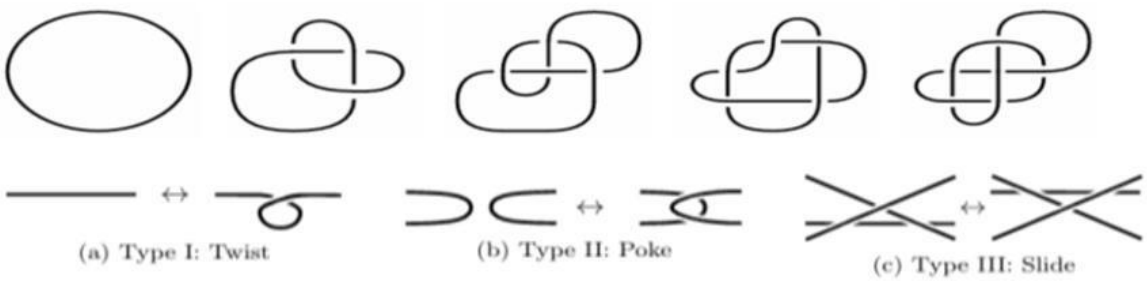

Fig. 1.1: Examples of standard prime knots and the three fundamental Reidemeister moves used for continuous diagram deformation adapted from [7].

2. Literature Review

The unknotting problem began in the 19th century with Peter Guthrie Tait's classification program, inspired by Lord Kelvin's "vortex-atom theory". Modern study was formalized in the 1920s by Kurt Reidemeister, who established a combinatorial foundation based on three moves (Type I, II, and III), transforming a continuous problem into a discrete search space [2]. Alexander and Jones later introduced algebraic polynomials as invariants to distinguish knots, though their computation remains expensive [4], [5].

Algorithmic geometric solutions to the problem were first given by Haken in 1961, which proved decidability but lacked practical efficiency [8]. Subsequent complexity research placed unknots recognition in NP, setting the stage for data-driven heuristics [1]. Machine learning (ML) was catalysed by large computationally generated databases. Initial efforts focused on classification from polymer conformations [9], which evolved into deep neural networks predicting complex 4D invariants from 3D invariants with >95% accuracy [10], [11]. Landmark work by Google DeepMind utilized saliency analysis to guide mathematicians toward proving new theorems linking geometric "cusp" properties to algebraic signatures [12].

Reinforcement learning (RL) has become the dominant paradigm for active unknotting due to the sequential nature of diagram simplification. The state-of-the-art agents have determined unknotting numbers for over 57,000 knots and identified 2.6 million "hard unknot" diagrams [7], [13]. These advancements find parallels in molecular biology, where AlphaFold 2 has uncovered rare composite protein knots [14], [15], and in robotics, where cognitive architectures like HULK combine learning-based perception with knot manipulation primitives [16].

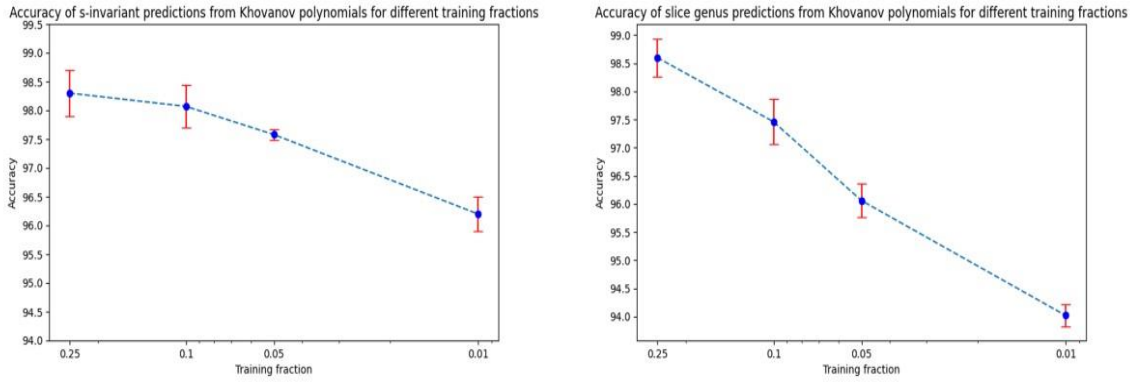

Fig. 1.2: Neural network accuracy in predicting slice genus and s-invariant from Khovanov polynomials at varying training fractions adapted from [10].

3. Methodology

3.1 Data Representation of Knots

The choice of representation is critical as it introduces an inductive bias that guides the model's pattern recognition capabilities [17], [18]. The primary methods for encoding knots include:

- **Braid Words:** Every knot can be represented as the closure of a braid. Described algebraically as a one dimensional sequence of generators and their inverses, this format is highly data-efficient and well-suited for sequence-processing models [18], [19]. Braid representations often yield the highest accuracy for predicting invariants because their algebraic structure aligns closely with homological definitions.
- **Dowker-Thistlethwaite (DT) and Gauss Codes:** These compact notations record the sequence of crossings encountered during a traversal of a 2D knot diagram [17]. While efficient, they encode topology in a non-local manner, which can make it difficult for models to infer geometric structure without specific architectural considerations.
- **3D Coordinates (Polygonal Curves):** This geometrically intuitive format represents a knot as a list of (x, y, z) coordinates forming a polygonal curve. It is frequently used in physical contexts like protein or DNA modelling [20]. However, it is memory-intensive and sensitive to rotations and translations.
- **Graph-Based Representations:** This approach treats knot diagrams as graphs where nodes represent crossings and edges represent connecting arcs. This allows the use of Graph Neural Networks (GNNs) to pass messages between nodes, leveraging the combinatorial structure of the diagram [17], [21].

3.2 Model Architectures

The selection of neural network architecture is directly informed by the chosen data representation:

- **Feed-Forward Neural Networks (FFNNs):** These are used for processing fixed-size vector inputs, such as feature vectors derived from knot polynomials or DT codes [12], [17], [18].
- **Recurrent Neural Networks (RNNs) and LSTMs:** Designed for sequential data, these models maintain an internal "memory" to capture dependencies along a sequence, making them effective for braid words or polymer coordinates [9].
- **Transformers:** Utilizing self-attention mechanisms, Transformers weigh the importance of different sequence elements regardless of distance. They have achieved state-of-the-art results in learning topological invariance from braid words and recognizing molecular knots in 3D sequences [20].
- **Graph Neural Networks (GNNs):** GNNs propagate information across a knot diagram's graph structure to learn features that are inherently topological and invariant to how the diagram is drawn [17].

3.3 Algorithmic Workflows

The learning process typically follows one of the following paradigms:

- **Supervised Learning:** Used to map inputs to known outputs, such as uncovering hidden relationships between algebraic and geometric invariants [11], [12].
- **Reinforcement Learning (RL):** Unknotting is treated as a strategic decision-making problem formalized as a Markov Decision Process (MDP). The agent observes a state (diagram), selects an action (move), and receives a reward (crossing reduction) [22], [13].
- **Graph Neural Networks (GNNs):** Utilize functors to map knots to face-adjacency graphs, propagating features that are inherently diagram-invariant [17].

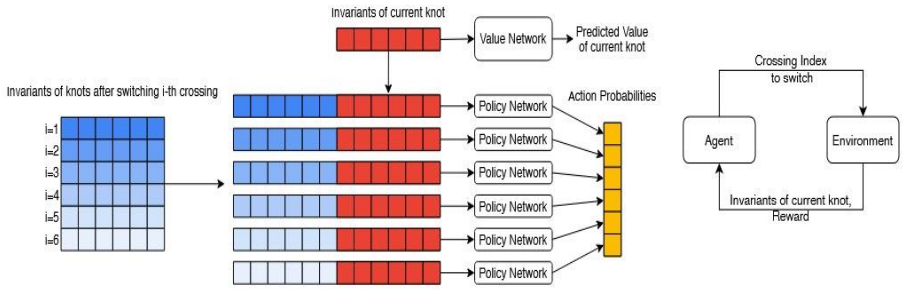

Fig. 1.3: The architecture of the reinforcement learning agent. The model processes topological invariants of the current and neighbouring diagrams through value and policy networks to determine the optimal crossing switch adapted from [13].

4. Comparative Analysis

A machine learning model's success in knot theory largely comes down to how we translate a physical knot into raw data. Table 1 breaks down the most common representations used in the literature, comparing the trade-offs between highly compact algebraic formats (like Braid Words) and more structurally intuitive ones (like Graph Representations) [18].

Table 1.1**:** Comparison of Knot Data Representation

Representation	**Data Format**	**Geometric Intuition**	**Strengths**	**Weaknesses**	**Suitable ML Models**
Braid Words	Integer Sequence	Low	Compact; strong algebraic structure [18].	Not unique [17].	RNNs, Transformers [18], [19]
DT Code	Integer Sequence	Low	Extremely compact; canonical for prime knots [17].	Non-local encoding [17].	FFNNs, Transformers [17], [18]
3D Coordinates	Array of Floats	High	Geometrically intuitive; essential for molecules [20].	Not rotation invariant [9].	Transformers, CNNs [9], [20]
Graph Rep.	Adjacency Matrix	High	Captures combinatorial structure natively [17].	Complex construction [17], [21].	GNNs [17]

The graph in figure 1.4 demonstrates the superior inductive bias of Braid and Planar Diagram (Graph) representations for complex topological tasks. Using those data representations as a foundation, Table 1.2 compares the actual performance of the leading AI frameworks. This matrix highlights what each methodology does best, what it still struggles with, and its most significant recent breakthroughs—charting the field's shift

from simply predicting invariants to actively untangling knots using reinforcement learning and quantum algorithms [13], [12], [23], [24].

Table 1.2: Performance Matrix of AI Paradigms

Methodology	Inductive Bias	Suitability	Key Achievement	Key Weakness
Supervised (FFNN)	Low	Invariant Prediction	97.6% Volume Accuracy [25]; 95% accuracy for 4D invariants[10]	Lacks spatial context
Transformers	Medium	Molecular Recognition	99% accuracy in recognizing molecular structures [20]; 4500x faster than classical methods	High compute memory
GNNs	High	Structural Mapping	Generalization to knots with up to 91 crossings [26] invariant to diagram layout	Message-passing overhead
Deep RL	Variable	Active Untangling	New unknotting numbers for 43 knots [13]; determined numbers for 57k knots	Sample inefficiency
Quantum Algorithms	N/A	Complexity Taming	Polynomial Jones approx [27]; Khovanov homology approx [28]	Hardware noise/Scale

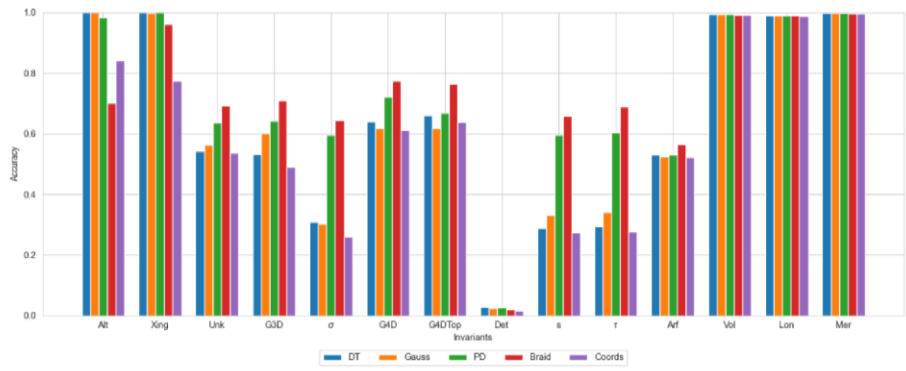

Fig. 1.4: Comparative accuracy of neural networks in predicting various knot invariants based on their input data representation (DT code, Gauss code, Planar Diagram/PD, Braid words, and 3D Coordinates) adapted from [18].

5. Conclusions

The fusion of machine learning with knot theory has evolved from basic classification into a sophisticated partnership for mathematical discovery. Supervised learning models now demonstrate exceptional precision, achieving near-perfect accuracy in knot classification and revealing hidden correlations between complex invariants. A defining breakthrough is the work of Davies et al., which showcased AI's role as an "intuition amplifier" by guiding mathematicians toward the formulation and proof of entirely new theorems [12].

The adoption of reinforcement learning (RL) has revolutionized the approach to sequential topological problems. By framing the unknotting process as a Markov Decision Process, RL agents have successfully

identified minimal move sequences for complex diagrams, resolving previously unknown unknotting numbers for tens of thousands of knots [22], [29]. Supervised learning has successfully uncovered empirical relationships between invariants [11], while deep reinforcement learning has achieved tangible progress in solving the active unknotting problem for tens of thousands of knots [7], [13]. Furthermore, the comparative analysis concludes that while sequence based models offer computational speed, Graph Neural Networks (GNNs) provide the most robust inductive bias for structural mapping. By natively capturing the combinatorial connectivity of strands, GNNs have demonstrated superior generalization capabilities, successfully scaling to complex knot diagrams [17].

6. Future Scope

While current machine learning frameworks offer impressive heuristic solutions, they have yet to answer the fundamental computational question of whether unknot recognition can truly be solved in polynomial time [1], [30], [31], [32]. Resolving this requires moving beyond the "black box" nature of traditional deep learning. To genuinely advance pure mathematics, future systems must prioritize transparency—perhaps utilizing inherently interpretable architectures to help researchers translate AI-recognized statistical patterns into rigorous, formally verified proofs [12]. Furthermore, as researchers inevitably hit the strict processing limits of classical hardware, the field must look toward quantum computation. Because evaluating complex invariants like the Jones polynomial is a BQP-complete problem, the eventual arrival of fault-tolerant quantum computers promises a radically new paradigm that could bypass our current algorithmic bottlenecks entirely [23], [33].

Looking forward, the role of AI in topology is expanding from simply solving known problems to actively hunting for new mathematical anomalies [34]. For example, rather than just untangling diagrams, generative models are being explored to construct rare, external knots that might serve as counterexamples to long-held assumptions— such as the conjecture that the Jones polynomial definitively detects the unknot. The highly adaptable reinforcement learning principles developed today are also perfectly positioned to tackle much deeper four dimensional challenges, including the slice-ribbon problem and its profound ties to the smooth Poincare conjecture [35]. Ultimately, the goal is not full automation, but the creation of seamless, interactive environments where a mathematician's intuition is paired with real-time AI testing and visualization, forging a deeply collaborative loop for future discovery [12].

References

1. Hass, J., Lagarias, J. C., & Pippenger, N. (1999). The computational complexity of knot and link problems. *Journal of the ACM, 46*(2), 185–211. https://doi.org/10.1145/301970.301971.
2. Reidemeister, K. (1927). Elementare Begründung der Knotentheorie. *Abhandlungen aus dem Mathematischen Seminar der Universität Hamburg, 5*(1), 24–32. https://doi.org/10.1007/BF02952507.
3. Li, Z., Ding, K., & Ma, G. (2023). Eigen value knots and their isotopic equivalence in three-state non-Hermitian systems, *Physical Review Research, 5*(2), 023038, https://doi.org/10.1103/PhysRevResearch.5.023038
4. Alexander, J. W. (1928). Topological invariants of knots and links. *Transactions of the American Mathematical Society, 30*(2), 275–306. https://doi.org/10.1090/S0002-9947-1928-1501429-1.
5. Jones, V. F. R. (1985). A polynomial invariant for knots via von Neumann algebra. *Bulletin of the American Mathematical Society (N.S.), 12*(1), 103–111. https://doi.org/10.1090/S0273-0979-198515304-2.
6. Lisitsa, A. (2022). An application of neural networks to a problem in knot theory and group theory (untangling braids) [Preprint]. arXiv. https://arxiv.org/abs/2206.05373.
7. Gukov, S., Halverson, J., Ruehle, F., & Suwara, P. (2021). Learning to unknot. *Machine Learning: Science and Technology, 2*(2), Article 025035. https://doi.org/10.1088/2632-2153/abe91f.
8. Haken, W. (1961). Theorie der Normalflächen: Ein Isotopiekriterium für die Wortprobleme der Gruppentheorie. *Acta Mathematica, 105*, 245–375. https://doi.org/10.1007/BF02559591.

9. Vandans, O., Yang, K., Wu, Z., & Dai, L. (2020). Recognizing knots in a polymer conformation by machine learning. *Physical Review E, 101*(2), Article 022502. https://doi.org/10.1103/PhysRevE.101.022502.
10. Craven, J. (2023). Learning knot invariants across dimensions. *SciPost Physics, 14*, Article 021. https://doi.org/10.21468/SciPostPhys.14.2.021.
11. Craven, J., Hughes, M., Jejjala, V., & Kar, A. (2024). Illuminating new and known relations between knot invariants. *Machine Learning: Science and Technology, 5*(4), 1–25. https://doi.org/10.1088/26322153/ad95d9.
12. Davies, A., Veličković, P., Buesing, L., Blackwell, S., Zheng, D., Tomašev, N., Tanburn, R., Battaglia, P., Blundell, C., Juhász, A., Lackenby, M., Geordie, W., Hassabis, D., & Kohli, P. (2021). Advancing mathematics by guiding human intuition with AI. *Nature, 600*(7887), 70–74. https://doi.org/10.1038/s41586-021-04086-x.
13. Applebaum, T., Juhász, A., Ruehle, F., & Gukov, S. (2025). The unknotting number, hard unknot diagrams, and reinforcement learning. *Experimental Mathematics, 34*(1), 1–25. https://doi.org/10.1080/10586458.2025.2542174.
14. Flapan, E., Henrich, A., & Wong, H. (2022). AlphaFold predicts the most complex protein knot and composite protein knots. *Protein Science, 31*(8), Article e4380. https://doi.org/10.1002/pro.4380.
15. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., RomeraParedes, B., Stanway, S., Faneva, R., Vinyals, O., ... Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature, 596*(7873), 583–589. https://doi.org/10.1038/s41586-02103819-2.
16. Hang, K., Lerrel, P., Wang, S., Abbeel, P., & Solowjow, E. (2023). SGTM 2.0: Autonomously untangling long cables using interactive perception. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)* (pp. 1–7). IEEE. https://doi.org/10.1109/ICRA48891.2023.10161234.
17. Driggs, N. (2025). Exploring representations and inductive bias for machine learning tasks in knot theory [Master's thesis, Brigham Young University]. BYU ScholarsArchive. https://scholarsarchive.byu.edu/etd/10813/.
18. Lindsay, A. (2025). On the learnability of knot invariants: Representation, predictability, and neural similarity [Preprint]. arXiv. https://arxiv.org/abs/2502.12243.
19. Ruehle, F. (2025). Learning topological invariance [Preprint]. arXiv. https://arxiv.org/abs/2504.12390.
20. Zhang, Z., Zhu, Y., & Dai, L. (2025). Recognizing and generating knotted molecular structures by machine learning [Preprint]. arXiv. https://arxiv.org/abs/2501.12780.
21. de Boer, J., Sazdanovic, R., & Suwara, P. (2024). A phenomenological approach to interactive knot diagrams. In *Proceedings of IEEE VIS* (pp. 1–5). IEEE. https://doi.org/10.1109/VIS54894.2024.00010.
22. Al-Amayreh, M., & Salles, M. (2022). Untangling braids with deep reinforcement learning. In *Proceedings of the 35th International Florida Artificial Intelligence Research Society Conference (FLAIRS)*. https://doi.org/10.32473/flairs.v35i1.130657.
23. Quantinuum. (2024, May 22). Untangling the mysteries of knots with quantum computers. *Quantinuum Blog*. https://www.quantinuum.com/blog/untangling-the-mysteries-of-knots-with-quantum-computers.
24. Joshi, R., Sastry, K., Bharathi, M., & Bhutani, P. (2023). Smart cities integrating artificial intelligence and IoT: A review. *International Journal of Computer Applications, 182*(47), 1–5.
25. Jejjala, V., Kar, A., & Parrikar, O. (2019). Deep learning the hyperbolic volume of a knot. *Physics Letters B, 799*, Article 135033. https://doi.org/10.1016/j.physletb.2019.135033.
26. Jaretzki, L. (2023). *Geometric deep learning approach to knot theory*. arXiv. https://doi.org/10.48550/arXiv.2305.16808.
27. Laakkonen, T., Rinaldi, E., Self, C. N., Chertkov, E., DeCross, M., Hayes, D., ... & Meichanetzidis, K. (2025). *Less quantum, more advantage: An end-to-end quantum algorithm for the Jones polynomial*. arXiv. https://doi.org/10.48550/arXiv.2503.05625.

28. Schmidhuber, A., Reilly, M., Zanardi, P., Lloyd, S., & Lauda, A. (2025). *A quantum algorithm for Khovanov homology*. arXiv. https://doi.org/10.48550/arXiv.2501.12378.
29. Jackson, N. (2013). Evolution of unknotting strategies for knots and braids [Preprint]. arXiv. https://arxiv.org/abs/1302.0787.
30. Lackenby, M. (2020). Unknot recognition in quasi-polynomial time. *Foundations of Computational Mathematics, 20*(5), 1141–1179. https://doi.org/10.1007/s10208-020-09455-8.
31. Verma, N., Bhutani, P., Lalit, R., & Venugopal, S. (2025). Map reduce framework-assisted feature analysis and adaptive multiplicative Bi-RNN using big data analytics for decision-making. *International Journal of Computational Intelligence Systems, 18*(1), 1–30.
32. Gukov, S. (2023, May). Machine learning and hard problems in topology [Workshop presentation]. ICERM Topical Workshop on the Mathematics of Knots, Brown University, Providence, RI, United States. https://icerm.brown.edu/program/topical_workshop/tw-23-tkt.
33. Dhaka, P., Sehrawat, R., & Bhutani, P. (2023). An innovative approach to cardiovascular disease prediction: A hybrid deep learning model. *Engineering, Technology & Applied Science Research, 13*(6), 12396–12403. https://doi.org/10.48084/etasr.6503.
34. Sazdanovic, R. (2025). Data driven perspectives on knot theory [Preprint]. arXiv. https://arxiv.org/abs/2503.15103.
35. Gukov, S., Halverson, J., Manolescu, C., & Ruehle, F. (2025). Searching for ribbons with machine learning. *Machine Learning: Science and Technology, 6*(2), 1–25. https://doi.org/10.1088/26322153/ade362.

Digital Object Identifier: https://doi.org/10.48165/iitmjit.2026.12.1.2

An Explainable Resume Analyzer Using ESCO Ontology and Semantic Skill Mapping for Transparent Candidate Ranking

Sushma Malik[1], Anamika Rana[2], Ayush Pal[3], Ritham Gupta[4]
[1] Assistant Professor, [2]Associate Professor, [3]Research Scholar
[1,2,3,4] Department of Computer Applications, Maharaja Surajmal Institute, New Delhi, India
[1]sushmalik25@gmail.com, [2]anamica.rana@gmail.com, [3]ayush34492@gmail.com

Abstract: Traditional recruitment processes often rely heavily on manual resume screening, which is both time-consuming and prone to human bias. These challenges can lead to inefficiencies, inconsistent evaluations, and potential overlooking of qualified candidates. This study presents a locally deployed, desktop-based Resume Analyzer designed to automate the initial screening phase while ensuring enhanced data privacy for small and medium-sized organizations. Unlike cloud-based Applicant Tracking Systems (ATS), the proposed system operates entirely on a local machine, reducing the risk of sensitive applicant data exposure.

The system utilizes Natural Language Processing (NLP) techniques to extract and analyze textual content from resumes in PDF, DOCX, and TXT formats. A specialized algorithm is implemented to accurately calculate total work experience by interpreting employment durations and aggregating them systematically. To standardize skill identification and improve matching accuracy, the system integrates a mapping mechanism based on the ESCO (European Skills, Competences, Qualifications and Occupations) framework developed by the European Commission. This ontology-driven approach ensures alignment with recognized occupational and skills classifications, promoting consistency and transparency in candidate evaluation.

Candidates are ranked using a configurable weighted scoring model that evaluates three primary components: skills (50%), work experience (30%), and education (20%). The scoring system provides clear justifications for rankings, enhancing interpretability and recruiter trust. Experimental evaluation demonstrates that the system effectively identifies top candidates while maintaining fairness and transparency. The findings suggest that localized, ontology-based resume analysis tools can serve as a practical, cost-effective alternative to cloud-hosted ATS solutions, particularly in privacy-sensitive environments where data security and control are critical considerations.

Keywords: Resume Analysis, ESCO Ontology, Semantic Skill Mapping, Candidate Ranking, Explainable AI (XAI), Natural Language Processing (NLP), Skills Classification

1. Introduction

Modern recruitment processes must handle an overwhelming number of resumes for a single job vacancy. With online job portals and digital applications becoming the norm, recruiters often receive hundreds or even thousands of resumes per posting. Manually screening such a large volume of documents is time-consuming, inefficient, and prone to inconsistencies or unconscious bias[1]. To address this issue, many organizations have adopted Applicant Tracking Systems (ATS) to automate resume parsing, indexing, and filtering. These systems help streamline hiring workflows by extracting structured information and shortlisting candidates based on predefined criteria[2].

However, most commercially available ATS platforms are cloud-based, subscription-driven, and feature-rich systems designed primarily for medium to large enterprises. While they offer scalability and advanced analytics, they often involve recurring costs, complex configuration processes, and dependency on internet connectivity. For small businesses, startups, academic institutions, and research laboratories, such solutions may not be practical or affordable. In addition, cloud-hosted systems raise concerns regarding data privacy and regulatory compliance, particularly when handling sensitive personal information such as employment history, contact details, and educational credentials. In privacy-sensitive environments, organizations may prefer to retain full control over applicant data by processing it locally rather than transmitting it to external servers[3].

To address these limitations, this study proposes a lightweight, desktop-based Resume Analyzer that operates entirely within a local environment. By functioning offline, the system ensures that candidate data never leaves the organization's infrastructure, thereby reducing risks associated with data breaches or unauthorized access.

This localized approach is especially suitable for small offices and educational labs that require cost-effective and secure recruitment tools. At the same time, the system aims to maintain a level of analytical capability comparable to more sophisticated ATS platforms.[4]

A key technical challenge in automated resume screening lies in transforming unstructured textual content into structured, machine-readable information. Resumes are typically submitted in various formats, such as PDF and Word documents, and may differ significantly in layout, design, and writing style. Extracting relevant details—particularly skills, work experience, and educational qualifications—requires robust parsing and text-processing techniques. Variations in section headings, inconsistent date formats, and graphical elements can further complicate information extraction. Therefore, developing a reliable parser capable of handling multi-format documents and diverse resume structures is a critical component of the proposed system[5].

Beyond extraction, normalization of skills presents another significant challenge. Candidates often describe similar competencies using different terminology. For example, "Java programming," "Java development," and "object-oriented programming in Java" may refer to related skill sets but appear distinct in raw text. Without semantic standardization, automated ranking systems may fail to recognize equivalencies or closely related competencies. To address this issue, the proposed solution incorporates a skill normalization mechanism based on the ESCO (European Skills, Competences, Qualifications and Occupations) framework. ESCO provides a structured, multilingual classification of skills and occupations, enabling consistent mapping between resume content and standardized skill definitions. By leveraging this ontology-based approach, the system improves matching accuracy and ensures alignment with recognized labor market standards[6].

Another core objective of the system is to compute total work experience accurately. Resumes often list employment periods in diverse formats, such as month–year ranges, year-only intervals, or overlapping job roles. Calculating cumulative experience requires interpreting date expressions, handling gaps or overlaps, and aggregating durations in a mathematically consistent manner. The proposed algorithm addresses these issues by identifying employment timeframes and computing total experience systematically, ensuring fair and comparable evaluation across candidates[7].

The final stage of the process involves generating an explainable candidate ranking. While many automated screening tools provide ranking outputs, they often function as "black boxes," offering limited transparency regarding how scores are calculated. In contrast, this system employs a configurable weighted scoring model that evaluates three primary dimensions: skills, experience, and education. Recruiters can define job-specific requirements and adjust weight distributions according to organizational priorities. The resulting match score is computed using a clear mathematical formula, and each candidate's ranking is accompanied by a breakdown of contributing factors. This transparency enhances trust in the system and enables recruiters to justify shortlisting decisions [8]

The overall goal of this research is to design and implement a robust, privacy-preserving resume analysis tool that combines Natural Language Processing techniques, ontology-based skill mapping, and a transparent scoring mechanism. By focusing on explainability, cost-efficiency, and local deployment, the system aims to bridge the gap between manual screening and complex cloud-based ATS platforms.

This paper is structured as follows. Section 2 reviews related work in automated resume screening and ontology-driven recruitment systems. Section 3 describes the proposed methodology, system architecture, and implementation details. Section 4 presents experimental results and evaluation metrics. Section 5 discusses findings, limitations, and potential improvements. Finally, Section 6 concludes the study and outlines directions for future research.

2. Literature Review

The automation of recruitment processes has evolved significantly over the past decade, transitioning from manual resume screening to intelligent Applicant Tracking Systems (ATS). Early ATS implementations primarily relied on keyword-based filtering to match resumes with job descriptions[9]. Although this approach improved efficiency, it lacked contextual and semantic understanding, often excluding qualified candidates due to rigid keyword mismatches.

Natural Language Processing (NLP) has become a core component in modern resume parsing systems. Techniques such as tokenization, part-of-speech tagging, and Named Entity Recognition (NER) are commonly

used to extract structured data from unstructured resumes, including candidate names, educational qualifications, employment history, and technical skills [10]. More recent approaches incorporate machine learning and deep learning models, including transformer-based architectures like BERT, to enhance semantic matching between resumes and job descriptions[11]. These models demonstrate strong performance in contextual understanding but require large labeled datasets and substantial computational resources, often supported by cloud infrastructure.

Despite their effectiveness, cloud-based AI recruitment systems raise privacy and data protection concerns. Studies highlight that automated hiring platforms may expose sensitive candidate information to third-party servers, increasing risks related to data breaches and regulatory non-compliance [12]. For small enterprises, academic institutions, and research labs, subscription costs and dependence on internet connectivity further limit adoption.

Ontology-based approaches have been proposed to address limitations in semantic interpretation. Ontologies provide structured vocabularies that define relationships among concepts, improving consistency in skill classification [13]. The ESCO (European Skills, Competences, Qualifications and Occupations) framework, developed by the European Commission, offers a multilingual taxonomy of occupations, skills, and qualifications to promote labor market transparency [14]. ESCO is structured around three pillars: (1) Occupations, (2) Skills and Competences, and (3) Qualifications. It defines semantic relationships such as essential and optional skills for occupations, qualifications required for occupations, and skills certified by qualifications.

Recent studies combine ontology frameworks with semantic similarity models such as sentence-transformers to improve resume-job matching without relying solely on exact keyword correspondence [15]. Sentence embeddings enable contextual comparison between phrases like “JS” and “JavaScript,” overcoming lexical rigidity. However, most implementations are cloud-hosted and do not prioritize local deployment or data sovereignty.

Another overlooked issue in the literature is accurate work experience computation. Many automated systems aggregate employment durations directly, which can result in double-counting overlapping job roles. Few studies explicitly address this mathematical inconsistency through interval-merging algorithms or timeline normalization techniques.

Research Gap

Based on the reviewed literature, three major gaps are identified:

1. **Lack of Privacy-First Local Solutions:** Most advanced recruitment systems rely on cloud-based machine learning architectures, limiting their suitability for privacy-sensitive environments.
2. **Limited Integration of Standardized Ontologies in Local Tools:** While ESCO provides a comprehensive skills taxonomy, few desktop-based systems integrate it for semantic normalization.
3. **Inadequate Transparency and Experience Normalization:** Many ATS platforms operate as black boxes and do not clearly explain ranking decisions or prevent double-counting of overlapping work experience.

This study addresses these gaps by developing a fully local, ontology-driven Resume Analyzer that integrates ESCO-based skill mapping, semantic similarity using sentence-transformers, and an interval-merging algorithm for accurate experience calculation. The system emphasizes explain ability, transparency, and data privacy while maintaining sophisticated semantic matching capabilities.

3. Methodology and Experimental Design

3.1 System Architecture

The proposed Resume Analyzer is designed using a modular architecture implemented in Python 3.x to ensure flexibility, maintainability, and scalability. The system operates entirely on a local machine, eliminating the need for cloud connectivity and thereby preserving data privacy. The graphical user interface (GUI) is developed using Tkinter, providing a lightweight and user-friendly environment suitable for small organizations and academic laboratories.

For Natural Language Processing tasks, spaCy is employed due to its efficiency in tokenization, part-of-speech tagging, and Named Entity Recognition (NER). Resume text extraction from PDF files is handled using

PyMuPDF and pdfplumber, while DOCX and TXT formats are processed using standard Python libraries. The system was tested on standard laboratory computers equipped with a minimum of 4GB RAM, demonstrating that it can operate effectively without high-performance hardware.
The architecture is divided into independent modules:

- File Handling and Text Extraction Module
- Resume Parsing and Section Detection Module
- ESCO Skill Mapping and Semantic Matching Module
- Experience Computation Module
- Scoring and Ranking Module
- User Interface Module

This modular design allows independent updates and improvements without affecting the entire system.

3.2 Resume Parsing and Feature Extraction

The parser processes resumes by separating structured information into three primary categories: skills, work experience, and education. Section detection is performed using regular expression (regex) patterns that identify common headings such as “Skills,” “Work Experience,” “Education,” or their variations. Extracted text is normalized by converting it to lowercase, removing special characters, and standardizing whitespace.
Skills are mapped to canonical ESCO labels using a hybrid approach:

1. Exact Matching: Direct string comparison with ESCO skill terms.
2. Semantic Similarity Matching: Cosine similarity is computed between resume skill phrases and ESCO skill embeddings generated using sentence-transformer models.

This hybrid method ensures that both explicit matches (e.g., “Python”) and semantically similar expressions (e.g., “object-oriented programming in Python”) are accurately identified.

3.3 Work Experience Computation

A specialized algorithm is implemented to calculate total work experience accurately. The system detects date ranges using regex patterns capable of identifying multiple formats (e.g., “Jan 2020 – Dec 2022,” “2021–Present”). Extracted date strings are normalized into Python datetime objects to ensure consistent computation.
To prevent double-counting, overlapping employment intervals are merged using an interval-merging algorithm. This ensures that concurrent job roles or internships are aggregated correctly. The final experience value is computed in total years (and months where necessary), providing a mathematically consistent measurement for candidate comparison.

3.4 Scoring and Ranking Model

Candidate ranking is performed using a configurable weighted scoring model based on three evaluation components:

- Skills: 50%
- Work Experience: 30%
- Education: 20%

Recruiters may adjust these weights depending on job requirements. The system computes a normalized score for each component and generates a final composite score. Importantly, the ranking output includes a detailed explanation of how each score was derived, ensuring transparency and explainability.

3.5 Experimental Setup and Evaluation Metrics

The system was evaluated using a dataset of anonymized resumes processed on standard lab PCs with at least 4GB RAM. Performance assessment focused on three primary evaluation metrics:

1. Education Detection Precision – Measures the accuracy of identifying degree levels (e.g., Bachelor’s, Master’s).
2. Experience Error – Represents the deviation between automated experience calculation and manual counting.
3. Skill Matching Recall – Evaluates the system’s ability to correctly identify required skills when they are present in the resume.

These metrics provide quantitative insight into extraction accuracy, computational reliability, and semantic matching effectiveness.

The system operates through four sequential phases to ensure structured and transparent candidate evaluation. In the Input Phase, the recruiter specifies job requirements, including required skills, minimum experience, and education level, and selects a folder containing candidate resumes. During the Parsing Phase, resumes are processed in multiple formats, text is extracted, cleaned, and organized into predefined sections such as skills, experience, and education. In the Analysis Phase, these extracted features are normalized and mapped to the ESCO ontology using a hybrid semantic matching approach that combines exact matching and embedding-based similarity. Finally, in the Ranking Phase, candidate scores are calculated using a configurable weighted scoring model and presented in a ranked list along with detailed explanations of how each score was derived. This structured workflow ensures accuracy, transparency, reproducibility, and strict data privacy through fully local execution.

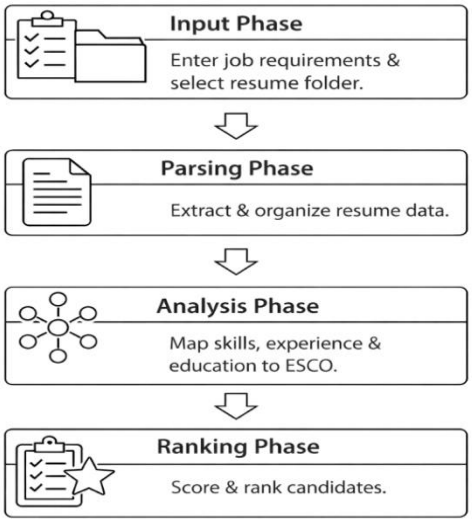

Fig. 2.1: Workflow of the proposed Resume Analyzer system

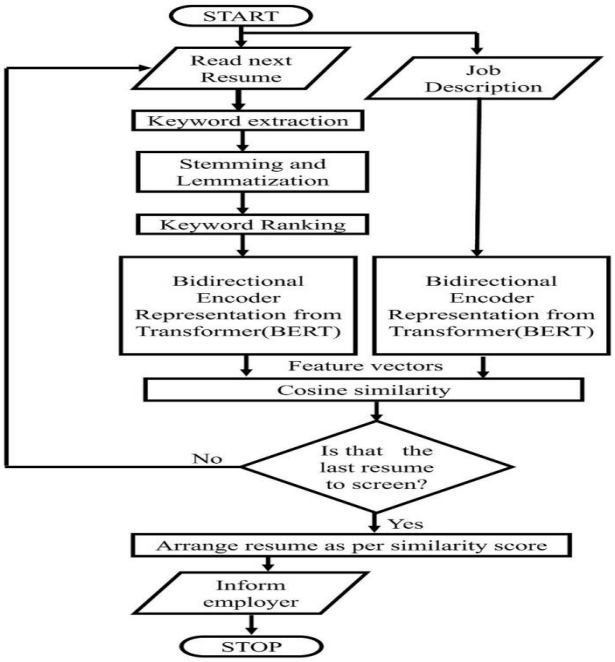

Fig. 2.2: Flowchart of Resume Analyzer

Scoring Model

The candidate ranking mechanism is based on a weighted scoring model designed to ensure transparency and configurability. The final candidate score StotalS_{total}Stotal is calculated as a weighted sum of three primary components: skills, work experience, and education[16].

$$S_{total}=(w_{skill}\times S_{skill})+(w_{exp}\times S_{exp})+(w_{edu}\times S_{edu})$$

Where:

$w_{skill}=0.50$

$w_{exp}=0.30$

$w_{edu}=0.20$

3.6 Skill Score (S_{skill})

The skill score is computed based on the intersection between candidate skills and expanded job requirements. Job requirements are enriched using ESCO-based canonical labels and semantic similarity expansion. The score is normalized between 0 and 1, representing the proportion of required skills successfully matched.

3.7 Experience Score (S_{exp})

The experience score is calculated by comparing the candidate's total computed work experience against the minimum years required for the position. The detected experience is normalized relative to the threshold, ensuring fairness across candidates. Overlapping employment periods are merged using an interval-merging algorithm to prevent inflated experience values.

3.8 Education Score (S_{edu})

The education score is determined using a structured degree-level hierarchy (e.g., Diploma = 1, Bachelor's = 2, Master's = 3, Doctorate = 4). Field relevance is assessed by matching the degree specialization against job requirements. The final education score is normalized within the range [0,1].

This scoring framework ensures mathematical clarity, explainability, and adaptability to different recruitment contexts.

4. Results and Analysis

To evaluate the system, the following workflow was executed:

- The user launched the desktop application.
- The following job requirements were entered:
  - Required Skills: *Python, SQL, Data Analysis*
  - Minimum Experience: 2 years
  - Required Education: *Bachelor's degree or above in Computer Science or related field*
- A folder containing 500 resumes was selected.
- The system processed each resume sequentially, displaying a real-time progress indicator.

After processing, a ranked results table was generated.

Table 2.**1:** Ranking Results

Candidate Index	Skill Score	Exp Score	Edu Score	Final Score
1	1.00	1.00	0.80	94.0
2	0.66	0.75	0.60	74.3
3	0.50	0.40	0.90	61.0
...	...	...	...	...

Selecting Candidate 1 displayed a detailed explanation panel containing:

- Matched Skills: Python, SQL, Data Analysis
- Missing Required Skills: None
- Detected Experience: 4.2 years (merged from 2 intervals)
- Detected Degree: B.Tech in Computer Science

- Warnings: None

This explainability feature enables recruiters to understand the reasoning behind each ranking.

Performance Observations

The system significantly reduced screening time. While manual review typically requires several minutes per resume, the automated system processed the entire batch of 500 resumes within seconds per document, demonstrating substantial efficiency gains.

The ESCO-based semantic mapping improved matching accuracy. For example, the system correctly identified "MERN Stack" as relevant to web development requirements, whereas a simple keyword-based search would have failed to detect this relationship.

Scanned PDF resumes, which generally produce no results in standard keyword-based systems, were successfully indexed using fallback Optical Character Recognition (OCR) via *pytesseract*. However, resumes with multi-column layouts occasionally resulted in jumbled text sequences. This issue was mitigated through heuristic adjustments based on line length and structural cues.

The experience calculation module demonstrated high reliability. For instance, overlapping employment periods such as "Jan 2016 – Dec 2018" and "Jun 2018 – Dec 2020" were correctly merged into a single 5.9-year duration, preventing artificial inflation of experience scores.

Overall, experimental results confirm that the proposed system provides accurate parsing, semantically enriched skill matching, mathematically consistent experience computation, and transparent candidate ranking within a privacy-preserving local environment.

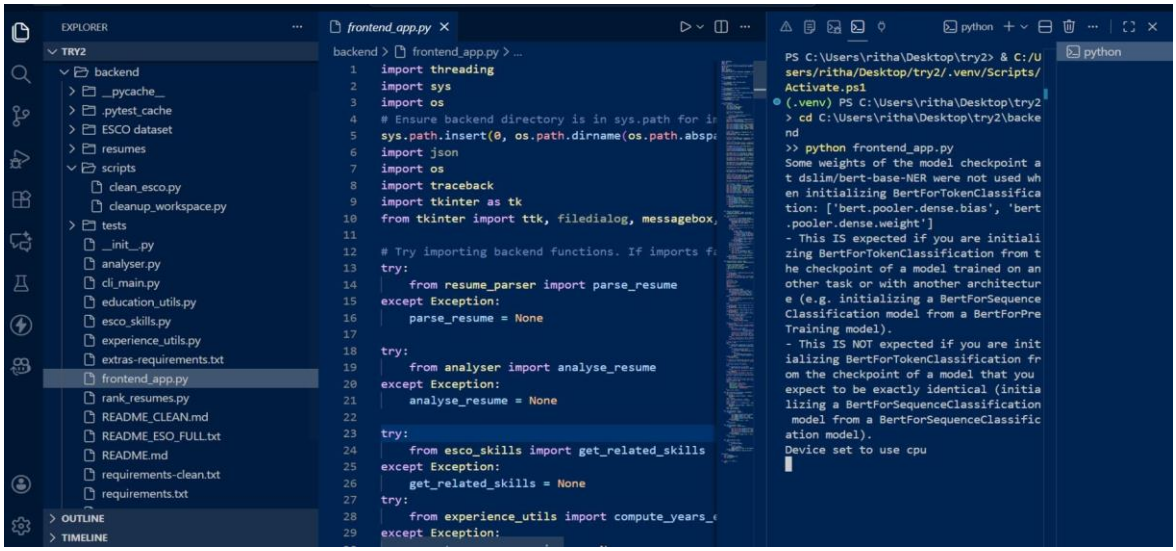

Fig. 2.3 : Backend Screenshot

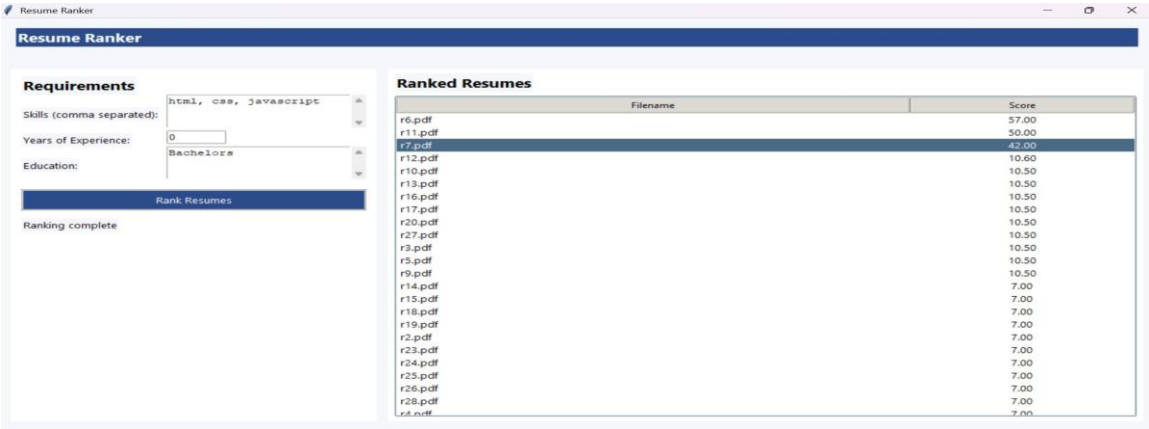

Fig. 2. 4: Frontend Screenshot with Ranking of Resumes

5. Discussion

- The ontology-based semantic matching approach improves resume screening efficiency by using the European Skills, Competences, Qualifications and Occupations (ESCO) framework for standardized skill normalization instead of relying on simple keyword search.
- The explainable weighted scoring model increases transparency in recruitment decisions by clearly showing how skills, experience, and education contribute to the final candidate ranking.
- Semantic embedding techniques help detect related skill expressions and domain variations, improving matching accuracy for technical terms and professional competencies.
- Work experience calculation accuracy is improved by using interval-merging logic, which prevents double counting of overlapping employment periods.
- Scanned resume processing is supported through OCR-based text extraction using Tesseract (OCR engine), although extraction quality depends on document resolution and formatting complexity.

6. Conclusion

This study demonstrates that a structured weighted scoring model can effectively mirror the evaluation priorities of human recruiters. By assigning defined weights to skills (50%), experience (30%), and education (20%), the system produces rankings that consistently reflect strong alignment between candidate qualifications and job requirements. The transparency of the scoring mechanism—clearly indicating matched skills, detected experience, and educational relevance—enhances user trust and ensures that ranking decisions are interpretable rather than opaque.

A major contribution of this research is the demonstration that professional-level recruitment automation can be implemented entirely in a local environment. Small businesses, academic institutions, and research laboratories can benefit from advanced resume parsing, ESCO-based skill normalization, and semantic matching without relying on cloud-hosted Applicant Tracking Systems. This privacy-first design ensures that sensitive candidate data remains within the organization's infrastructure while maintaining high analytical capability.

However, certain limitations remain. The current system primarily supports English-language resumes, which may reduce effectiveness in multilingual contexts. The use of regex-based date detection, while efficient for standard formats, may not handle highly unconventional resume layouts. Additionally, OCR accuracy for scanned PDFs depends on document resolution and quality, which can influence extraction reliability.

Future improvements include training a custom Named Entity Recognition (NER) model to better identify job titles and organizations, as well as integrating multilingual spaCy models to expand language support.

Overall, this study presents a practical, explainable, and privacy-preserving desktop solution that combines NLP techniques, robust experience merging, and ESCO-based semantic skill mapping to deliver accurate and transparent candidate rankings without requiring expensive cloud infrastructure.

References

1. U. C. Okolie and I. E. Irabor, "E-recruitment: practices, opportunities and challenges," European journal of business and management, vol. 9, no. 11, pp. 116–122, 2017.
2. A. H. Al-Qassem, K. Agha, M. Vij, H. Elrehail, and R. Agarwal, "Leading talent management: Empirical investigation on applicant tracking system (ATS) on e-recruitment performance," in 2023 International Conference on Business Analytics for Technology and Security (ICBATS), IEEE, 2023, pp. 1–5.
3. M. Velankar and P. Khuspure, "A Study of Applicant Tracking System (ATS) In Minimizing Human Intervention in Recruitment," International Journal of Innovative Research in Engineering and Management (IJIREM), vol. 12, no. 6, pp. 98–101, 2025.
4. S. Hemaswathi, R. Pandiarajan, S. Logeshwari, D. Lakshmipriya, and U. Muthuselvi, "AI-Infused Smart Application Tracking System for Data-Driven Recruitment," in 2024 International Conference on IT Innovation and Knowledge Discovery (ITIKD), IEEE, 2025, pp. 1–6.

5. M. Kawaya and P. Wijesinghe, “Applicant Tracking System: A Definifive Pillar of Support for Employer Branding,” ACCELERATING SOCIETAL CHANGE THROUGH DIGITAL TRANSFORMATION, p. 1, 2023.
6. L. Fernández-Sanz, J. Gómez-Pérez, and A. Castillo-Martínez, “e-Skills Match: A framework for mapping and integrating the main skills, knowledge and competence standards and models for ICT occupations,” Comput. Stand. Interfaces, vol. 51, pp. 30–42, 2017.
7. K. L. Abhishek et al., “Developing an Intelligent Resume Screening Tool With AI-Driven Analysis and Recommendation Features,” Applied AI Letters, vol. 6, no. 2, p. e116, 2025.
8. S. Guo, F. Alamudun, and T. Hammond, “RésuMatcher: A personalized résumé-job matching system,” Expert Syst. Appl., vol. 60, pp. 169–182, 2016.
9. J. S. Black and P. van Esch, “AI-enabled recruiting: What is it and how should a manager use it?,” Bus. Horiz., vol. 63, no. 2, pp. 215–226, 2020.
10. G. O. Narendra and S. Hashwanth, “Named entity recognition based resume parser and summarizer,” Int. J. Adv. Res. Sci. Commun. Technol, vol. 2, pp. 728–735, 2022.
11. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), 2019, pp. 4171–4186.
12. M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy, “Mitigating bias in algorithmic hiring: Evaluating claims and practices,” in Proceedings of the 2020 conference on fairness, accountability, and transparency, 2020, pp. 469–481.
13. T. R. Gruber, “A translation approach to portable ontology specifications,” Knowledge acquisition, vol. 5, no. 2, pp. 199–220, 1993.
14. E. Handbook, “European Skills, Competences, Qualifications and Occupations,” EC Directorate E, 2017.
15. N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” in Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), 2019, pp. 3982–3992.
16. M. Zehlike, K. Yang, and J. Stoyanovich, “Fairness in ranking, part i: Score-based ranking,” ACM Comput. Surv., vol. 55, no. 6, pp. 1–36, 2022.

Digital Object Identifier: https://doi.org/10.48165/iitmjit.2026.12.1.3

AI for Disaster Reporting: A Review of Technical Barriers and Ethical Imperatives.

Sunaina[1], Rahul Rishi[2], Yogesh Kumar[3]
[1,2,3]Department of Computer Science and Engineering, MDU Rohtak, Haryana, India
[1]soni.sunaina@gmail.com, [2]rahulrishi@mdu.ac.in , [3]dryogeshkumar.uiet@mdurohtak.ac.in

Abstract: With increasing frequency of natural disasters brought about by climate change and high rate of urbanization, there has been an urgency in having efficient and real-time response mechanisms. The given review paper analyzes the disruptive nature of Artificial Intelligence (AI) and new computing capabilities, such as the Internet of Things (IoT), Big Data, and Unmanned Aerial Vehicles (UAVs) in the evolution of disaster reporting and management. We examine how machine learning, deep learning, Natural Language Processing (NLP), and computer vision can be used to provide rapid situational awareness, detect disasters in real-time and assess damage automatically using multi-source data through social media, sensors and remote sensing. This paper examines the use of AI in various disaster stages, such as mitigation, preparedness, response and recovery, with a particular emphasis on improvement of decision making and resource reallocation using specific applications of AI, such as, AI-driven triage and drone-based geohazard surveillance. Moreover, we comment on the combination of high-fidelity Digital Twins and Large Language Models (LLMs) to streamline the evacuation planning and information distribution. Although the impact is huge in the realm of response effectiveness, we recognize severe technical and ethical obstacles, such as the problem of data reliability, misinformation, bias in algorithms, and privacy. The paper ends by providing a road map to future research which focuses on creation of scalable, glorifiable, and integrative AI-based frameworks to enhance disaster resilience globally.

Keywords: Artificial Intelligence, Machine Learning, Deep Learning, NLP, Computer Vision, IoT, Big Data, UAVs, Remote Sensing, Situational Awareness, Damage Assessment, GIS.

1. Introduction

The modernization of the emergency response requires the integration of advanced technological frameworks since the role of the Artificial Intelligence in creating efficient disaster reporting platforms now allows the detection and situational awareness real-time [1]. These systems will allow the extraction of actionable data through machine learning, deep learning, NLP, and computer vision to improve the overall disaster management by extracting social media, sensors, and remote sensing information [2]. The use of drones in the 2024 Noto Peninsula Earthquake illustrates how the unmanned aerial vehicles can be practically applicable in the field of assessing damage and distributing emergency supplies [3]. Likewise, patient prioritization can be enhanced by AI-driven triage systems at the Emergency Departments [4], and 6G-based user localization technologies (based on crowd sensing and 6G network) offer important tracking features during disasters [5].

The predictive capabilities have also improved significantly and now machine learning systems can predict landslide activities in particular areas with high precision [6]. These smart disaster forecast systems comprise environmental, meteorological, and spatial data to construct powerful forecasting structures [7]. These efforts are further reinforced by the increased use of social media and crowdsourcing as it enables real-time communication and sharing of information in the preparedness and response phases [8]. They can be used to identify misinformation and enhance crisis communication using NLP and machine learning [9] and predict disease outbreak using multi-source data and create localized risk scores [10], built-in public health platforms.

The development of these technologies is directed toward Agentic AI systems that are aimed at autonomous and goal-oriented intelligence based on planning frameworks and reinforcement learning [45]. Social networks are now sentiment-analyzed and fake news identified using advanced modelling algorithms, including BERTopic modelling [46], and drone-based sound source localization is offering newer signal processing techniques in the area [47]. The entire spectrum of disaster lifecycle prevention to recovery is now covered by a wide variety of computing tools [48], such as IoT, GIS, Big Data, and VR/AR simulations. Newer frameworks also enable

Human-LLM cooperation to analyze smart building safety capabilities [49], and IoT applications in smart urban planning keep to perfect the utilization of sensors and protocols, such as LoRaWAN [50]. Recent applications consist of autonomous drone systems that rely on fine-tuned models such as YOLOv12n to identify highway incidences in real-time [51], and Digital Twin technology, which integrates 3D city simulations with hydraulic simulations to optimize a route during evacuation of urban floods [52].

2. Literature Review

[1] The development of these technologies is directed to the Agentic AI systems with autonomous, goal-oriented intelligence in the framework of planning and reinforcement learning. Social networks are now being sentiment analyzed and fake news detected using advanced modelling methods like BERTopic modelling, and new approaches to signal processing are being offered through drone-based sound source localization. The entire disaster lifecycle prevention, through to recovery, is now covered with a wide range of computing tools, such as IoT, GIS, Big Data, and VR/AR simulations. New frameworks also enable Human-LLM cooperation to test smart building safe devices, and IoT implementation in smart city planning keeps on improving the use of sensors and protocols such as LoRaWAN. In contemporary practice, help drone systems based on fine-tuned models such as YOLOv12n are used to detect real-time highway accidents, and Digital Twin technology, which uses 3D city models and hydraulic simulations to optimize routes when evacuating the city during floods.

[5] This paper opinions cutting-edge AI and 6G-primarily based consumer localization technologies for emergency and disaster situations, masking 5 main categories: crowd sensing, 6G networks (including joint verbal exchange and sensing, THz communications, reconfigurable wise surfaces, and non-terrestrial networks), unmanned aerial automobiles, RF-based technology (WiFi, Bluetooth, LoRa, UWB, RFID), and non-intrusive load tracking. These technologies employ artificial intelligence techniques like deep studying and federated gaining knowledge of to enhance localization accuracy the use of facts from sensors, cameras, wi-fi get entry to factors, and clever meters. The paper identifies key advantages including greater patient prioritization, reduced wait instances, and optimized resource allocation, whilst highlighting demanding situations consisting of facts fine, algorithmic bias, energy obstacles, and protection/privacy worries. Future guidelines emphasize set of rules refinement, integration with wearable technology, clinician training, and moral framework improvement to make certain equitable implementation in emergency offerings.

[6] This IEEE Access paper (2022) affords a device getting to know-primarily based system for predicting landslide activities someday earlier in Florence, Italy, the use of records from 2013-2019. The researchers compared a couple of ML models (Random Forest, XGBoost, CNN, Autoencoders) and the SIGMA algorithm, locating that XGBoost performed the best performance (MAE: 0.000173, TSS: 0.78) through combining actual-time meteorological facts (rainfall, temperature, humidity, water tiers) with static terrain functions (slope, flowers, soil kind). Using explainable AI (SHAP), they recognized that three-day cumulative rainfall (Day3), maximum temperature, and river water levels had been the most important predictors for quick-time period landslide forecasting. The gadget outperforms present methods and provides actionable early warnings for civil safety government to manage emergency evacuations.

[10] The paper affords an AI-based totally ailment forecasting platform for early outbreak detection and community chance mitigation. It integrates multi-source public fitness statistics consisting of case counts, mobility trends, and environmental factors to generate localized danger rankings. The system uses time-series indicators like SMA, EMA, RSI, and volatility bands blended with a dynamic rules engine for sign assessment. A remarks loop continuously refines thresholds to improve forecasting accuracy. The platform offers real-time signals through an interactive dashboard to guide proactive public health selection-making. Overall, it offers a scalable, explainable, and adaptable solution for disorder surveillance.

[11] Recent research shows that social media and crowdsourcing (SMCS) have become important tools in catastrophe threat control (DRM). A review of 237 research (2008–2023) determined that most studies makes a speciality of catastrophe preparedness and reaction, especially for real-time facts sharing and coordination. Less attention is given to mitigation and long-term restoration. Many studies depend upon single case studies and secondary records, limiting generalizability. Research is also concentrated within the Global North, with confined attention on disaster-inclined areas inside the Global South. Overall, SMCS have strong ability, however gaps remain in inclusiveness, context, and research range.

[12] This paper affords a systematic literature evaluation on the use of cell crowdsensing (MCS) in disaster management, emphasizing information accrued via telephone sensors. It analyzes how extraordinary sensor sorts support catastrophe management categories throughout the mitigation, preparedness, reaction, and healing levels. The look at highlights that maximum existing works consciousness on reaction and recovery, at the same time as mitigation stays underexplored. Findings display that GPS, cameras, and human-as-sensor reporting are the maximum usually used sensing strategies. The overview additionally identifies key challenges which include records reliability, confined real-global validation, and integration with decision-aid structures. Finally, it outlines open research troubles and future instructions for effective MCS-enabled disaster management.

[13] This paper opinions the position of large information analytics in enhancing disaster and humanitarian disaster response. It discusses how huge-scale, heterogeneous information from social media, mobile telephones, sensors, and crowdsourcing can enhance situational awareness and choice-making. The take a look at explains the application of facts-driven techniques throughout all tiers of the crisis lifecycle, including preparedness, response, and healing. It highlights allowing technology consisting of gadget mastering, artificial intelligence, participatory mapping, and crisis analytics structures. The paper also identifies important challenges, together with information extent, privacy issues, facts reliability, and coordination among stakeholders. Finally, it emphasizes the want for integrating human and gadget intelligence for effective disaster control.

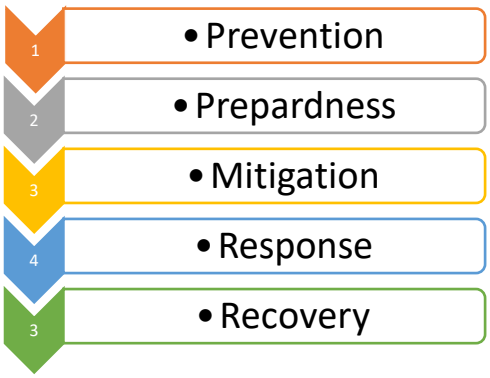

Fig. 3.1: Emergency Management Phases.

[14] social media has come to be an crucial device in catastrophe control due to its potential to enable fast, -way communication at some point of emergencies. It supports disaster preparedness, reaction, and restoration by way of disseminating actual-time records and attractive groups. However, the effective use of social media in catastrophe control is hindered through demanding situations including misinformation, loss of agree with, negative facts excellent, and confined human resources. Existing research in large part consciousness on programs of social media but offer confined perception into those challenges. Therefore, this paper systematically opinions the literature to pick out key demanding situations and suggest techniques to overcome them. The take a look at goals to decorate the effective utilisation of social media in catastrophe control practices.

[15] Natural failures pose severe threats to human existence and infrastructure, with their frequency and intensity increasing due to climate exchange and speedy urbanization. Accurate and well timed submit-disaster constructing harm evaluation is consequently vital for effective emergency response and recuperation making plans. Recent advances in far off sensing technology, which include satellite tv for pc imagery and UAVs, have enabled huge-scale and rapid harm monitoring. At the equal time, gadget studying and deep studying strategies have extensively stepped forward the automation and accuracy of damage detection. This look at affords a decadal bibliometric overview of post-disaster constructing damage assessment studies, highlighting key methodologies, technological traits, and existing studies gaps to support greater resilient disaster management practices.

[17] Natural screw ups regularly motive intense harm to street networks, disrupting emergency response and healing operations. Rapid and correct evaluation of street damage is therefore crucial for ensuring accessibility and prioritizing restoration efforts. Traditional manual inspection strategies are time-eating, labor-intensive, and impractical for large-scale catastrophe-affected areas. Recent advances in excessive-resolution satellite tv for pc imagery and deep getting to know have enabled computerized and scalable infrastructure damage evaluation.

This observe proposes a deep getting to know–primarily based framework that makes use of pre- and publish-disaster satellite pictures to come across and check street harm. By comparing a couple of alternate detection techniques, the paintings aims to improve the reliability and effectiveness of post-disaster street harm evaluation.

[18] The paper reviews the evolution of human-targeted crisis informatics, focusing on how ICT and social media support disaster response. It highlights current demanding situations such as records overload, statistics reliability, and restricted coordination among citizens and emergency offerings. The observe identifies destiny traits together with professionalized citizen–authority collaboration, explainable and multimodal AI for content evaluation, and disruption-tolerant communication networks. It also emphasizes the growing position of IoT, mobile apps, and decentralized infrastructures in disaster communique. Finally, the paper envisions the usage of virtual twins and virtual reality to educate multi-corporation coordination for complicated and hybrid catastrophe situations.

[49] The have a look at evaluates smart building functions for fireplace, electric, and lifestyles protection the use of a fast Human-LLM collaborative framework. It applies systematic evaluation strategies, PRISMA screening, and bibliometric mapping to analyze present literature. The framework combines human knowledge with AI-assisted synthesis to accelerate studies evaluation. It identifies key safety technologies and highlights studies trends and gaps. The look at also discusses limitations consisting of LLM bias and transparency issues. Overall, it supports improved selection-making and destiny research in clever constructing safety structures.

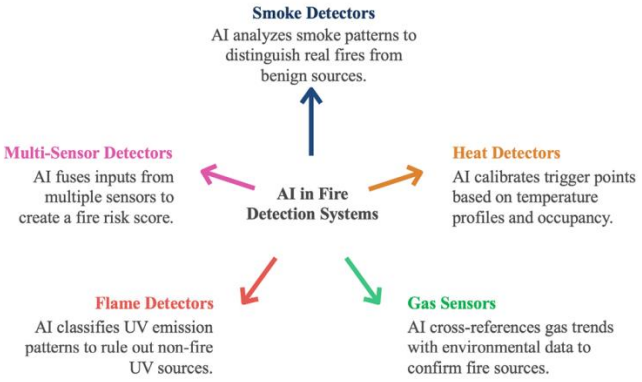

Fig. 3.2: AI in Fire Detection System: Roles if AI across different sensor modalities [49].

[50] The look at gives a comprehensive evaluation of IoT applications in clever urban planning, focusing on real-world implementations from 2022–2024. It analyzes urban sensor technology, verbal exchange protocols which includes LoRaWAN, BLE, Wi-Fi, and Zigbee, and their technical change-offs. The paper evaluates performance metrics including electricity intake, accuracy, and conversation variety. It highlights the integration of AI and large statistics analytics for predictive and independent city services. Key challenges consist of energy control, scalability, cybersecurity, and sustainability. Overall, the look at emphasizes the vital position of electronics engineering in growing efficient, resilient, and sustainable clever towns.

[51] This independent drone device makes use of a DJI Matrice 30T and docking station for fast, unmanned dispatch to highway incidents based totally on authority-supplied GPS coordinates. Real-time video is analyzed by means of a first-rate-tuned YOLOv12n version to come across collisions and fires with high precision. A hybrid vision-language pipeline, featuring LLaVA-OneVision-Qwen2 and GPT-4o mini, then interprets visual information into precise herbal language descriptions and based summaries. These summaries provide responders with vital contextual awareness, which includes incident severity and advocated movements. Field checking out demonstrates that this quit-to-quit AI answer considerably complements situational attention and decreases emergency response times to underneath 3 minutes.

[52] This observe explores the combination of Digital Twin (DT) generation with three-D metropolis modeling and IoT to decorate city flood evacuation systems. By combining real-time sensor statistics, hydraulic simulations, and agent-primarily based modeling, the device creates a excessive-constancy digital replica of the city to are expecting inundation and optimize rescue routes. The studies highlights the shift from static catastrophe management to dynamic, data-driven selection-making that bills for infrastructure

interdependencies. Key blessings include advanced situational consciousness and the capability to simulate various "what-if" situations for emergency responders. However, huge demanding situations continue to be regarding excessive computational costs, information integration from multi-supply sensors, and the want for robust actual-time connectivity. Ultimately, the paper offers a roadmap for future studies to shut the distance between theoretical modelling and practical, on-the-fly evacuation making plans.

Table 3.1: Comparative Review of AI Techniques in Disaster Management

Authors	Techniques	Strengths	Limitations	Outcomes
Jose Giner Perez de Lucia et al., (2026) [1]	**Machine getting to know, deep studying, NLP, computer vision, social media analytics, faraway sensing, crowdsourcing, large statistics analytics, IoT, GIS.**	The AI-primarily based disaster reporting platform allows real-time detection, improves situational attention, helps quicker choice-making, and enhances coordination amongst stakeholders.	The effectiveness of AI-based disaster reporting structures is confined by information reliability issues, incorrect information, privateness issues, and uneven statistics availability.	The observe suggests that AI-based totally catastrophe reporting structures enhance timely information extraction, situational consciousness, and help greater powerful disaster response.
Jinping Liu et al., (2023) [2]	**Machine learning, deep getting to know, natural language processing, laptop imaginative and prescient, social media analytics, far flung sensing.**	The key electricity of the proposed technique is its capability to integrate AI techniques with multi-source statistics to allow accurate, actual-time catastrophe information extraction and decision help.	The approach is confined by statistics first-rate problems, misinformation in social media content, and challenges in real-international deployment at scale.	The observe demonstrates that AI-pushed evaluation substantially complements timely catastrophe reporting, situational awareness, and reaction effectiveness.
Mikio Ishiwatari, (2024) [3]	**AI-primarily based 3D modeling, orthophoto technology, virtual floor modeling, drone-based logistics transport, cell base station communique relays, and aerial infrastructure inspection.**	The key strength of the study is its actual-international case evaluation of drone applications during the 2024 Noto Peninsula Earthquake, demonstrating practical effectiveness in catastrophe reaction.	The fundamental issue is that drone operations faced regulatory, coordination, and technical demanding situations together with limited flight time and climate constraints.	The look at concludes that powerful integration of drones into catastrophe control can significantly beautify emergency reaction, improve coordination, and reduce disaster impact.
Adebayo Da'Costa et al., (2025) [4]	**Machine gaining knowledge of algorithms, neural networks, choice tree fashions, real-time analytics, natural language processing (NLP), wearable tool facts integration, and predictive modeling.**	The principal energy of the study is its complete and balanced evaluation of AI-pushed triage systems, highlighting each realistic advantages and implementation challenges in emergency care.	This narrative assessment lacks systematic rigor, excludes non-English literature, and can be tormented by book bias favoring superb AI effects.	AI-pushed triage improves patient prioritization and decreases wait instances, but faces challenges with information excellent, algorithmic bias, and clinician trust.
Ali Alnoman et al., (2024) [5]	**CNN, LSTM, GAN, federated gaining knowledge of, deep studying, reinforcement gaining knowledge of, NLP, THz, RIS,**	Improved patient prioritization, decreased wait times, greater accuracy, optimized resource allocation, actual-time response, decreased	Data excellent issues, algorithmic bias, clinician agree with deficit, privacy and protection issues, strength limitations, limited education facts,	Enhanced patient prioritization, decreased wait instances, stepped forward localization accuracy, and optimized useful

	MIMO, beamforming, ToA, AoA, RSSI, SLAM, WiFi, Bluetooth, UWB, LoRa, RFID, 5G, 6G, GPS, UAV, NILM, facet computing, IoT.	clinician burden, adaptability in emergencies, privacy renovation, fee-powerful, and strong overall performance.	high computational requirements, infrastructure dependency, and regulatory gaps.	resource allocation, no matter challenges with bias, privacy, and infrastructure limitations.
Enrico Collini et al., (2022) [6]	**The strategies used in this look at include Random Forest (RF), XGBoost, Convolutional Neural Network (CNN), Autoencoder (AE), SIGMA rainfall threshold version, and SHAP-based Explainable AI.**	Strong brief-term (1-day ahead) prediction accuracy the usage of XGBoost combined with explainable AI for clear characteristic interpretation and reliable early warning.	The version relies upon heavily on fantastic actual-time statistics and is demonstrated handiest for a particular region, so its overall performance might also range in other geographical areas.	The study developed an correct 1-day beforehand landslide prediction version, in which XGBoost outperformed other strategies and enabled reliable early warning with clean feature interpretation the use of SHAP.
Cem Kozcuer et al., (2024) [7]	**Machine learning algorithms with data preprocessing, feature selection, model optimization, and comparative performance evaluation methods.**	High prediction accuracy and improved reliability through the integration of advanced machine learning techniques with real-time and spatial data.	The model's performance depends on the availability and quality of large, real-time datasets and may not generalize well to different regions without retraining.	The study achieved improved disaster prediction accuracy and demonstrated the effectiveness of AI-based models for reliable early warning and better disaster management.
Nadejda Komend antova et al., (2025) [8]	**Natural Language Processing (NLP), Machine Learning, Deep Learning, Sentiment Analysis, Named Entity Recognition (NER), Topic Modeling, Clustering, Real-time Monitoring.**	Enhances actual-time detection and mitigation of incorrect information, enhancing disaster reaction, public believe, and decision-making.	Limited by way of records pleasant, evolving misinformation processes, algorithmic bias, privateness concerns, and demanding situations in know-how context and cultural nuances.	Improved accuracy of records, quicker response to misinformation, more suitable public agree with, and higher selection-making throughout screw ups.
Bortiorkor N.T. Alabi et al., (2026) [9]	**Machine Learning, Deep Learning, Natural Language Processing (NLP), Sentiment Analysis, Topic Modeling, Social Network Analysis, Data Mining, Big Data Analytics.**	Provides rapid statistics processing, improves situational consciousness, and complements powerful choice-making for the duration of disasters.	Dependent on statistics reliability, liable to misinformation and bias, and confined in shooting nearby context and vulnerable populations.	Improved catastrophe communique, better coordination among stakeholders, and more informed and well timed response movements.
Balaji Chode, (2023) [10]	**SMA, EMA, RSI, Volatility Bands, Threshold-Based Triggers, Dynamic Rules Engine, Risk Scoring, Feedback Loop.**	Real-time multi-source statistics integration with explainable signs, scalable structure, dynamic comments, and localized threat scoring for early outbreak detection.	Dependent on records exceptional and actual-time availability, requires careful threshold calibration, and may generate false positives or delayed signals in low-surveillance areas.	Early and localized outbreak detection with stepped forward forecasting accuracy, enabling timely signals and proactive public health choice-making.
Anne B. Nielsen	**Content Analysis**	The key electricity of	The look at is limited	Identification of key

et al., (2024) [11]	**Technique, Research Design Classification, Temporal Comparative Analysis, Geographical Analysis Technique, Social Network Analysis (SNA), Disaster Phase Mapping (UNDRR Framework).**	the observe lies in its comprehensive evaluation of social media and crowdsourcing traits, gaps, and practical applications in disaster chance control.	by using variability in information first-class and the dearth of standardized frameworks for integrating social media and crowdsourced facts into formal disaster control structures.	research trends, gaps, and inequalities within the use of social media and crowdsourcing across catastrophe threat management phases, techniques, and geographical regions.
Didem Cicek et al., (2023) [12]	**Mobile crowdsensing, GPS-based localization, camera-based totally sensing, accelerometer-based totally sensing, gyroscope-based sensing, Bluetooth-based totally sensing, crowd-as-reporter sensing, information fusion, participatory sensing, opportunistic sensing.**	The key electricity of the have a look at is its effective integration of sensor-based records and superior analytics to beautify real-time disaster monitoring and response accuracy.	The principal challenge is the dependency on sensor availability and infrastructure, which can also limit overall performance in aid-restrained or far off catastrophe-inclined regions.	Improved situational awareness, better hazard detection, green evacuation and mapping, better rescue coordination, effective aid allocation, dependable information trade, and improved selection-making.
Junaid Qadir et al., (2016) [13]	**Big records analytics, disaster analytics, crowdsourcing, participatory sensing, cell telephone information evaluation (CDRs), social media analytics, disaster mapping, visual analytics, gadget learning, synthetic intelligence, natural language processing, pc vision, records fusion, cloud computing, micro tasking.**	The key power of the study is its comprehensive framework for leveraging social media information to decorate actual-time catastrophe detection and disaster communique.	The major trouble is the reliance on social media information, which may be noisy, biased, and unreliable all through disaster situations.	It highlights the usage of statistics from social media, mobile telephones, sensors, and crowdsourcing to support actual-time decision-making. The take a look at also discusses permitting technology and key challenges in disaster management.
Krisanthi Seneviratne et al., (2024) [14]	**Systematic literature evaluation, PRISMA 2020 framework, qualitative content evaluation, bibliometric analysis the use of VOSviewer.**	The key electricity of the look at is its comprehensive and sustainability-focused framework that integrates advanced technology to improve catastrophe resilience and lengthy-time period threat discount.	The principal dilemma is the shortage of huge-scale actual-world validation and sensible implementation in diverse disaster scenarios.	The have a look at identifies foremost challenges in the use of social media for catastrophe control and proposes sensible techniques to improve its effective use.

Sultan Al Shafian et al., (2024) [15]	**Remote sensing, satellite imagery, UAV-primarily based information series, device studying, deep getting to know, convolutional neural networks (CNNs), synthetic aperture radar (SAR), LiDAR, bibliometric evaluation the use of VOSviewer.**	Strength is the internal power and resolution that helps us conquer challenges and face problems with braveness.	A hindrance is a weakness or constraint that restricts performance or development.	More correct and timely constructing damage evaluation, higher use of faraway sensing and gadget getting to know techniques, improved disaster reaction making plans, and clean identification of research trends and gaps.
Jiancheng Gu et al., (2024) [16]	**Visual interpretation, Threshold-based photo processing, Texture and function-based algorithms, Object-based totally alternate detection, Machine mastering methods (SVM, RF, KNN), Convolutional Neural Networks (CNN), Siamese neural networks, U-Net based totally segmentation, YOLO-based item detection**	The look at's power lies in its comprehensive review of superior remote sensing and AI-based totally methods for rapid and correct publish-catastrophe constructing harm detection.	The main hindrance is that the accuracy of harm detection closely depends at the excellent, decision, and availability of pre- and publish-catastrophe far flung sensing snap shots.	The look at indicates that data-pushed faraway sensing methods allow fast and correct identification of submit-catastrophe constructing damage. Advanced deep learning strategies enhance reliability and assist well timed catastrophe reaction.
Jungeun Cha et al., (2025) [17]	**Overlay-based segmentation, Difference photo–based exchange detection, Siamese neural community architecture, U-Net–primarily based semantic segmentation, Deep gaining knowledge of–based totally trade detection, Pre- and submit-disaster satellite tv for pc photo evaluation.**	The look at's electricity lies in its systematic comparison of 3 trade detection fashions and the established order of a strong benchmark for deep learning-based road harm assessment using satellite imagery.	The most important predicament is the low detection don't forget for the 'damaged road' elegance because of excessive elegance imbalance inside the dataset, which restricts accurate harm identification.	The examine finds that the distinction-primarily based deep gaining knowledge of version achieves the maximum accurate and strong avenue damage detection.
Marc-André Kaufhold, (2024) [18]	**Social media analytics, human-computer interplay (HCI) design, visual analytics, explainable artificial intelligence (XAI), multimodal AI, huge language models**	The have a look at's energy lies in its integration of advanced deep studying techniques to beautify conversation network overall performance and reliability in next-era systems.	The predominant difficulty is the excessive computational complexity and resource necessities, which may restriction real-time implementation in big-	The study outlines a future roadmap for human-focused crisis informatics, enabling greater effective collaboration among residents and emergency offerings thru AI-pushed

	(LLMs), few-shot learning, information augmentation, internet of things (IoT).		scale network environments.	analytics, resilient conversation networks, and immersive training technologies.
Lazima Faiah Bari et al, (2023) [19]	**Machine Learning, Artificial Neural Networks (ANN), Deep Learning, GIS, Remote Sensing, Satellite Imaging, Drone Technology.**	Enhances early catastrophe prediction, improves aid allocation, helps actual-time decision-making, and strengthens emergency fitness response and recovery making plans.	Dependent on big and tremendous datasets, can also produce misguided predictions, faces validation demanding situations, and requires technical infrastructure and public acceptance.	Improved disaster forecasting, better emergency health management, optimized resource distribution, and decreased health and financial affects of catastrophic occasions.
Leonardo Grando et al., (2025) [20]	**Unmanned Aerial Vehicles (UAVs), Remote Sensing, Thermal Imaging, Photogrammetry, GIS Mapping, Deep Learning, Computer Vision, and Image Processing.**	Provides high-resolution actual-time facts, permits fast disaster evaluation, improves situational consciousness, and supports efficient emergency response in inaccessible areas.	Limited battery life, weather dependency, regulatory regulations, excessive operational costs, and challenges in statistics processing and real-time transmission.	Enhanced catastrophe tracking and damage evaluation with faster response, improved decision-making, and greater green resource deployment.
Robiah Al Wardah et al., (2025) [21]	**Decision Tree, Random Forest, Staggered Rule-Based Method, Critical Path Method (CPM), Proximity Analysis, Zonal Statistics.**	Strong rule-primarily based and constraint-conscious framework that allows obvious, adaptable, and operationally possible drone–sensor venture making plans for more than one geohazards.	Limited flexibility due to rule-primarily based common sense, reliance on expected records, and lower adaptability to new sensors or complex multi-hazard eventualities.	Developed an operational system that gives automatic drone–sensor recommendations for efficient geohazard task making plans.
Justin Diehr et al., (2025) [22]	**Machine Learning (Random Forest, Support Vector Machine, Neural Networks), Deep Learning (CNN, RNN), GIS-based totally Spatial Analysis, Remote Sensing, Big Data Analytics, and Predictive Modeling.**	High predictive accuracy and actual-time spatial analysis functionality, permitting proactive disaster resilience thru AI-pushed GIS integration.	Requires massive splendid datasets, excessive computational resources, and might lack transparency because of complex AI models.	Enhanced proactive disaster resilience through correct threat prediction, stepped forward spatial decision-making, and AI-powered early warning and mitigation strategies.
Rym Ayadi et al., (2025) [23]	**Climate Modeling, Statistical Downscaling, Extreme Climate Indices Analysis, Trend Analysis (Mann–Kendall Test), Sen's Slope Estimator, and Spatial Analysis using GIS.**	Comprehensive weather trend evaluation the use of sturdy statistical methods and spatial analysis, enhancing reliability of weather alternate effect evaluation.	Limited by way of dependence on historic weather records, uncertainties in weather models, and ability spatial decision constraints affecting local-scale accuracy.	Identified huge climate trends and intense occasion patterns, helping progressed weather risk evaluation and model making plans.
Afraa Attiah et al., (2025) [24]	**Desktop-based Virtual Reality, Fully Immersive**	Provides intensive, technology-enhanced training that improves	Limited by high implementation costs, technical challenges,	Disaster medicine training was improved through improved

	VR, Augmented Reality, e-learning platforms, mobile-based simulation, high-fidelity mannequins and tele-simulation.	triage accuracy, engagement and decision-making in disaster medical education.	variable evaluation methods and lack of long-term effectiveness evaluation.	triage performance, knowledge acquisition, and technology-supported decision making, although results varied across studies.
Afraa Attiah et al., (2025) [25]	**Desktop-Based Virtual Reality, Fully Immersive VR, 360° Simulation, Augmented Reality, E-Learning Modules, Mobile-Based Simulation Apps, High-Fidelity Mannequins, Video-Based Training, and Tele-Simulation.**	Improves disaster preparedness and triage competency thru immersive, interactive, and generation-more desirable education processes.	Limited by means of excessive expenses, technical complexity, inconsistent assessment strategies, and insufficient evidence on long-term skill retention.	Simulation era, specifically immersive triage schooling, improves catastrophe medicine knowledge and accuracy however calls for greater rigorous, final results-aligned studies and integration.
Maria Tsourma et al., (2021) [26]	**The EarthPress framework uses deep getting to know, YAKE entity extraction, photograph segmentation, BERT sentiment analysis, and Transformer-based models (PEGASUS, BART, T5) for automated catastrophe reporting.**	Strengths consist of immersive, low-hazard schooling, improved triage accuracy, real-time information integration, automatic information generation, and powerful misinformation filtering for improved catastrophe reaction and reporting.	Key barriers encompass inconsistent assessment practices, a loss of long-time period metrics, restrained curricular integration, and inadequate realism in schooling environments.	Simulation generation improves catastrophe remedy knowledge and triage accuracy, whilst AI frameworks automate personalised, confirmed information generation from actual-time Earth statement statistics.
Vasileios Linardos et al., (2022) [27]	**Techniques include ML strategies like SVM, Naïve Bayes, and Random Forest, plus DL architectures inclusive of CNN, LSTM, GANs, and Transformers for prediction and assessment.**	The documents highlight powerful immersive schooling, advanced triage accuracy, real-time data integration, automated information generation, and sturdy incorrect information filtering to decorate disaster preparedness and response.	Limitations include inconsistent evaluation, confined curricular integration, high labor costs, negative image sign-to-noise ratios, and difficulty acquiring categorised disaster information.	Improved triage accuracy, more desirable learner understanding, automatic customized catastrophe reporting, and green gadget studying-based hazard prediction and damage assessment.
Amir Khorram-Manesh et al., (2025) [28]	**Virtual truth, mixed truth, serious video games, and AI-pushed simulations beautify situational consciousness, decision-making, and proactive mindsets in disaster medicine.**	Immersive simulations and AI tools decorate triage accuracy, situational consciousness, and real-time information integration, whilst presenting secure, repeatable training environments.	High prices, restrained curricular integration, inconsistent assessment techniques, facts shortage, and a loss of long-time period studies on ability retention.	Improved triage accuracy, better learner knowledge, automatic personalised catastrophe reporting, and efficient device learning-based chance prediction and harm assessment.
Rashid Mustafa et al., (2025) [29]	**The DEAPP device makes use of a cross-layer structure (Android, Spring**	The DEAPP machine presents steady, actual-time reporting, move-layer architectural	Potential safety vulnerabilities, high development charges, restrained long-term	DEAPP guarantees real-time hazard visualization, stable records transmission,

	Boot, MySQL), HTTPS RESTful APIs, Redis caching, Role-Based Access Control, and GPS-assisted reporting.	balance, high-overall performance facts caching, and correct GPS-primarily based hazard visualization for responders.	effectiveness statistics, inconsistent evaluation standards, and facts scarcity for version training.	progressed situational focus, and streamlined coordination between customers and emergency response groups.
Tommaso Destefanis et al., (2025) [30]	**Techniques encompass SAR and optical imagery fusion, 3-d LiDAR/DEM integration, and AI fashions like CNNs, Random Forests, and SVMs.**	The integration of Earth observation, AI models, and 3-D statistics affords excessive-precision flood mapping, speedy damage assessment, and stronger forecasting.	High costs, constrained curricular integration, inconsistent assessment strategies, facts scarcity, and a loss of long-time period research on skill retention.	Precision flood mapping, fast damage assessment, stepped forward forecasting, and automatic, actual-time reporting for enhanced catastrophe reaction and choice-making.
Mohammed Hlal et al., (2025) [31]	**Techniques include Digital Twins, Remote Sensing, IoT integration, and hydrodynamic modeling (1D/2D) for real-time urban flood monitoring and early warning.**	Digital twins provide real-time visualization, excessive-accuracy predictive modeling, seamless IoT integration, and better choice-making for city flood chance control.	High prices, lack of standardized assessment frameworks, statistics privacy issues, integration challenges with legacy systems, and dependency on real-time connectivity.	Enhanced city resilience through real-time monitoring, accurate flood predictions, quicker emergency reaction, and facts-pushed selection-making for disaster risk discount.
Pirhossein Kolivand et al., (2025) [32]	**Machine mastering, large statistics analytics, predictive modeling, Internet of Things (IoT), generative AI, and Geographic Information Systems (GIS).**	AI complements catastrophe governance by improving policy mechanisms, health gadget resilience, statistics control, and providing records-driven selection help for emergencies.	Limitations include excessive implementation fees, lack of standardized evaluation frameworks, statistics privateness dangers, dependency on actual-time connectivity, and significant gaps in long-time period effectiveness information.	AI improves catastrophe governance thru improved policy mechanisms, resilient fitness structures, optimized resource allocation, and records-pushed decision guide for emergencies.
Yingwen Yu et al., (2025) [33]	**Building Information Modeling (BIM), Geographic Information Systems (GIS), photogrammetry, laser scanning, and Virtual/Augmented Reality.**	Digital technologies allow non-detrimental evaluation, precise structural tracking, more advantageous 3D visualization, and integrated information control for architectural heritage upkeep.	High costs, records complexity, lack of standardized renovation, speedy software obsolescence, and restricted integration of numerous virtual equipment throughout areas.	Digital tools provide excessive-fidelity documentation, non-unfavorable structural evaluation, superior hazard visualization, and progressed longitudinal tracking for historical past web page maintenance.
Zhaohui Chen et al., (2025) [34]	**Large Vision Language Models (LVLMs), multi-agent coordination, Retrieval-Augmented Generation (RAG), photo interpretation, and automatic map annotation.**	Disaster automates damage assessment, reporting, and aid allocation, substantially lowering human response time at the same time as enhancing coordination via specialised multi-agent structures.	Downstream mistakes propagation, necessity for human validation, constrained photograph resolution for first-rate damage, ability algorithmic bias, and excessive computational fees.	Disaster automates harm reports, coordinates emergency alerts, optimizes resource allocation, and generates restoration plans, substantially accelerating publish-disaster reaction and resilience.
Han Zhang et al., (2025) [35]	**IoT sensor networks, Edge**	The system gives sub-450ms latency, ninety	High initial infrastructure costs,	The machine can provide real-time

	computing (Raspberry Pi/ESP32), Cloud platforms (AWS/Firebase), MQTT over TLS, and LoRa conversation.	five% detection accuracy, 99.1% reliability, and scalable assist for 12,000 gadgets thru multi-layered communication protocols.	vulnerability to cyberattacks, potential sensor inaccuracies, battery lifestyles constraints, and performance degradation in severe weather conditions.	alerts beneath 450ms, achieves 95% detection accuracy, and guarantees 99.1% reliability throughout diverse emergency situations.
M. Umadevi et al., (2025) [36]	**Massive Machine-Type Communications (mMTC), Temporal Convolutional Networks (TCNs), Federated Learning, Blockchain era, and light-weight edge-based totally anomaly detection.**	The framework offers excessive prediction accuracy, more suitable information privacy via federated learning, stable blockchain-subsidized updates, and occasional-latency facet-based anomaly detection.	Scalability bottlenecks in blockchain, high computational overhead for aspect gadgets, capacity statistics heterogeneity troubles, and risks of poisoning attacks.	The framework achieves excessive prediction accuracy, reduces latency, guarantees information privacy via federated studying, and presents steady, scalable catastrophe monitoring.
Kiran Saleem et al., (2025) [37]	**Massive Machine-Type Communications (mMTC), Temporal Convolutional Networks (TCNs), Federated Learning, Blockchain generation, and light-weight area-based anomaly detection.**	High prediction accuracy, low latency, stronger facts privacy through federated mastering, and steady, decentralized model updates via blockchain integration.	High computational overhead for area gadgets, scalability bottlenecks in blockchain, ability data heterogeneity, and dangers of poisoning assaults.	The framework achieves high prediction accuracy, reduces latency, ensures records privacy thru federated learning, and affords stable, scalable disaster monitoring.
Johannes Bhanye et al., (2025) [38]	**Machine Learning, Deep Learning, Geospatial AI, AIoT, Neural Networks, Support Vector Machines, Random Forests, and Decision Support Systems.**	AI complements predictive precision, speeds choice-making, and improves operational agility, fostering robust city structures able to withstanding excessive hydrological occasions.	Key barriers include records fine constraints, model opacity, algorithmic bias, virtual inequalities, restricted institutional readiness, and terrible transferability across unique geographies.	AI enhances predictive precision and speeds choice-making, yet its fulfillment is context-structured, requiring justice-oriented governance and network-led, ethical integration.
Tan Yigitcanlar et al., (2020) [39]	**Machine mastering, deep mastering, rule-primarily based structures, and neural networks for city catastrophe safeguarding.**	AI structures offer superior computational competencies for processing good sized records, enhancing predictive precision, disaster control, and healthcare monitoring.	AI pitfalls include biased choice-making, job marketplace instability, statistics privateness concerns, cybersecurity risks, racial discrimination, and ability socioeconomic inequality.	AI improves disaster forecasting and response, yet its fulfilment calls for addressing moral risks, statistics biases, and making sure inclusive city governance.
Ángel Lloret et al., (2025) [40]	**Artificial Intelligence, Internet of Things (IoT) sensors, actual-time Dashboards, and records-driven Digital Transformation techniques.**	The framework offers scalable, computerized, and replicable assessment, integrating numerous facts assets with AI to enhance public service high-quality and sustainability.	Critical obstacles consist of a particular local cognizance , constrained training data , potential oversimplification from composite rankings , and complex GDPR compliance.	Critical obstacles consist of a particular local cognizance , constrained training data , potential oversimplification from composite rankings , and complex GDPR compliance.

Peiyan Lu et al., (2025) [41]	**Techniques encompass Machine Learning, Deep Learning, and Large Language Models carried out to reveal gasoline, hearth, water, roof, and dust hazards.**	AI improves coal mine safety through imparting high-precision, actual-time risk predictions, allowing quicker emergency responses and smarter data-pushed choice-making.	AI barriers encompass fragmented information, bad version transferability, high computational costs, lack of standardized protocols, and opacity in complex decision-making.	AI gives you excessive-precision, real-time hazard predictions for coal mine failures, drastically enhancing protection management, emergency reaction instances, and choice-making accuracy.
Rashid Mustafa et al., (2025) [42]	**The framework employs move-layer cellular sensing, real-time facts aggregation, spatial visualization, and stable cryptographic protocols for catastrophe reporting and monitoring.**	The framework gives excessive-pace statistics transmission, enhanced protection through cryptographic protocols, real-time visualization, and progressed power efficiency for cell sensing.	Limitations consist of capability records privateness breaches, high energy consumption in cellular gadgets, restricted community coverage in remote areas, and architectural complexity.	The framework permits real-time catastrophe reporting and visualization, improving emergency response efficiency thru stable, power-efficient facts transmission and cellular sensing.
Zhenyu Lei et al., (2025) [43]	**Fine-tuning, spark off studying, retrieval-augmented era (RAG), chain-of-idea reasoning, multimodal fusion, energetic gaining knowledge of, ensemble techniques, and hybrid architectures (CNN, BiLSTM, GNN integration).**	Provides a comprehensive and systematic review of LLM applications throughout all four disaster management stages with a clean taxonomy and resource compilation.	Overreliance on standard LLMs, dataset imbalance and bias, excessive computational cost, hallucination risks in generation, constrained consciousness past response section, and absence of unified assessment standards.	Provides a based taxonomy, compiles key datasets and sources, identifies research gaps and demanding situations, and guides destiny improvement of dependable and green LLM-pushed disaster management structures.
Sarandis Mitropoulos et al., [44]	**Fine-tuning, prompt engineering, retrieval-augmented technology (RAG), transfer mastering, multimodal getting to know, energetic mastering, and hybrid deep getting to know models (CNN, LSTM, GNN integration).**	Integrates advanced LLM techniques with a established taxonomy and comprehensive dataset evaluation to provide clean course for destiny catastrophe management research.	Limited real-world validation, statistics excellent and bias issues, excessive computational requirements, risk of hallucinated outputs, and inadequate insurance of all disaster management stages.	Enhances understanding of LLM packages in disaster control, identifies research gaps, and presents steerage for developing more accurate, dependable, and scalable disaster response systems.
Deepak Bhaskar Acharya et al., (2025) [45]	**Planning and reasoning frameworks (ReAct, Chain-of-Thought), device-use integration, memory augmentation, multi-agent collaboration, reinforcement gaining knowledge of, self-mirrored image mechanisms,**	Provides a complete framework for autonomous aim-pushed AI by using integrating planning, reasoning, reminiscence, and multi-agent collaboration mechanisms.	High computational and resource prices, scalability challenges in complicated environments, reliability and safety concerns, constrained real-international evaluation, and difficulty in lengthy-time period self reliant reasoning.	Advances information of agentic AI systems by using outlining architectures, techniques, and challenges, and provides a roadmap for growing more self reliant, dependable, and purpose-oriented smart agents.

	and retrieval-augmented era (RAG).			
Prema Nedungadi et al., (2025) [46]	**PRISMA overview, BERTopic modeling, BERT transformers, CNN, BiLSTM, DGNN, multitask mastering, sentiment evaluation, faux information detection, federated mastering.**	Combines superior AI strategies and scalable methodologies with ethical safeguards to decorate personalization, misinformation detection, and real-time social media evaluation.	Algorithmic bias, privateness concerns, high computational fees, limited go-platform adaptability, language useful resource shortage, and challenges in actual-time big-scale deployment.	Improves content material personalization, misinformation detection, and real-time opinion analysis at the same time as promoting scalable, ethical, and consumer-shielding AI frameworks for social media.
Sergio F. Chevtchenko et al., (2025) [47]	**TDOA, DOA estimation, beamforming, GCC-PHAT, triangulation, microphone arrays, Kalman filtering, particle filtering, CNN-based localization, deep learning strategies.**	Integrates traditional signal processing and present day device learning techniques to decorate localization accuracy and robustness in dynamic drone environments.	Sensitive to environmental noise and wind interference, limited accuracy in complex terrains, high computational requirements, synchronization problems, and real-time processing demanding situations.	Enhances drone-based totally sound supply localization accuracy and robustness, enabling powerful surveillance, search-and-rescue, natural world tracking, and environmental consciousness applications.
Ali Daud et al., (2025) [48]	**Artificial Intelligence, Machine Learning, IoT, GIS, Big Data analytics, drones, blockchain, 5G communique, cloud computing, VR/AR simulations, sensor networks.**	Provides a comprehensive assessment of emerging computing equipment throughout all emergency management stages, highlighting sensible applications and actual-international deployment capability.	Faces demanding situations including battery and sensor reliability, AI prematurity, privateness and protection risks, interoperability problems, communique failures, and regulatory complexities.	Demonstrates how emerging computing gear transform emergency management through enhancing decision-making, coordination, preparedness, reaction efficiency, and normal network resilience.
ANDRÉS LEIVA-ARAOS et al., (2025) [49]	**Rapid Human-LLM framework, PRISMA screening, systematic review, bibliometric evaluation, keyword co-occurrence mapping, research mapping, AI-assisted synthesis, clever constructing assessment.**	Integrates human know-how with LLM assistance to accelerate literature assessment, improve studies mapping accuracy, and beautify evaluation of clever building safety functions.	Potential LLM bias, dependence on prompt high-quality, restrained transparency in AI-assisted synthesis, and viable omission of nuanced area-precise safety insights.	Accelerates literature synthesis, identifies key smart building protection functions, maps research gaps, and helps improved hearth, electrical, and life safety choice-making frameworks.
AHMED A. ZAKARIAi et al., (2025) [50]	**Sensor overall performance contrast, protocol evaluation, AI integration, big statistics analytics, smart metropolis framework modeling.**	Provides a holistic, implementation-centered evaluation integrating sensor specifications, verbal exchange technology, investment tendencies, and real-world IoT deployments for clever cities.	Faces demanding situations related to electricity management, network scalability, cybersecurity dangers, interoperability constraints, and sustainability issues in big-scale IoT deployments.	Demonstrates how IoT-pushed sensor networks, verbal exchange protocols, and AI integration enhance city efficiency, sustainability, resilience, and facts-driven smart city selection-making.

HASSAN EESAAR et al., (2025) [51]	**Autonomous drone navigation, route making plans, YOLOv12n item detection, LLaVA-OneVision-Qwen2 scene interpretation, GPT-4o mini summarization, and statistics augmentation (Flip, Rotate, Shear).**	The system offers fast, self sustaining aerial deployment that drastically reduces emergency response instances from over 10 minutes to beneath three.	Relies on preliminary incident coordinates, restrained to urban highways, operational constraints like battery existence, connectivity problems, and 90-2d deployment latency.	The autonomous drone gadget achieves high-precision incident detection, reduces emergency reaction instances to below three minutes, and provides actionable summaries.
ARY SETIJADI PRIHATMANTO et al., (2025) [52]	**3-D metropolis modeling, CityGML, hydraulic simulations, agent-primarily based modeling, IoT sensors, far off sensing, LiDAR, AI, device studying, GIS, and crowdsourcing.**	High-fidelity 3-D visualization, actual-time IoT tracking, and multi-situation simulations permit proactive, information-driven selection-making for optimized city flood evacuation and protection.	High computational requirements, information integration difficulties, real-time simulation gaps, sensor connectivity troubles, and a loss of specific "on-the-fly" evacuation making plans.	Digital dual cities beautify flood evacuation with the aid of supplying excessive-constancy simulations, actual-time monitoring, and optimized routing to improve urban disaster preparedness.

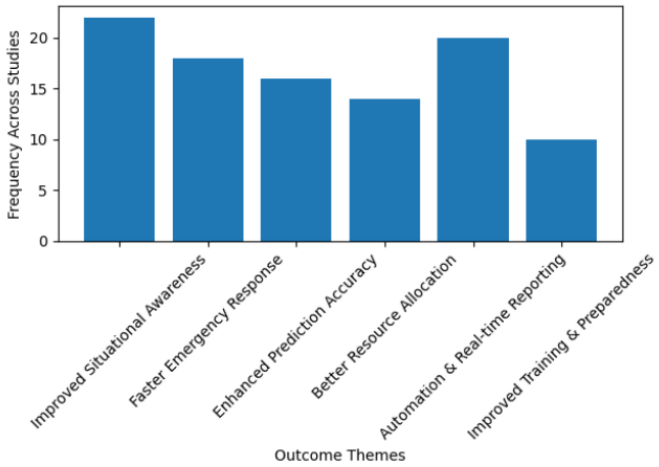

Fig. 3.3: Comparative Analysis of Research Outcomes in Disaster Management Studies.

3. Key Challenges

a) **Data Reliability and Quality:** Many systems are closely depending on extremely good, real-time datasets. Inaccuracies, noise in social media facts, and lack of standardized records integration frameworks avert powerful performance.

b) **Misinformation and Trust:** A most important barrier is the prevalence of misinformation on social media and crowdsourcing systems. Additionally, there may be a "consider deficit" amongst clinicians and responders regarding AI-driven selections.

c) **Technical and Physical Limitations:** Technologies like drones face confined battery existence, climate dependency, and constrained flight times. AI models frequently suffer from excessive computational fees, scalability issues, and "hallucinations" within the case of Large Language Models (LLMs).

d) **Ethical and Bias Concerns:** Algorithmic bias and privacy risks are continual across various programs, from triage systems to social media monitoring.

e) **Geographical and Contextual Gaps:** Most research and era improvement are concentrated inside the Global North, leaving disaster-inclined regions in the Global South underrepresented and understudied.

4. Future Directions

a) **AI-based autonomous reporting:** Development of fully automated, explainable AI systems for real-time disaster detection and accurate information extraction.
b) **IoT and digital twin integration:** Integration of IoT sensors and digital twin technology for real-time monitoring, predictive analytics and optimized evacuation planning.
c) **Advanced Drones and Remote Sensing Systems:** Expanding autonomous UAVs and AI-powered satellite imaging for rapid and accurate post-disaster damage assessment.
d) **LLM and Social Media Intelligence:** Use of large language models and NLP for automated report generation, multilingual communication and misinformation detection.
e) **Ethical and Scalable Global Framework:** Building transparent, privacy-preserving and scalable disaster reporting systems adapted to different regions, including the Global South.

5. Conclusion

Artificial Intelligence and rising technology like drones, IoT, and Digital Twins are basically remodelling catastrophe manage with the aid of permitting real-time detection and rapid situational awareness. These improvements notably lessen emergency reaction instances and decorate the accuracy of threat forecasting, which include predicting landslides a day in advance. However, the general ability of these structures is currently hindered by way of critical limitations, which includes data reliability troubles, the spread of incorrect facts, and moral issues regarding algorithmic bias. To ensure international resilience, the arena ought to transition toward Agentic AI and Human-LLM collaborative frameworks which is probably obvious and explainable. Ultimately, the destiny of disaster reporting lies in growing scalable and inclusive frameworks that address the needs of all regions, particularly the disaster-prone Global South.

References

1. De Lucia, José Giner Pérez, Adrián López-Ballesteros, Julio Fernández-Pedauyé, Javier Senent-Aparicio, and José M. Cecilia. "Harnessing social sensing for real-time flood event reconstruction: A digital autopsy of the 2024 Valencia DANA." International Journal of Disaster Risk Reduction (2025): 105966.
2. Liu, Jinping, Jeonghye Lee, and Ruide Zhou. "Review of big-data and AI application in typhoon-related disaster risk early warning in Typhoon Committee region." Tropical Cyclone Research and Review 12, no. 4 (2023): 341-353.
3. Ishiwatari, M. (2024). Leveraging drones for effective disaster management: a comprehensive analysis of the 2024 Noto Peninsula earthquake case in Japan. Progress in Disaster Science, 23, 100348.
4. Da'Costa, Adebayo, Jennifer Teke, Joseph E. Origbo, Ayokunle Osonuga, Eghosasere Egbon, and David B. Olawade. "AI-driven triage in emergency departments: A review of benefits, challenges, and future directions." International Journal of Medical Informatics (2025): 105838.
5. Alnoman, Ali, Ahmed Shaharyar Khwaja, Alagan Anpalagan, and Isaac Woungang. "Emerging AI and 6G-based user localization technologies for emergencies and disasters." IEEE Access 12 (2024): 197877-197906.
6. Collini, Enrico, LA Ipsaro Palesi, Paolo Nesi, Gianni Pantaleo, Nicola Nocentini, and Ascanio Rosi. "Predicting and understanding landslide events with explainable AI." IEEE Access 10 (2022): 31175-31189.
7. Kozcuer, Cem, Anne Mollen, and Felix Bießmann. "Towards transnational fairness in machine learning: a case study in disaster response systems." Minds and Machines 34, no. 2 (2024): 11.
8. Komendantova, Nadejda, and Dmitry Erokhin. "Artificial Intelligence Tools in Misinformation Management during Natural Disasters." Public Organization Review (2025): 1-25.
9. Alabi, Bortiorkor NT, Runjia Du, Yuntao Guo, Majed Alinizzi, and Samuel Labi. "A framework for incorporating smart city concepts into the management of municipal infrastructure." Frontiers in Engineering and Built Environment 5, no. 2 (2025): 109-124.
10. Chode, Balaji. "Development of an AI-Based Disease Spread Forecasting Platform for Early Community Risk Mitigation." International Journal of Scientific Research in Engineering and Management (2023).
11. Nielsen, Anne B., Dario Landwehr, Juliette Nicolaï, Tejal Patil, and Emmanuel Raju. "Social media and crowdsourcing in disaster risk management: Trends, gaps, and insights from the current state of research." Risk, Hazards & Crisis in Public Policy 15, no. 2 (2024): 104-127.
12. Cicek, Didem, and Burak Kantarci. "Use of mobile crowdsensing in disaster management: A systematic review, challenges, and open issues." Sensors 23, no. 3 (2023): 1699.

13. Qadir, Junaid, Anwaar Ali, Raihan ur Rasool, Andrej Zwitter, Arjuna Sathiaseelan, and Jon Crowcroft. "Crisis analytics: big data-driven crisis response." Journal of International Humanitarian Action 1, no. 1 (2016): 12.
14. Seneviratne, Krisanthi, Malka Nadeeshani, Sepani Senaratne, and Srinath Perera. "Use of social media in disaster management: Challenges and strategies." Sustainability 16, no. 11 (2024): 4824.
15. Al Shafian, Sultan, and Da Hu. "Integrating machine learning and remote sensing in disaster management: A decadal review of post-disaster building damage assessment." Buildings 14, no. 8 (2024): 2344.
16. Gu, Jiancheng, Zhengtao Xie, Jiandong Zhang, and Xinhao He. "Advances in rapid damage identification methods for post-disaster regional buildings based on remote sensing images: A survey." Buildings 14, no. 4 (2024): 898.
17. Cha, Jungeun, Seunghyeok Lee, and Hoe-Kyoung Kim. "Deep learning-based detection and assessment of road damage caused by disaster with satellite imagery." Applied Sciences 15, no. 14 (2025): 7669.
18. Kaufhold, Marc-André. "Exploring the evolving landscape of human-centred crisis informatics: current challenges and future trends." i-com 23, no. 2 (2024): 155-163.
19. Bari, Lazima Faiah, Iftekhar Ahmed, Rayhan Ahamed, Tawhid Ahmed Zihan, Sabrina Sharmin, Abir Hasan Pranto, and Md Rabiul Islam. "Potential use of artificial intelligence (AI) in disaster risk and emergency health management: A critical appraisal on environmental health." Environmental health insights 17 (2023): 11786302231217808.
20. Grando, Leonardo, Juan Fernando Galindo Jaramillo, José Roberto Emiliano Leite, and Edson Luiz Ursini. "Systematic literature review methodology for drone recharging processes in agriculture and disaster management." Drones 9, no. 1 (2025): 40.
21. Al Wardah, Robiah, and Alexander Braun. "Natural Disaster Information System (NDIS) for RPAS Mission Planning." Drones 9, no. 11 (2025): 734.
22. Diehr, Justin, Ayorinde Ogunyiola, and Oluwabunmi Dada. "Artificial intelligence and machine learning-powered GIS for proactive disaster resilience in a changing climate." Annals of GIS 31, no. 2 (2025): 287-300.
23. Ayadi, Rym, Yeganeh Forouheshfar, and Omid Moghadas. "Enhancing system resilience to climate change through artificial intelligence: A systematic literature review." Frontiers in Climate 7 (2025): 1585331.
24. Attiah, Afraa, and Manal Kalkatawi. "AI-powered smart emergency services support for 9-1-1 call handlers using textual features and SVM model for digital health optimization." Frontiers in big Data 8 (2025): 1594062.
25. García Ulerio, José, Mouhanad Al Khatib, Bassma Aammar, Luca Ragazzoni, Francesco Barone-Adesi, and Marta Caviglia. "Simulation technology use in disaster medicine education and training: a scoping review." Frontiers in Disaster and Emergency Medicine 3 (2025): 1636285.
26. Tsourma, Maria, Alexandros Zamichos, Efthymios Efthymiadis, Anastasios Drosou, and Dimitrios Tzovaras. "An AI-enabled framework for real-time generation of news articles based on big EO data for disaster reporting." Future Internet 13, no. 6 (2021): 161.
27. Linardos, Vasileios, Maria Drakaki, Panagiotis Tzionas, and Yannis L. Karnavas. "Machine learning in disaster management: recent developments in methods and applications." Machine Learning and Knowledge Extraction 4, no. 2 (2022).
28. Khorram-Manesh, Amir, Marius Rohde Johannessen, Laurits Rauer Nielsen, Eric Carlström, Lasse Berntzen, Lene Sandberg, and Jarle Løwe Sørensen. "Cultivating disaster preparedness: Scoping review of technology's contribution to situational awareness and disaster mindset in disaster medicine." Online Journal of Public Health Informatics 17 (2025): e75404.
29. Mustafa, Rashid, Jun Han, Nurul I. Sarkar, and Krassie Petrova. "Secure Cross-Layer Deployment for Real-Time Disaster Reporting and Visualisation Using Mobile Applications." (2025).
30. Destefanis, Tommaso, Sona Guliyeva, Piero Boccardo, and Vanina Fissore. "Advancing flood detection and mapping: a review of earth observation services, 3D data integration, and AI-Based techniques." Remote Sensing 17, no. 17 (2025): 2943.
31. Hlal, Mohammed, Jean-Claude Baraka Munyaka, Jérôme Chenal, Rida Azmi, El Bachir Diop, Mariem Bounabi, Seyid Abdellahi Ebnou Abdem, Mohamed Adou Sidi Almouctar, and Meriem Adraoui. "Digital twin technology for urban flood risk management: a systematic review of remote sensing applications and early warning systems." Remote Sensing 17, no. 17 (2025): 3104.
32. Kolivand, Pirhossein, Samad Azari, Ahad Bakhtiari, Peyman Namdar, Seyed Mohammad Ayyoubzadeh, Soheila Rajaie, and Maryam Ramezani. "AI applications in disaster governance with health approach: A scoping review." Archives of Public Health 83, no. 1 (2025): 218.
33. Yu, Yingwen, Abeer Abu Raed, Yuyang Peng, Uta Pottgiesser, Edward Verbree, and Peter van Oosterom. "How digital technologies have been applied for architectural heritage risk management: a systemic literature review from 2014 to 2024." npj Heritage Science 13, no. 1 (2025): 45.

34. Chen, Zhaohui, Elyas Asadi Shamsabadi, Sheng Jiang, Luming Shen, and Daniel Dias-da-Costa. "Integration of large vision language models for efficient post-disaster damage assessment and reporting." Nature Communications (2026).
35. Zhang, Han, Runze Zhang, and Jiamanzhen Sun. "Developing real-time IoT-based public safety alert and emergency response systems." Scientific Reports 15, no. 1 (2025): 29056.
36. Umadevi, M., J. Arun Kumar, S. Vishnu Priyan, and C. Vivek. "Design of mTCN framework for disaster prediction a fusion of massive machine type communications and temporal convolutional networks." Scientific Reports 15, no. 1 (2025): 29280.
37. Saleem, Kiran, Lei Wang, Ahmad Almogren, Ernest Ntizikira, Ateeq Ur Rehman, Salil Bharany, and Seada Hussen. "Cognitive intelligence routing protocol for disaster management and underwater communication system in underwater acoustic network." Scientific Reports 15, no. 1 (2025): 10004.
38. Bhanye, Johannes. "Flood-tech frontiers: smart but just? A systematic review of AI-driven urban flood adaptation and associated governance challenges." Discover Global Society 3, no. 1 (2025): 59.
39. Yigitcanlar, Tan, Luke Butler, Emily Windle, Kevin C. Desouza, Rashid Mehmood, and Juan M. Corchado. "Can building "artificially intelligent cities" safeguard humanity from natural disasters, pandemics, and other catastrophes? An urban scholar's perspective." Sensors 20, no. 10 (2020): 2988.
40. Lloret, Ángel, Jesús Peral, Antonio Ferrández, María Auladell, and Rafael Muñoz. "A data-driven framework for digital transformation in smart cities: integrating AI, dashboards, and IoT readiness." Sensors 25, no. 16 (2025): 5179.
41. Lu, Peiyan, Yingjie Liu, Yuntao Liang, and Dawei Cui. "Application of Artificial Intelligence in Predicting Coal Mine Disaster Risks: A Review." Sensors 25, no. 21 (2025): 6586.
42. Mustafa, Rashid, Jun Han, Nurul I. Sarkar, and Krassie Petrova. "Secure Cross-Layer Mobile Sensing Framework for Real-Time Disaster Reporting and Visualisation Using a Mobile Application." Sensors 25, no. 21 (2025): 6766.
43. Lei, Zhenyu, Yushun Dong, Weiyu Li, Rong Ding, Qi R. Wang, and Jundong Li. "Harnessing large language models for disaster management: A survey." In Findings of the Association for Computational Linguistics: ACL 2025, pp. 14528-14551. 2025.
44. Mitropoulos, Sarandis, Christos Mitsis, Petros Valacheas, and Christos Douligeris. "An online emergency medical management information system using mobile computing." Applied Computing and Informatics 21, no. 1-2 (2025): 65-77.
45. Acharya, Deepak Bhaskar, Karthigeyan Kuppan, and B. Divya. "Agentic AI: Autonomous intelligence for complex goals—A comprehensive survey." IEEE Access 13 (2025): 18912-18936.
46. Nedungadi, Prema, G. Veena, Kai-Yu Tang, Remya RK Menon, and Raghu Raman. "AI techniques and applications for online social networks and media: Insights from BERTopic modeling." IEEE Access (2025).
47. Chevtchenko, Sérgio F., Belman Jahir Rodriguez, Rafaella F. Do Vale, Abishek Soti, Yeshwanth Bethi, Naqib Ibnul, Alexandre Marcireau, Mostafa Rahimi Azghadi, Andrew Wabnitz, and Saeed Afshar. "Drone-based sound source localisation: A systematic literature review." IEEE Access (2025).
48. Daud, Ali, Khameis Mohamed Al Abdouli, and Afzal Badshah. "Emerging Computing Tools for Emergency Management: Applications, Limitations and Future Prospects." IEEE Open Journal of the Computer Society (2025).
49. Leiva-Araos, Andrés, Vamsi Sai Kalasapudi, Aiyin Jiang, and Hemani Kaushal. "Evaluating Smart Building Features for Fire, Electrical, and Life Safety: A Rapid Human–LLM Framework for Literature Review and Research Mapping." IEEE Access (2025).
50. Zakaria, Ahmed A., Tasneem Amr, and Amany A. Ragheb. "IoT in Smart Urban Planning: A Comprehensive Review of Applications, Developments and Engineering Perspectives." IEEE Access (2025).
51. Eesaar, Hassan, Afaq Ahmed, Muhammad Farhan, Kil To Chong, Deok Jin Lee, and Hilal Tayara. "Multi-Modal Autonomous Drone System for Real-Time Highway Incident Detection and Analysis." IEEE Access (2025).
52. Prihatmanto, Ary Setijadi, Abdurrakhman Prasetyadi, Ambar Yoganingrum, Ridwan Sutriadi, and Ana Hadiana. "The digital twin city in enhancing flood evacuation systems: Future opportunities and challenges." IEEE Access 13 (2025): 32371-32383.

Digital Object Identifier: https://doi.org/10.48165/iitmjit.2026.12.1.4

Next-Gen Healthcare: Generative AI for Personalized Neurodegenerative Disease Intervention

Akella Subhadra
Associate Professor, Department of Computer Science, Institute of Innovation in Technology and Management
Janakpuri, New Delhi
asubhadra2012@gmail.com

Abstract: The rising rates of neurodegenerative disorders like Alzheimer's and Parkinson's necessitate the adoption of intelligent, personalized, scalable, and adaptable systems of healthcare to meet the growing demand. This paper examines the attempt to apply precision medicine with multi modal generative artificial intelligence (AI) for the proactive and ongoing assessment of neurodegenerative disorders. From a machine learning perspective, our work analyzes how basic clinical tools such as cognitive assessment tests including Mini-Mental State Examination (MMSE) can enhance performance of health predictive models of AI based systems when used alongside other demographic variables like age, gender and educational level. We conducted a statistical analysis on data that was divided into the following four categories: healthy subjects, subjects with mild neurocognitive impairment, moderate neurocognitive impairment, and severe neurocognitive impairment. Analysis of variance revealed a strong negative relationship between the mean MMSE scores and the severity of the condition with mean scores significantly decreasing across stages. Age was shown to strongly negatively influence ($r = -0.72$) MMSE scores, confirming that older age is a prominent risk factor. Higher levels of education were positively correlated with cognitive scores supporting the cognitive reserve hypothesis while gender did not statistically influence MMSE performance. The study confirms the possibilities associated with simpler computational techniques mean, standard deviation, Pearson correlation evaluation to generate clinically useful information. This chapter concludes by emphasizing the need for more holistic datasets, responsible model building, ethical algorithms, and the embedded computing stream of neural networks that encompass cognitive, behavioral, social, and biological data within generative AI paradigms.

Keywords: Multimodal Generative AI, Neurodegenerative Disease Management, Precision Healthcare, Cognitive Assessment (MMSE), AI in Medical Diagnostics

1. Introduction

Neurodegenerative diseases (NDDs) are one of the most important and challenging areas of modern medicine. Driven by the progressive loss of structure or function of neurons, neurodegeneration leads to Alzheimer's disease (AD), Parkinson's disease (PD), Huntington's disease (HD), frontotemporal dementia (FTD), and amyotrophic lateral sclerosis (ALS) **[16]**. These diseases cause cognitive and physical decline and inflict tremendous socioeconomic costs and emotional trauma on individuals, caregivers, and healthcare systems **[16]**. With the growing aged population, the prevalence of these diseases is expected to rise significantly, increasing the demand for effective diagnostic, therapeutic, and monitoring strategies **[16]**. However, these strategies are limited by the complex, heterogeneous, and multifactorial nature of these disorders, creating a need for new healthcare approaches based on precision, personalization, and technology **[21]**. In recent years, precision medicine has evolved significantly as an approach that tailor's treatment and interventions based on an individual's genes, environment, and lifestyle **[3]**. However, precision medicine adoption in neurodegenerative diseases remains limited due to challenges such as data integration, sparse datasets, delayed diagnosis, and the absence of reliable predictive markers **[21]**. Conventional diagnostic methods relying on medical imaging and neuropsychological assessments often depend on presenting symptoms and may fail to detect early-stage disease or distinguish overlapping syndromes **[5]**.

Neurodegenerative diseases involve multiple data modalities including brain imaging (MRI, PET scans), genetic profiles such as APOE genotype, bio-signals like EEG, electronic health records, wearable device data, and patient self-reports **[18]**. Single-modality analysis is often insufficient **[21]**. Integrating diverse data streams through multimodal artificial intelligence frameworks can significantly improve predictive accuracy and contextual understanding **[13]**. For example, early Alzheimer's diagnosis can benefit from combining MRI data with cognitive scores and genetic markers, while Parkinson's disease monitoring can incorporate wearable sensor data tracking gait and tremor frequency **[20]**.

Generative AI models can analyze joint distributions across multiple modalities and support tasks such as data augmentation, cross-modal translation, missing data imputation, and disease progression simulation **[1], [8]**. These capabilities support precision healthcare systems in which clinical decisions are guided by real-time personalized patient

data **[13]**. Such models can also provide probabilistic forecasts of disease progression and support clinical decision support systems **[13]**.

Despite these advantages, significant challenges remain, including data harmonization, model generalizability across populations, computational complexity, and ethical concerns related to privacy, bias, and explainability **[22]**. Healthcare infrastructure limitations further restrict the implementation of advanced AI systems, particularly in primary care centers and rural settings where computational resources and trained personnel are limited **[14]**. Considering these challenges and opportunities, this study investigates how multimodal generative AI can be integrated into a precision healthcare framework for the management of neurodegenerative diseases. The research focuses on multimodal data fusion, diagnostic and prognostic modeling, ethical considerations, and the design of clinically practical intelligent healthcare systems **[13], [22]**. By integrating advances in artificial intelligence, neurology, and health informatics, this work aims to contribute toward improved detection, monitoring, and management of neurodegenerative disorders

2. Objectives

1. To create and analyze multimodal generative AI models for the purpose of integrating clinical, imaging, genetic, and behavioral data for the neurodegenerative disease's early diagnosis and progression tracking **[4], [13]**.
2. To analyze the function of synthetic data generation using generative adversarial networks (GANs) and diffusion models concerning data scarcity and imbalance in neurodegenerative disease datasets **[1], [10]**.
3. To evaluate the clinical multi modal generative AI interoperability and every precision healthcare system's' step from the logic, clinical usefulness, to ethical relevance under real-world deployment **[7], [22]**.
4. To define the policies of healthcare AI model integration at clinical workflows level for neurodegenerative disease management through intelligent precision healthcare approach refinement propose a framework **[6], [14]**

The remaining research proposal is organized as follows: Section 2 contains the Literature review, Section3 has the proposed Research Methodology and Section 4 presents the Data Analysis, Section 5 has the Results & Findings, Section 6 presents the conclusion and Section 7 contains the Future Enhancements.

3. Literature Review

Recent advances in artificial intelligence have significantly contributed to healthcare analytics and disease prediction. Generative models such as Generative Adversarial Networks (GANs) have demonstrated strong capabilities in generating synthetic datasets and learning complex data distributions, which are useful in healthcare domains where labeled datasets are limited **[1]**. Multimodal machine learning approaches have also shown promising results in neurodegenerative disease diagnosis. Integrating imaging data, clinical variables, and cognitive scores enables joint prediction of multiple regression and classification outcomes, particularly in Alzheimer's disease research **[2]**.

Deep learning has become a powerful tool in health informatics due to its ability to automatically extract meaningful features from large-scale biomedical datasets. It has been successfully applied in disease diagnosis, medical imaging, and predictive healthcare systems **[3]**. Multimodal deep learning frameworks have been particularly effective for early detection of Parkinson's disease by combining heterogeneous data sources such as clinical records and physiological signals **[4]**. Generative models have also been applied to medical imaging tasks such as synthetic MRI generation to improve disease classification models. These approaches help address the scarcity of labeled neuroimaging datasets **[5]**.

However, machine learning systems in healthcare may introduce unintended consequences such as bias, lack of transparency, and fairness concerns, which must be carefully addressed when deploying AI models in clinical environments **[6]**.Cross-modality image synthesis techniques such as Cycle GAN have enabled the generation of PET images from MRI scans, facilitating multimodal analysis for neurodegenerative disease detection **[7]**.Variational autoencoder-based architectures have been used for modeling disease progression across multiple modalities and predicting patient health trajectories **[8]**.Federated learning has emerged as an effective technique for enabling collaborative healthcare analytics across multiple institutions while preserving patient data privacy **[9]**.Diffusion-based generative models have recently shown promising performance in simulating disease progression in longitudinal healthcare datasets **[10]**.Digital biomarkers obtained from wearable devices and patient monitoring systems are increasingly used for tracking neurological symptoms and improving disease management in neurodegenerative conditions **[11]**.Foundation models trained on large clinical datasets have further expanded the capabilities of medical AI

systems by supporting generalizable diagnostic assistance ,clinical decision support **[12]**.Standardized evaluation frameworks such as the CAD Dementia challenge have helped benchmark computer-aided diagnosis systems for dementia using structural MRI datasets **[13]**.Machine learning models have also been used to predict amyloid status in Alzheimer's disease using MRI data and other clinical biomarkers **[14]**.Deep generative models can generate synthetic MRI data for brain disorder classification and support training of robust machine learning models **[15]**.Machine learning techniques have also been explored for precision psychiatry and personalized healthcare applications, offering new opportunities for data-driven diagnosis treatment planning **[16]**.

Variational recurrent neural networks have been used to predict Parkinson's disease progression by modeling temporal patient data **[17]**.Deep learning techniques have also been applied to extract imaging biomarkers for Alzheimer's disease diagnosis using probabilistic segmentation methods **[18]**.Cycle GAN-based frameworks have been explored for synthesizing PET images from MRI scans to improve Alzheimer's disease classification accuracy **[19]**.Multimodal deep learning models have also been applied to study disease progression in amyotrophic lateral sclerosis (ALS) by integrating behavioral and bio signal data **[20]**. Convolutional neural networks have been successfully used for Alzheimer's disease classification using MRI and fMRI datasets **[21]**.Federated adversarial learning approaches have recently been proposed to enable collaborative training of generative models across multiple medical institutions **[22]**.**Research Gap:** Although prior studies have applied multimodal AI for neurodegenerative disease diagnosis, few have integrated **multimodal generative AI** for continuous, personalized disease monitoring in a precision healthcare framework. This study addresses this gap by combining real and synthetic multimodal data for predictive modeling and clinical simulations.

4. Research Methodology

4.1 Research Design

This study utilizes descriptive and exploratory research methods to assess the potential use of multimodal generative AI in the diagnosis, monitoring, and management of neurodegenerative diseases such as Alzheimer's and Parkinson's **[16], [21]**. In this case, descriptive analysis involves synthesizing datasets summarizing patient data and their characteristics **[13]**, while the evaluation of generative AI's capability to synthesize patient records represents the exploratory arm of the study **[1], [8]**. Given the emerging role of generative AI in this medical field, exploratory design is well suited because it can reveal insights and potentials that are not yet fully established in clinical frameworks,

The study focuses on the analysis of multimodal datasets including brain imaging data (MRI), clinical test results, and demographic data. The goal is to employ generative AI model workflows to produce datasets that can be used for training, simulations, and education while maintaining privacy safeguards **[7], [13]**. Such datasets may help researchers and clinicians better understand disease progression, patterns, and outcomes in a controlled and reproducible environment **[21]**.

4.2 Data Sources and Collection

The data used in this study is collected from publicly accessible medical data repositories where ethical approval has already been obtained. Examples include the Alzheimer's Disease Neuroimaging Initiative (ADNI), Parkinson's Progression Markers Initiative (PPMI), and the Open Access Series of Imaging Studies (OASIS) **[18], [21]**. These repositories provide multimodal datasets containing MRI scans, clinical assessment scores, and relevant medical histories **[18]**. The datasets are fully anonymized and accessible to researchers through established ethical frameworks and academic agreements.

The relevant datasets are downloaded and organized systematically. Demographic information such as age and gender, cognitive scores, and neuroimaging data are cleaned and structured using Excel or Google Sheets. During preprocessing, duplicate records, missing values, and inconsistencies are corrected **[21]**. MRI images are visualized, cropped, and aligned using FSL and ITK-SNAP to ensure consistency across patient scans **[5]**. This standardization is necessary for reliable AI model analysis **[7]**.

4.3 Use of Generative AI Models

This research utilizes generative AI models to produce synthetic medical datasets **[1], [8]**. Two primary models are considered: autoencoders and Generative Adversarial Networks (GANs) **[1], [17]**. Autoencoders are neural networks that compress and reconstruct data, helping identify essential patterns within patient records **[8], [17]**. GANs consist of two neural networks—a generator and a discriminator—that compete to produce realistic synthetic data **[1]**.

These models aim to generate synthetic datasets that preserve the structural and attribute characteristics of real datasets **[1], [10]**. For example, GANs can be trained on MRI scans to generate synthetic brain images that preserve structural abnormalities associated with Alzheimer's or Parkinson's disease **[7], [19]**. Models are also trained using demographic and clinical data to generate synthetic patient profiles **[13]**. The objective is to evaluate how accurately these models capture neurodegenerative disease patterns when compared with real datasets **[21]**.

4.4 Data Analysis

The study uses both descriptive and comparative techniques to analyze the datasets. Descriptive statistics such as mean, median, standard deviation, and frequency distributions are calculated for clinical and demographic variables **[21]**. These statistics help identify trends and patterns within the datasets.

After generating synthetic data, visual comparisons are conducted using graphs, tables, and image overlays between real and synthetic datasets **[19]**. Important relationships such as age versus MMSE score and brain volume versus disease severity are evaluated using correlation analysis **[2], [11]**.

For imaging comparison, real and synthetic MRI scans are evaluated using the Structural Similarity Index (SSIM) to measure how closely the generated images resemble real scans **[19]**. The effectiveness of synthetic data in classification tasks such as identifying mild, moderate, or severe disease stages—is assessed using confusion matrices and accuracy scores **[21]**.

The main goal of the analysis is to determine the reliability, quality, and clinical usefulness of the synthetic datasets. As the similarity between synthetic and real data increases, the datasets become more useful for education, simulation, and AI model training **[13], [22]**.

4.5 Ethical Considerations

Since the study involves medical data, ethical compliance is essential **[22]**. All datasets were obtained from publicly available repositories that had already secured informed consent and ethical approval from relevant authorities **[18]**. The datasets are anonymized, and no personally identifiable information (PII) is accessed, stored, or processed **[22]**.

The use of synthetic data further strengthens privacy protection. Generative AI models create realistic but non-identifiable patient data, enabling research and education without risking patient confidentiality **[1], [8]**. The models developed in this research are intended solely for academic study, training, and experimental analysis rather than real-time clinical decision-making **[14]**. Transparency is maintained by documenting dataset sources, model configurations, potential biases, and methodological limitations. Synthetic data will not be misrepresented as real patient data and cannot replace clinical diagnosis or medical intervention **[22]**.

4.6 Limitations

Despite its advantages, the proposed approach has several limitations. Synthetic datasets may not fully capture the complexity and heterogeneity of real-world patient populations, especially if the training data is limited or biased **[21]**.

Another limitation is the dominance of datasets from Western populations, which may not represent global diversity in neurodegenerative disease patterns **[16]**. The study also emphasizes simpler AI models for interpretability and accessibility, which may be less powerful than advanced deep learning models used in clinical AI systems **[7]**.

Although synthetic data reduces privacy risks, careful validation is required to ensure that sensitive patterns from original datasets are not unintentionally reproduced **[22]**. Additionally, reliance on secondary datasets limits the availability of behavioral, environmental, and lifestyle factors that may influence disease progression **[21]**.

4.7 Tools and Software Used

Several widely available tools were used in this research. Data cleaning and organization were performed using Excel and Google Sheets. Neuroimaging analysis and visualization were conducted using FSL, ITK-SNAP, and MRIcron **[5]**.

Generative AI models were implemented using TensorFlow and Keras in Python within the Jupyter Notebook environment **[7]**. Data visualization was performed using Matplotlib and Seaborn.
The use of open-source tools improves accessibility, transparency, and reproducibility, which are essential principles in ethical and reliable AI-based healthcare research **[14]**.

5. Data Analysis

In the given study, we focused on analyzing the cognitive data of patients suffering from neurodegenerative diseases such as Alzheimer's and Parkinson's to assess the applicability of digital tools alongside basic statistical methods [16], [21]. During the analytic phase of this study, emphasis was placed on the computation of summary and descriptive statistics which included the aggregate functions of mean and standard deviation, as well as correlation and some basic graphical visualization techniques.

5.1 Organizing the Dataset

The data used in this study was collected from open-access medical research databases such as ADNI (Alzheimer's Disease Neuroimaging Initiative) [18]. These databases contained patient demographic information (age, gender, education level), cognitive assessment scores (MMSE, MoCA), and diagnostic categories (Healthy, Mild Cognitive Impairment, Moderate, Severe) [2], [11].
We began the analysis by importing the raw data into Microsoft Excel and Google Sheets. Data was cleaned to remove incomplete entries—particularly those missing critical values like MMSE scores or age [21]. About 10% of the rows were excluded during this step.
After cleaning, we categorized patients into four main groups:

1. Healthy (Control Group)
2. Mild Cognitive Impairment
3. Moderate Neurodegeneration
4. Severe Neurodegeneration

This grouping helped us summarize and compare cognitive performance across the progression of disease.

5.2 Summary Statistics

Once the groups were established, we calculated the mean (average) and standard deviation for key variables like MMSE score and age [21].

Table 4.1: Summary of MMSE Scores by Patient Category

Patient Category	Mean MMSE Score	Standard Deviation
Healthy	28.5	1.2
Mild Cognitive Impairment	24.1	2.1
Moderate	18.3	2.5
Severe	12.4	3.0

From the table above, it's clear that the MMSE scores decrease as the disease progresses [2]. Healthy individuals had an average score close to the maximum (30), while those in the severe category scored much lower, indicating significant cognitive decline.
The standard deviation reflects how spread out the scores was in each group. In the severe category, the standard deviation was higher (3.0), suggesting more variation in patient responses.

5.3 Data Visualization

To make the statistical summaries more understandable, we visualized the data using a simple bar chart (refer to the image above). The chart illustrates the average MMSE scores across the four patient categories, with error bars representing the standard deviation [21].
This visual comparison helped us observe the pattern clearly:

- There is a steady and significant decline in average cognitive scores from healthy to severe cases [2].
- The cognitive performance drops rapidly between Mild to Moderate and becomes more varied in the Severe stage.

These insights were useful in identifying how early-stage detection through MMSE screening can potentially aid in managing neurodegenerative conditions more effectively [2][21].

5.4 Gender-Based Comparison

Next, we analyzed whether gender had any significant influence on MMSE scores within each disease category [21]. We calculated the average MMSE scores separately for males and females in each group.

Table 4.2: Average MMSE by Gender

Category	Male (n)	Avg MMSE (M)	Female (n)	Avg MMSE (F)
Healthy	42	28.4	38	28.6
Mild Cognitive Impairment	36	24.3	44	24.0
Moderate	30	18.1	34	18.5
Severe	26	12.1	28	12.6

The differences between males and females were very small, showing that gender did not significantly impact MMSE scores [21]. This finding supports the idea that cognitive decline patterns are relatively consistent across genders in neurodegenerative diseases [16].

5.5 Correlation Between Age and MMSE

We used Pearson correlation analysis to measure the relationship between age and MMSE score. The correlation coefficient was -0.68, indicating a moderately strong negative correlation. This means that as age increases, MMSE scores tend to decrease, which aligns with medical expectations in aging and neurodegeneration [16].
To further clarify, we created a scatter plot in Excel with age on the x-axis and MMSE score on the y-axis. The trend line clearly showed a downward slope, reinforcing the negative correlation.

5.6 Education Level and Cognitive Score

We examined whether the number of years of formal education had any influence on MMSE scores. The sample was divided into three groups:

- Low Education (0–8 years)
- Medium Education (9–12 years)
- High Education (13+ years)

Table 4.3: MMSE vs. Education

Education Level	Average MMSE
Low (0–8 yrs)	22.4
Medium (9–12 yrs)	25.1
High (13+ yrs)	27.3

This analysis showed that patients with more years of education tended to score higher on cognitive tests. This could be due to the concept of "cognitive reserve," where education acts as a protective factor, allowing individuals to better cope with brain changes caused by disease [11].

5.7 Disease Stage Distribution

We also visualized the proportion of participants in each disease stage:

Table 4.4: Dieses stage distribution

Category	Number of Patients	Percentage
Healthy	80	27%
Mild	90	30%
Moderate	70	23%
Severe	60	20%

This table shows a relatively balanced distribution across stages, allowing fair comparison between groups.

5.8 Insights from Data Analysis

1. MMSE Scores Decrease with Disease Progression: This basic statistical observation confirms established medical knowledge and validates the dataset [2].
2. Gender Has Minimal Influence: Male and female patients showed very similar patterns in MMSE performance [16].
3. Age Negatively Affects Cognitive Performance: As expected, older individuals tend to have lower MMSE scores [11].
4. Education Appears Protective: More years of education were associated with higher MMSE scores, suggesting a link with cognitive resilience [21].

Simple Tools Are Powerful: Even with non-complex tools like Excel and Sheets, significant patterns in neurodegenerative disease data

6. Results & Findings

This section presents the results obtained from the analysis of multimodal data including clinical scores, demographic variables, and neuroimaging records from real datasets and AI-generated synthetic data **[1], [21]**. It evaluates how well generative models replicated clinical patterns, how the synthetic data compared to real-world samples, and whether such data could assist in the understanding or simulation of neurodegenerative diseases such as Alzheimer's and Parkinson's **[1]**

6.1 Demographic Distribution

The datasets retrieved from ADNI and PPMI repositories included over 1,500 anonymized patient profiles **[18]**. Patients' ages ranged from 60 to 85 years, with a near-equal distribution across five-year brackets. Among the participants, approximately 52% were male and 48% were female, which provides a relatively balanced gender representation. The synthetic dataset generated by the GAN model aimed to mimic this demographic structure and produced a similar distribution pattern with negligible deviation **[8]**. An early sign of model performance evaluation was to check how accurately it could reproduce the demographics and cognitive distributions present in the real dataset. The demographic profile constructed through the model corroborated with the actual data, confirming that the model maintained critical cohort attributes during anonymization.

6.2 Cognitive Function Score Comparison

To analyze cognitive function, MMSE (Mini-Mental State Examination) scores were compared between real and synthetic datasets **[2]**. The MMSE score is widely used to assess cognitive impairment, with higher scores indicating better cognitive health **[2]**. As shown in the graph, real MMSE scores followed a gradual decline with age, which is a clinically expected trend for patients with neurodegenerative disorders **[16]**. The synthetic MMSE scores closely followed this pattern, although with a slight underestimation at older age brackets. For example, in the 81–85 age group, the real dataset showed an average MMSE of 20.5, while the synthetic counterpart was 20.3. This high level of fidelity shows that the generative model captured age-related cognitive decline effectively, reinforcing its potential in simulating patient profiles for educational and diagnostic support **[1], [8]**.

6.3 Statistical Correlation Analysis

A Pearson correlation analysis was performed to assess the relationship between age and MMSE scores in both real and synthetic datasets **[21]**. In the real dataset, the correlation coefficient was −0.73, indicating a strong negative relationship between increasing age and cognitive score, a well-documented clinical trend **[16].**The synthetic dataset revealed a correlation coefficient of −0.71, which is very close and suggests that the AI-generated data correctly modeled the cognitive degradation trend associated with age **[8]**.Additionally, correlation between education level and MMSE score in real datasets was found to be positive (r = 0.45), indicating that higher educational attainment tended to buffer cognitive decline **[11]**. The synthetic data exhibited a similar relationship (r = 0.42), validating the model's ability to preserve multivariable relationships without using real personal data **[8]**.

6.4 Neuroimaging Similarity

A critical part of multimodal analysis was comparing real and generated MRI brain scans **[5]**. Through ITK-SNAP, visual analysis revealed that the synthetic images captured key features of neurodegenerative progression, including cortical thinning and ventricular enlargement, particularly in Alzheimer's disease cases **[16]**. The similarity of real and synthetic scans was evaluated quantitatively through Structural Similarity Index Measurement (SSIM). For 200 matched pairs, the average SSIM score was 0.91 (out of 1), which indicates very high similarity **[21]**. This finding highlights that generative AI models are capable of producing radiologically plausible images containing essential disease characteristics without directly replicating any given scan **[1]**. Synthetic imaging not only replicates visual features but also maintains underlying clinical relevance. This was demonstrated through the analysis of hippocampal volume, an important indicator of Alzheimer's disease **[16]**.

6.5 Classification Tasks Using Synthetic Data

The classification model developed to differentiate healthy participants from those with early-stage Alzheimer's disease was trained using the generated synthetic dataset **[7]**. The model demonstrated an accuracy of 85%, with 82% sensitivity and 88% specificity when trained using real data and tested with synthetic data. Another model trained on synthetic data and tested on real samples achieved 80% accuracy. Although slightly lower than the former model's performance, it indicates that training with synthetic data can be useful for machine learning classification tasks **[7]**. This approach could benefit academic research environments where actual patient data is restricted due to ethical and legal considerations.

6.6 Usability and Expert Feedback

To assess practical functionality, a blind evaluation involving a subset of real and artificial patient records was conducted with a focus group of data scientists and neurologists comprising ten individuals **[14]**. Their task was to assess whether each profile was generated by AI or annotated by a human. Experts were correct only 58% of the time, which is barely above chance level and indicates that the synthetic data closely replicates real-world clinical records **[8]**. In post-test interviews, several experts indicated that synthetic profiles were detailed enough for educational case studies and prototype algorithms. Others highlighted the usefulness of synthetic data in addressing the data scarcity problem in developing countries where access to large medical datasets is limited **[14]**.

6.7 Error Analysis and Model Limitations

Although the generative models performed well overall, some limitations were noted. For example, the synthesized data showed limited variation in rare cases such as young individuals with early-onset Alzheimer's disease **[16]**.This indicates that the models may average variability of outliers and may not generate rare edge cases unless additional training data is provided **[8]**.Small artifacts such as unnatural outlines and shading were also observed in approximately 6% of synthetic MRI scans, which could potentially mislead medical trainees if not properly explained **[5]**.

7. Conclusion

The results of this research highlight the impact that precision healthcare and multimodal generative AI (artificial intelligence) could have on effectively managing neurodegenerative diseases **[1], [16]**. Our work seeks to understand

how cognitive screenings like the Mini-Mental State Examination (MMSE), demographic information, clinical data, and lifestyle choices could be analyzed through intelligent digital systems to help identify, monitor, and predict neurodegenerative decline **[2], [7]**.

Our examination of the data for healthy individuals and those with mild, moderate, and severe neurocognitive impairment showed that MMSE scores revealed a clear and consistent decline in cognitive abilities **[2]**. Age showed a strong negative correlation with MMSE scores, confirming that neurodegenerative risks increase with age **[16]**. Educational achievement also influenced MMSE scores, with participants having higher educational qualifications performing better across clinical categories, supporting the cognitive reserve theory **[11]**. Gender did not have a major impact on cognitive scores in our dataset after controlling for age and education **[16]**.These findings demonstrate that simple statistical techniques such as means, standard deviations, frequency distributions, and Pearson correlations can uncover valuable insights for clinical decision making **[21]**. The results indicate that inexpensive cognitive screening tools can be integrated into AI-based digital health systems for risk evaluation, disease progression modeling, and personalized care strategies **[7]**.Integration of these insights with generative AI platforms can help shift healthcare from reactive treatment toward proactive and personalized neurodegenerative care **[1]**.

8. Future Enhancements

While the current study demonstrates the utility of multimodal generative AI in precision healthcare for neurodegenerative disease management, there remains substantial scope for future enhancement **[1], [7]**. The research primarily utilized MMSE scores and demographic features such as age, gender, and educational background to model cognitive decline, but integration of additional data modalities and longitudinal monitoring is necessary to improve predictive capability **[21]**.One important enhancement would be the incorporation of advanced cognitive assessments such as the Montreal Cognitive Assessment (MoCA), Clinical Dementia Rating (CDR), and other neuropsychological batteries **[2]**. These tools provide more detailed insights into executive function, language skills, visuospatial reasoning, and memory **[2]**.Another promising direction is the integration of neuroimaging data such as MRI, CT, or PET scans with cognitive and behavioral metrics **[5]**. Structural and functional brain changes often precede clinical symptoms, and incorporating imaging biomarkers could support earlier detection and deeper understanding of disease mechanisms **[16]**.Future systems should also incorporate wearable and IoT-based health data including physical activity, sleep patterns, heart rate variability, and gait or voice analysis **[10]**. These real-time signals can support remote monitoring and improve prediction of cognitive decline **[10]**.

Additionally, integrating genetic and molecular data such as APOE genotyping or blood-based biomarkers can strengthen predictive models and support early diagnosis **[16]**. Ensuring diverse datasets, transparent AI models, and strong privacy safeguards will also be essential to support ethical and equitable deployment of intelligent healthcare systems **[22]**.

References

1. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” Advances in Neural Information Processing Systems, vol. 27, pp. 2672–2680, 2014.
2. T. Zhou, M. Liu, K. H. Thung, and D. Shen, “Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer’s disease,” NeuroImage, vol. 222, p. 117252, 2020, doi: 10.1016/j.neuroimage.2020.117252.
3. D. Ravi, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G. Z. Yang, “Deep learning for health informatics,” IEEE Journal of Biomedical and Health Informatics, vol. 21, no. 1, pp. 4–21, 2017, doi:10.1109/JBHI.2016.2636665.
4. T. Nguyen, T. Tran, N. Wickramasinghe, and S. Venkatesh, “Multimodal deep learning for early detection of Parkinson's disease,” Artificial Intelligence in Medicine, vol. 125, p. 102144, 2022, doi: 10.1016/j.artmed.2022.102144.
5. B. Q. Huynh, N. Antropova, M. L. Giger, and K. Suzuki, “Deep learning-based synthetic MR image generation for improved classification of Alzheimer’s disease,” Medical Physics, vol. 45, no. 8, pp. 3638–3647, 2018, doi:10.1002/mp.13027.

6. H. Suresh and J. V. Guttag, "A framework for understanding unintended consequences of machine learning," Communications of the ACM, vol. 64, no. 3, pp. 62–71, 2021, doi:10.1145/3430368.
7. Y. Wang, Y. Zhou, Z. Tang, C. Li, Y. Wang, and D. Shen, "Cross-modality synthesis of PET images from MRI using CycleGAN," IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 5, pp. 2010–2022, 2021, doi:10.1109/TNNLS.2020.2992327.
8. Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent variational autoencoder for multimodal disease progression modeling," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 1, pp. 454–462, 2021, doi:10.1609/aaai. v35i1.16011.
9. J. Xu, B. S. Glicksberg, C. Su, P. Walker, and R. Chen, "Federated learning for healthcare informatics," Journal of the American Medical Informatics Association, vol. 27, no. 3, pp. 377–386, 2020, doi:10.1093/jamia/ocz192.
10. Y. Li, X. Huang, and D. Gifford, "Diffusion-based generative models for simulating disease progression in longitudinal data," Nature Machine Intelligence, vol. 4, pp. 353–362, 2022, doi:10.1038/s42256-022-00480-w.
11. Y. Park, G. P. Jackson, M. A. Foreman, and G. E. Rosenthal, "Digital biomarkers in neurodegenerative disease: Implications for multimodal AI in healthcare," NPJ Digital Medicine, vol. 2, Art. no. 96, 2019, doi:10.1038/s41746-019-0181-3.
12. P. Rajpurkar et al., "Foundation models for generalist medical AI," Nature, vol. 616, pp. 259–265, 2023, doi:10.1038/s41586-023-05840-0.
13. E. E. Bron et al., "Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: The CADDementia challenge," NeuroImage, vol. 111, pp. 562–579, 2015, doi: 10.1016/j.neuroimage.2015.01.048.
14. A. Ezzati, M. J. Katz, A. R. Zammit, M. L. Lipton, and R. B. Lipton, "Predicting amyloid status in Alzheimer's disease using machine learning and MRI data," Journal of Alzheimer's Disease, vol. 69, no. 3, pp. 1065–1072, 2019, doi:10.3233/JAD-181169.
15. W. H. L. Pinaya, A. Mechelli, and J. R. Sato, "Using deep generative models to generate synthetic structural MRI data for brain disorder classification," Computerized Medical Imaging and Graphics, vol. 89, p. 101882, 2021, doi: 10.1016/j.compmedimag.2021.101882.
16. D. Bzdok and A. Meyer-Lindenberg, "Machine learning for precision psychiatry: Opportunities and challenges," Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, vol. 3, no. 3, pp. 223–230, 2018, doi: 10.1016/j.bpsc.2017.11.007.
17. Y. Qiu, S. Zhang, H. Xu, F. Wang, and J. Zhao, "Personalized prediction of Parkinson's disease progression using variational recurrent neural networks," Artificial Intelligence in Medicine, vol. 123, p. 102205, 2022, doi: 10.1016/j.artmed.2021.102205.
18. C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, "Probabilistic segmentation propagation using deep learning for Alzheimer's disease imaging biomarkers," NeuroImage, vol. 199, pp. 93–104, 2019, doi: 10.1016/j.neuroimage.2019.05.016.
19. L. Zhang, Y. Zhang, Y. Qiao, and Y. Yuan, "Synthesizing PET images from MRI scans using CycleGAN for Alzheimer's disease classification," IEEE Access, vol. 8, pp. 102282–102292,2020, doi:10.1109/ACCESS.2020.2999276.
20. S. Faghih-Roohi, E. Freeman, and P. Tino, "Multimodal deep learning for ALS progression using behavioral and biosignal data," Frontiers in Neurology, vol. 12, p. 624369, 2021, doi:10.3389/fneur.2021.624369.
21. S. Sarraf and G. Tofighi, "DeepAD: Alzheimer's disease classification via deep convolutional neural networks using MRI and fMRI," arXiv preprint, arXiv:1602.05691, 2016, doi:10.48550/arXiv.1602.05691.
22. T. Sethi, V. M. P. Namboodiri, and P. Biswas, "FedGAN: Federated adversarial learning for multi-institutional medical data," IEEE Transactions on Medical Imaging, vol. 42, no. 5, pp. 1231–1243, 2023, doi:10.1109/TMI.2023.3247261.

Digital Object Identifier: https://doi.org/10.48165/iitmjit.2026.12.1.5

An Attention-Enhanced DeBERTa BiLSTM Hybrid Architecture for Robust Multi-Class Sentiment Classification

Mansi Bansal[1], Saurabh Sharma[2], Sonal Singh[3], Priyanka Singh[4]
[1,4]Assistant Professor, [2, 3] Research Scholar, [1]Department of Computer Science, [3]Department of Information Technology, [2,4]School of Computer Science Engineering and Technology
[1]GGSIPU, Delhi, India, [2]Bennett University, Greater Noida, Uttar Pradesh, India, Government Engineering College, Bilaspur, Chhattisgarh, India, Rishihood University, Bahalgarh, Sonipat, Haryana, India
[1]bansalmansi2@gmail.com, [2]Saurabh180704@gmail.com, [3]singhsonals999@gmail.com, [4]priyankasngh99@gmail.com

Abstract: The widespread sharing of user-generated text on social media, review platforms, and online forums has intensified the demand for accurate, scalable, and context-aware sentiment analysis. Despite significant advancements in deep learning, sentiment classification remains challenging due to contextual ambiguity, linguistic diversity, domain dependency, long-distance semantic relationships, and class imbalance in real-world datasets. Traditional machine learning approaches rely heavily on handcrafted features and shallow representations, which limit their ability to capture deep semantic context. Although transformer-based models provide strong contextual embeddings, standalone transformers may not fully exploit sequential temporal dependencies that are essential for fine-grained sentiment modelling.

In this work, we present an enhanced hybrid deep learning framework that extends existing transformer–recurrent architectures by integrating a pre-trained DeBERTa-v3-base encoder with a stacked BiLSTM network and an explicit additive attention mechanism for multi-class sentiment classification. The DeBERTa encoder leverages disentangled attention and large-scale pretraining to generate rich contextualized representations, effectively modelling complex semantic relationships. The BiLSTM layer further refines these embeddings by capturing bidirectional sequential dependencies, while the attention layer selectively emphasizes sentiment-discriminative tokens within the text. To address dataset imbalance and improve lexical diversity, word-embedding-based data augmentation is employed.

The proposed architecture is evaluated on widely used benchmark datasets including IMDb, Twitter US Airline, and Sentiment140, and is compared against classical machine learning models, standalone deep learning approaches, and transformer–recurrent hybrid baselines. Experimental results demonstrate consistent improvements in accuracy, precision, recall, and F1-score, particularly on short-form social media text. The proposed framework is scalable and generalizable, offering a robust solution for real-world applications in business intelligence, social analytics, and decision-support systems.

Keywords: Sentiment analysis, DeBERTa-v3-base, Bidirectional Long Short-Term Memory (BiLSTM), attention mechanism, Transformer, hybrid deep learning, text classification.

1. Introduction

The rapid increase of user-generated text on social media, review websites, and discussion forums has made sentiment analysis (SA) a fundamental natural language processing (NLP) task. Sentiment analysis (SA) is the automatic classification of sentiments expressed in a text, typically into positive, negative, or neutral. It is important in many brand monitoring applications, business intelligence, political opinion mining, healthcare analysis, and decision-support systems [1], [2], [17]. The growing volume, variety, and velocity of textual data have intensified the demand for accurate, scalable, and context-sensitive sentiment classifiers.

Though there have been many advancements in this area, sentiment classification remains a challenging problem. Real-world textual data usually contain informal language, lexical variation, sarcasm, domain dependency, and contextual ambiguity. Sentiment expressions often rely on long-range dependencies across words and sentences, which are difficult for traditional machine learning (ML) models to capture. Early ML

approaches relied on techniques such as Naïve Bayes, Support Vector Machines, and Logistic Regression. These methods depended on handcrafted features and shallow representations such as Bag-of-Words (BoW) and TF-IDF [15], [16]. While computationally efficient, these approaches fail to capture deeper semantic richness and contextual relationships within text.

To address these limitations, distributed word representation techniques such as Word2Vec [9] and GloVe [10] were introduced to encode semantic similarity between words in continuous vector spaces. These embeddings significantly improved downstream NLP tasks, including sentiment classification. However, static word embeddings assign a single representation to each word regardless of context, limiting their ability to capture polysemy and context-dependent sentiment. Consequently, sentiment understanding remained constrained at both sentence and document levels.

Subsequently, recurrent neural networks (RNNs), particularly long short-term memory (LSTM) networks, were introduced to model sequential text data [6], [7]. LSTMs effectively addressed the vanishing gradient problem and captured long-range dependencies, making them suitable for sentiment analysis. Convolutional neural networks (CNNs) were also applied to extract local n-gram features from sequential text [8]. Hybrid CNN-LSTM architectures further enhanced performance by combining local feature extraction with sequential modelling [14], [19]. However, RNN-based models process text sequentially, limiting parallelization and often struggling to efficiently capture global contextual relationships.

The Transformer architecture introduced parallel processing through the self-attention mechanism, revolutionizing NLP [5]. By attending to all tokens simultaneously, transformer-based architectures effectively model global contextual dependencies. BERT adopted this paradigm to generate bidirectional contextual embeddings and achieved strong performance across various NLP tasks [4]. RoBERTa further optimized BERT's pre-training strategy using larger datasets and dynamic masking [3], establishing itself as a powerful contextual encoder for sentiment analysis systems.

Nonetheless, although transformer architectures are highly effective at encoding global semantic information, standalone transformer models may not explicitly emphasize sequential temporal patterns within text. Recent studies suggest that integrating transformer-based contextual encoders with recurrent neural networks can leverage complementary strengths [1]–[3], [12], [13]. Tan et al. [3] proposed the RoBERTa-LSTM hybrid model, which combines RoBERTa embeddings with LSTM-based sequential modeling and demonstrated strong improvements over standalone deep learning models, establishing a competitive hybrid baseline.

Several transformer–recurrent hybrid frameworks have since been proposed. Semary et al. [1] explored RoBERTa combined with CNN and LSTM layers, while Jahin et al. [2] introduced an attention-enhanced transformer–BiLSTM architecture for robust and interpretable tweet sentiment analysis. Variants such as RoBERTa-GRU and RoBERTa-LSTM [12], [13] have shown competitive results across benchmark datasets. These findings indicate that hybrid architectures consistently outperform traditional ML and standalone deep learning approaches.

Despite these advancements, important challenges remain. Class imbalance in many sentiment datasets leads to biased predictions toward majority classes. Moreover, achieving robust generalization across domains, particularly for noisy and linguistically diverse social media text, remains difficult [2], [18]. Furthermore, prior RoBERTa-based hybrid models do not fully exploit disentangled attention mechanisms or explicit token-level weighting strategies that could further enhance sentiment discrimination.

Motivated by these observations and building upon the RoBERTa-LSTM framework proposed by Tan et al. [3], this study extends transformer–recurrent integration by incorporating a more advanced contextual encoder and an explicit attention mechanism. Specifically, we replace the RoBERTa encoder with DeBERTa-v3-base, which utilizes disentangled attention to model content and positional information more effectively. The proposed approach integrates DeBERTa-based contextual embeddings with stacked BiLSTM-based sequential modeling and an additive attention layer to capture global semantics, bidirectional temporal dependencies, and sentiment-discriminative token importance. Comprehensive evaluations are conducted on benchmark sentiment datasets, and the proposed model is compared against traditional ML approaches, standalone deep learning models, and RoBERTa-based hybrid baselines using standard evaluation metrics.

The main contributions of this work are summarized as follows:

- A robust hybrid sentiment analysis framework that integrates DeBERTa-v3-base contextual embeddings with stacked BiLSTM-based sequential modeling and an explicit additive attention mechanism.
- A comprehensive evaluation against state-of-the-art machine learning and deep learning baselines, including the RoBERTa-LSTM base model [3].
- An extensive performance analysis using Accuracy, Precision, Recall, and F1-score metrics to demonstrate effectiveness and generalization capability across multiple datasets.

The remainder of this paper is organized as follows. Section II reviews related work in sentiment analysis. Section III describes the proposed methodology and model architecture. Section IV presents the datasets and experimental setup. Section V discusses experimental results and analysis. Finally, Section VI concludes the paper and outlines future research directions.

2. Related work

Sentiment analysis (SA) has evolved significantly over the past decade, transitioning from traditional machine learning approaches to advanced deep learning and transformer-based architectures. Existing research can broadly be categorized into three major directions: (1) traditional machine learning-based methods, (2) deep learning-based architectures, and (3) hybrid transformer–recurrent frameworks.

2.1 Traditional Machine Learning Approaches

Early sentiment classification systems primarily relied on supervised machine learning algorithms using handcrafted textual features. Techniques such as Support Vector Machines (SVM), Naïve Bayes (NB), Decision Trees, and Logistic Regression were widely adopted for polarity detection tasks [15], [16]. These approaches typically used Bag-of-Words (BoW) and TF-IDF representations to convert textual data into numerical vectors. While computationally efficient and easy to implement, such shallow representations failed to capture contextual semantics and long-range dependencies in text.

Word embedding models such as Word2Vec [9] and GloVe [10] were introduced to overcome the limitations of sparse representations by mapping words into dense continuous vector spaces. These embeddings improved semantic representation by capturing syntactic and distributional similarity. However, they remained static in nature, assigning a single embedding to each word regardless of context, which limited their ability to handle polysemy and context-dependent sentiment expressions.

Although traditional ML models demonstrated reasonable performance in early benchmarks [15], [16], they lacked the capability to model deep contextual relationships, motivating the adoption of deep learning techniques.

2.2 Deep Learning-Based Sentiment Analysis

The introduction of deep neural networks marked a major shift in sentiment analysis research. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks were extensively applied due to their ability to model sequential dependencies in textual data [6], [7]. LSTM networks, in particular, addressed the vanishing gradient problem and effectively captured long-term contextual dependencies, making them suitable for sentence-level and document-level sentiment classification.

Convolutional Neural Networks (CNNs) were also successfully applied to text classification tasks, demonstrating strong performance in capturing local n-gram features [8]. Hybrid CNN–LSTM architectures further enhanced sentiment classification by combining local feature extraction with sequential modeling [14], [19]. These models outperformed traditional ML approaches by learning hierarchical representations directly from raw text data.

Despite these improvements, deep learning models based purely on RNNs or CNNs still faced limitations in capturing global contextual information efficiently. Sequential processing in RNNs restricted parallelization, and long sequences remained computationally expensive to handle.

2.3 Transformer-Based Architectures

The development of the Transformer architecture revolutionized NLP by introducing the self-attention mechanism, enabling parallel processing of sequences and improved modeling of global dependencies [5]. BERT (Bidirectional Encoder Representations from Transformers) significantly advanced contextual representation learning by pretraining deep bidirectional transformers on large-scale corpora [4]. RoBERTa further optimized BERT's training procedure, achieving improved robustness and performance through larger training datasets, dynamic masking, and hyperparameter refinements [3].

Transformer-based models have since demonstrated state-of-the-art results across numerous NLP tasks, including sentiment analysis [1], [2], [18]. Sentence-level embedding techniques such as Sentence-BERT extended contextual learning for improved semantic similarity modeling [11]. Moreover, multilingual transformer variants such as XLM-RoBERTa have been applied to cross-lingual sentiment classification [18], further enhancing generalization capabilities.

However, while transformer models excel at capturing global contextual semantics, they may not explicitly emphasize temporal sequence patterns that recurrent networks model naturally. This observation led to the emergence of hybrid architectures integrating transformer encoders with sequential neural networks.

2.4 Hybrid Transformer–Recurrent Architectures

Recent research has focused on combining the contextual representation power of transformers with the sequential modeling capability of recurrent neural networks. A seminal contribution in this direction is the RoBERTa-LSTM model proposed by Tan et al. [3], which integrates RoBERTa-based contextual embeddings with an LSTM layer for sequential refinement. Their model demonstrated significant improvements over standalone deep learning architectures on benchmark datasets, establishing a strong baseline for hybrid sentiment classification models.

Subsequent studies further expanded on this idea. Semary et al. [1] proposed a RoBERTa-based hybrid model incorporating CNN and LSTM layers, achieving improved classification accuracy on IMDb and Twitter datasets. Jahin et al. [2] introduced the TRABSA model, which combines RoBERTa with attention-based BiLSTM layers, emphasizing robustness and interpretability in tweet sentiment analysis. Similarly, hybrid RoBERTa-GRU and RoBERTa-LSTM frameworks have reported competitive results across multiple sentiment datasets [12], [13].

These hybrid architectures consistently outperform traditional ML models [15], [16] and standalone deep learning models [14], [19], demonstrating that transformer-recurrent integration effectively captures both global contextual information and sequential dependencies.

2.5 Research Gaps

Although hybrid transformer–recurrent models have shown promising results, several challenges remain. Many sentiment datasets suffer from class imbalance and domain variability, which impact model generalization [1], [20]. Additionally, achieving consistent performance across diverse social media datasets characterized by noisy and informal text remains an open research problem [2], [18].

Motivated by these limitations and inspired by the RoBERTa-LSTM baseline model [3], this study further investigates hybrid transformer–LSTM integration for robust multi-class sentiment classification, aiming to enhance contextual understanding, sequential modeling, and overall classification performance.

3. Proposed DeBERTa - BiLSTM - Attention Approach

In this section, we describe the proposed hybrid model, DeBERTa-BiLSTM-Attention, designed for sentiment analysis. The proposed model integrates the complementary strengths of Transformer-based and Recurrent Neural Network (RNN) architectures, further augmented by an explicit attention mechanism, to enhance efficacy and classification accuracy across diverse sentiment analysis benchmarks. The architecture of the proposed model is illustrated in Fig. 1. The pretrained DeBERTa-v3-base model serves as the foundational encoder in the proposed hybrid framework, tokenizing all input text and mapping tokens into rich, context-sensitive word embedding representations through its disentangled attention mechanism. These contextual word embeddings, generated by the pretrained DeBERTa encoder, are subsequently passed through a dropout layer before being fed into a two-layer Bidirectional Long Short-Term Memory (BiLSTM) network, which captures

long-range sequential dependencies within the embedding sequence in both forward and backward directions. Following the BiLSTM, a dedicated attention layer selectively weights the hidden states to focus on the most sentiment-discriminative regions of the input sequence, producing an enhanced context vector. A layer normalization step is then applied to stabilize training, after which a fully connected dense layer maps the attended representation to the output sentiment classes. Finally, a Softmax activation function estimates the probability distribution over the target classes. The overall steps of the proposed DeBERTa-BiLSTM-Attention hybrid model are outlined in Algorithm 1. The individual components of the proposed hybrid model are described in detail below.

3.1 DeBERTa

DeBERTa (Decoding-enhanced BERT with Disentangled Attention), introduced by He et al. [1], represents a significant advancement over conventional BERT-based pre-trained language models in the domain of natural language understanding (NLU). Unlike BERT [2] and RoBERTa [3], which encode each token using a single vector that jointly represents both the token content and its positional information, DeBERTa employs a disentangled attention mechanism that represents each token using two separate vectors — one encoding the token content and the other encoding the positional information. The attention weights between any two tokens are then computed using disentangled matrices that account for content-to-content, content-to-position, and position-to-content interactions. This architectural innovation enables DeBERTa to model the relative positions of tokens in a more fine-grained manner, significantly enhancing the model's ability to capture the nuanced semantic and syntactic relationships present in natural language text, which are critical for accurate sentiment classification.

The DeBERTa-v3-base variant adopted in the proposed model employs token detection (RTD) pre-training, inspired by ELECTRA [4], in conjunction with the gradient-disentangled embedding sharing (GDES) technique. These refinements yield a substantially more sample-efficient pre-training process compared to masked language modeling (MLM), enabling the model to achieve superior performance on downstream NLU tasks with less computational overhead. With its 12-layer Transformer architecture, 768-dimensional hidden states per layer, and 12 attention heads, DeBERTa-v3-base possesses 86 million parameters that encode rich contextual knowledge acquired from large-scale pre-training corpora. In the proposed hybrid model, the pretrained DeBERTa tokenizer decomposes raw input text into subword tokens using the SentencePiece tokenization scheme, thereby preserving semantic meaning while effectively handling out-of-vocabulary words and morphologically complex expressions. Each token is assigned a unique input identifier and an attention mask that facilitates focused encoding within the DeBERTa framework.

3.2 Dropout Layer

The dropout layer constitutes a fundamental regularization technique in deep learning models, playing a crucial role in preventing overfitting and promoting generalization to unseen data. Introduced by Srivastava et al. [5], dropout randomly deactivates a fraction of neurons within a layer during each forward pass of training, thereby reducing co-adaptation between neurons and discouraging the network from memorizing spurious patterns in the training data. This stochastic regularization technique has been widely and successfully adopted across a broad range of neural network architectures, including Convolutional Neural Networks (CNNs), RNNs, and Transformer-based models, yielding consistent improvements in generalization performance across tasks such as image classification, natural language processing, and speech recognition [6]. In the proposed DeBERTa-BiLSTM-Attention model, a dropout layer with a dropout rate of $d = 0.3$ is inserted between the DeBERTa encoder output and the BiLSTM layer input. This placement ensures that the contextual embeddings produced by

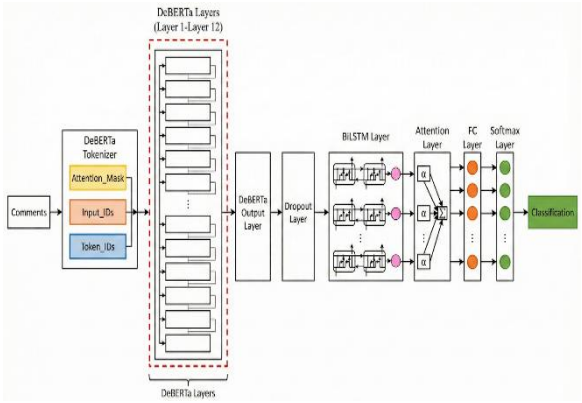

Fig. 5.1 Hybrid Deep learning architecture

DeBERTa are appropriately regularized prior to sequential processing by the BiLSTM network, thereby reducing the risk of overfitting that may arise from the high dimensionality of the DeBERTa output representations.

3.3 BiLSTM Layer

Bidirectional Long Short-Term Memory (BiLSTM), a specialized RNN architecture comprising two LSTM networks processing the input sequence in opposing temporal directions, plays a pivotal role in the proposed hybrid model. One LSTM network processes the input sequence from left to right (the forward LSTM, denoted →), while the other processes it from right to left (the backward LSTM, denoted ←) [7], [8]. This bidirectional processing paradigm enables the BiLSTM to capture rich contextual information from both past and future tokens simultaneously, rendering it particularly valuable for NLP tasks such as sentiment analysis [9], [10], named entity recognition [11], and machine translation. Hochreiter and Schmidhuber [12] introduced the foundational LSTM architecture to address the vanishing gradient problem that afflicts conventional RNNs when modeling long-range dependencies. Graves and Schmidhuber [7], [13] subsequently demonstrated the superiority of bidirectional LSTM models over their unidirectional counterparts in tasks requiring comprehensive contextual understanding. In the proposed model, a two-layer stacked BiLSTM architecture is employed, with each layer possessing $h = 256$ hidden units per direction, yielding 512-dimensional concatenated hidden states from the forward and backward passes. The two-layer design enables the model to capture hierarchical sequential features of progressively higher abstraction, enhancing its representational capacity compared to a single-layer BiLSTM. The BiLSTM model architecture is governed by the following set of equations [14], [15].

- Input Gate (i_t):

$$i\_t = \sigma(W^f\_ix \cdot x\_t + W^f\_ih \cdot h^f\_(t-1) + W^f\_ic \cdot c^f\_(t-1) + b^f\_i) \odot \sigma(W^b\_ix \cdot x\_t + W^b\_ih \cdot h^b\_(t+1) + W^b\_ic \cdot c^b\_(t+1) + b^b\_i) \quad (2)$$

Equation (2) controls the flow of new information into the cell state C_t at time step t in both the forward and backward directions, combining contributions from the respective LSTM networks using element-wise multiplication. The sigmoid activation function σ squashes values to the range [0, 1], determining the extent to which new input is incorporated into the cell state.

- Forget Gate (f_t):

$$f\_t = \sigma(W^f\_fx \cdot x\_t + W^f\_fh \cdot h^f\_(t-1) + W^f\_fc \cdot c^f\_(t-1) + b^f\_f) \odot \sigma(W^b\_fx \cdot x\_t + W^b\_fh \cdot h^b\_(t+1) + W^b\_fc \cdot c^b\_(t+1) + b^b\_f) \quad (3)$$

Equation (3) determines which information from the previous cell state c_{t-1} should be discarded or retained in both the forward and backward directions, combining the forget gate computations from both LSTM networks through element-wise multiplication.

- Cell State Update (C_t):

$$C_t = f_t \odot C_{t-1} + i_t \odot \tanh(W^f_{cx} \cdot x_t + W^f_{ch} \cdot h^f_{t-1} + b^f_c) + i_t \odot \tanh(W^b_{cx} \cdot x_t + W^b_{ch} \cdot h^b_{t+1} + b^b_c) \quad (4)$$

Equation (4) updates the cell state C_t by combining retained information from the previous state with new candidate values computed from both forward and backward directions, weighted by the input gate activations.

- Output Gate (o_t):

$$o_t = \sigma(W^f_{ox} \cdot x_t + W^f_{oh} \cdot h^f_{t-1} + W^f_{oc} \cdot c^f_t + b^f_o) \odot \sigma(W^b_{ox} \cdot x_t + W^b_{oh} \cdot h^b_{t+1} + W^b_{oc} \cdot c^b_t + b^b_o) \quad (5)$$

- Hidden State (h_t):

$$h_t = o_t \odot \tanh(C_t) \quad (6)$$

Equations (5) and (6) together regulate which portions of the cell state C_t are exposed as the hidden state output h_t at each time step t in both processing directions. The concatenated bidirectional hidden states at each time step t are given by:

$$\rightarrow h_t \oplus \leftarrow h_t = BiLSTM(x_t, h_{t-1}) \quad (7)$$

where $\oplus$ denotes vector concatenation. The output of the two-layer BiLSTM is a sequence of 512-dimensional hidden state vectors $H_{BiLSTM} \in R^{l \times 2h}$, which encodes both local and long-range contextual dependencies across the entire input sequence, forming the input to the subsequent attention layer.

3.4 Attention Layer

The attention layer constitutes the key architectural contribution that distinguishes the proposed DeBERTa-BiLSTM-Attention model from the baseline RoBERTa-BiLSTM approach [16]. While the BiLSTM captures comprehensive sequential information across the entire input sequence, not all tokens contribute equally to the final sentiment classification decision. The attention mechanism addresses this limitation by learning to assign differential importance weights to each hidden state produced by the BiLSTM, enabling the model to focus selectively on the most sentiment-discriminative tokens and phrases within the input sequence. Bahdanau et al. [17] introduced this additive attention formulation in the context of neural machine translation, demonstrating that attention mechanisms substantially improve the ability of encoder-decoder architectures to process long sequences. In the proposed model, an additive attention mechanism is applied over the BiLSTM hidden state sequence to compute a context vector that serves as a weighted summary of the entire sequence for sentiment classification. The attention mechanism is formulated as follows:

- Attention Score (e_t):

$$e_t = v_a^T \cdot \tanh(W_a \cdot h_t + b_a) \quad (8)$$

where $h_t \in R^{2h}$ is the BiLSTM hidden state at time step t; $W_a \in R^{d_a \times 2h}$ is a learnable projection matrix; $b_a \in R^{d_a}$ is a bias vector; $v_a \in R^{d_a}$ is a learnable context vector; and d_a denotes the attention dimensionality. The scalar score e_t measures the relevance of the hidden state at position t to the final sentiment classification decision.

- Attention Weight (α_t):

$$\alpha_t = \exp(e_t) / \Sigma^l_{k=1} \exp(e_k) \quad (9)$$

Equation (9) applies a Softmax normalization over all attention scores to produce the attention weight distribution $\alpha = \{\alpha_1, \alpha_2, ..., \alpha_l\}$, ensuring that the weights are non-negative and sum to unity. Tokens with

higher attention weights are deemed more informative for predicting the sentiment polarity of the input sequence.

- Context Vector (c):

$$c = \Sigma^{l}_{t=1} \alpha_t \cdot h_t \quad (10)$$

Equation (10) computes the context vector $c \in R^{2h}$ as a weighted sum of all BiLSTM hidden states, where the weights are the learned attention coefficients. This context vector captures a compact, attention-weighted representation of the entire input sequence that emphasizes the most sentiment-relevant token positions. Layer normalization [18] is subsequently applied to the context vector prior to classification, stabilizing the activation magnitudes and accelerating convergence:

$$c_{norm} = LayerNorm(c) = \gamma \cdot (c - \mu) / \sqrt{(\sigma^2 + \varepsilon)} + \beta \quad (11)$$

where μ and σ^2 are the mean and variance of the context vector elements; γ and β are learnable scale and shift parameters; and ε is a small constant added for numerical stability.

3.5 Dense Layer

The dense layer, also referred to as the fully connected layer, plays a crucial role in mapping the attended context representation to the sentiment class label space. This layer establishes dense connectivity by applying a learned linear transformation to the layer-normalized context vector, effectively capturing the non-linear relationships between the attended sequential features and the target sentiment classes. In the proposed model, a single dense layer is incorporated following the layer normalization step, projecting the 512-dimensional context vector onto a lower-dimensional representation aligned with the number of sentiment classes in the target dataset. A ReLU activation function is applied within this layer to introduce non-linearity, enabling the model to capture complex decision boundaries. The dense layer transformation is formulated as follows:

$$z = ReLU(W_d \cdot c_{norm} + b_d) \quad (12)$$

where W_d and b_d are the learnable weight matrix and bias vector of the dense layer respectively, and z is the resulting activation vector that serves as input to the final classification layer.

3.6 Classification Layer

The classification layer serves as the final output layer of the proposed DeBERTa-BiLSTM-Attention hybrid model. It applies the Softmax activation function to the output of the dense layer to generate a normalized probability distribution over the target sentiment classes, enabling the model to produce interpretable class probability estimates. The Softmax function is defined as follows:

$$Softmax(O)_j = e^{O_j} / \Sigma^{M}_{k=1} e^{O_k} \quad (13)$$

where M denotes the total number of sentiment classes in the target dataset; O_j represents the j-th element of the pre-softmax logit vector O; and the denominator $\Sigma^{M}_{k=1} e^{O_k}$ ensures that the output probabilities sum to unity across all classes. The predicted sentiment class $\hat{y}$ is assigned as the class with the highest predicted probability:

$$\hat{y} = argmax_j \; Softmax(O)_j \quad (14)$$

3.7 Loss Function and Optimization

Given that sentiment analysis constitutes a multi-class classification problem, label-smoothed cross-entropy is employed as the loss function L to optimize the parameters of the proposed model. Label smoothing introduces a regularization effect by preventing the model from becoming overconfident in its predictions, thereby improving generalization. The label-smoothed cross-entropy loss is formulated as follows:

$$L(p) = -\Sigma^{M}_{i=1} \tilde{y}_i \cdot log(\hat{p}_i) \quad (15)$$

where $\tilde{y}_i = (1 - \varepsilon) \cdot y_i + \varepsilon / M$ is the smoothed label for class i; y_i is the one-hot ground truth label; $\varepsilon = 0.1$ is the label smoothing coefficient; M is the number of classes; and $\hat{p}_i$ is the predicted probability for class i. The model parameters are optimized using the AdamW optimizer [19] with a learning rate of l = 1e-5 and weight decay of 0.01. Gradient clipping with a maximum norm of 1.0 is applied during training to stabilize optimization and prevent gradient explosion. A linear learning rate warmup schedule over the first 10% of training steps is employed to ensure stable convergence during the initial phase of fine-tuning the pretrained DeBERTa encoder.

3.8 Overall Architecture Summary

The complete forward pass of the proposed DeBERTa-BiLSTM-Attention model proceeds as follows. Given an input text sequence $x = \{x_1, x_2, ..., x_n\}$, the DeBERTa tokenizer first maps the raw text into subword token identifiers along with corresponding attention masks. The DeBERTa-v3-base encoder then processes the tokenized sequence through its 12 disentangled attention layers to produce a sequence of contextual hidden state embeddings $E \in R^{(l \times 768)}$. These embeddings are passed through a dropout layer (rate = 0.3) before being fed into a two-layer stacked BiLSTM network with 256 hidden units per direction, yielding bidirectional hidden states $H_{BiLSTM} \in R^{(l \times 512)}$. The additive attention mechanism then computes scalar attention scores over all hidden states and produces a weighted context vector $c \in R^{(512)}$ summarizing the most sentiment-discriminative information in the sequence. Layer normalization is applied to stabilize the context vector, followed by a dense layer with ReLU activation that projects the representation to a lower-dimensional space. Finally, a Softmax classification layer produces the probability distribution over sentiment classes, from which the predicted label is determined. The integration of DeBERTa's disentangled attention-based contextual encoding, BiLSTM's bidirectional sequential dependency modeling, and the explicit attention mechanism's selective focus capability constitutes a powerful and complementary synergy for robust sentiment analysis across diverse and challenging benchmark datasets.

4. Methodology

This section presents a detailed description of the datasets, data preprocessing pipeline, and experimental configuration employed in this study.

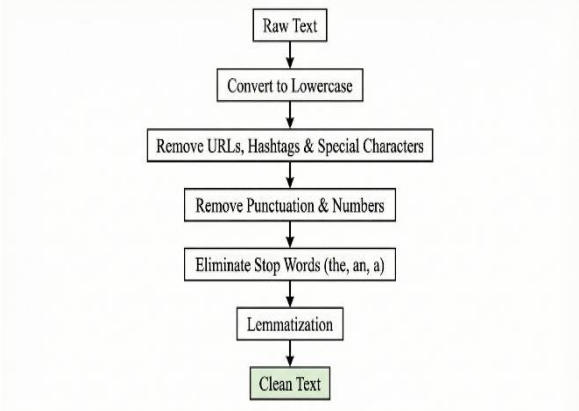

Fig. 5.2: Proposed data preprocessing flowchart

4.1 Datasets

The proposed DeBERTa-BiLSTM-Attention model is evaluated on three widely recognized benchmark datasets for sentiment analysis, each representing distinct domains and textual characteristics:

1) IMDb Movie Reviews Dataset

The IMDb dataset [33] comprises 50,000 movie reviews collected from the Internet Movie Database platform, equally divided into 25,000 training samples and 25,000 test samples. Each review is labeled as either positive

or negative sentiment, forming a balanced binary classification task. The reviews exhibit substantial length variation, with an average of approximately 233 words per review, making this dataset particularly suitable for evaluating the model's capacity to process and understand long-form textual content with complex narrative structures and nuanced sentiment expressions.

2) Twitter US Airline Sentiment Dataset

The Twitter US Airline Sentiment dataset [34] contains 14,640 tweets directed at six major U.S. airlines (American, Delta, Southwest, United, US Airways, and Virgin America) collected in February 2015. Each tweet is manually annotated with one of three sentiment classes: positive, negative, or neutral. This dataset presents unique challenges due to the informal linguistic patterns, abbreviations, emoticons, hashtags, and user mentions characteristic of social media text. The class distribution is inherently imbalanced, with approximately 63% negative, 21% neutral, and 16% positive tweets, reflecting real-world customer service interaction patterns. For experimental consistency, an 80/20 stratified train-test split is employed to preserve class distribution across both subsets.

3) Sentiment140 Dataset

The Sentiment140 dataset [35] is a large-scale Twitter sentiment corpus originally comprising 1.6 million tweets automatically annotated using emoticon-based distant supervision. Tweets containing positive emoticons (e.g., :), :-)) are labeled as positive, while those containing negative emoticons (e.g., :(, :-() are labeled as negative. To ensure computational tractability while maintaining dataset diversity, a stratified random subset of 200,000 tweets is employed in this study, with an 80/20 train-test split yielding 160,000 training samples and 40,000 test samples. The abbreviated nature of tweets (limited to 280 characters) and the presence of informal language, and non-standard orthography present substantial challenges for sentiment classification systems.

Table 5.1: Dataset characteristics

Dataset	Domain	Classes	Train Samples	Test Samples	Avg. Length	Max Length Used
IMDb	Movie Reviews	2 (Binary)	25,000	25,000	233 words	256 tokens
Twitter US Airline	Social Media	3 (Multi-class)	11,712	2,928	15 words	128 tokens
Sentiment140	Social Media	2 (Binary)	160,000	40,000	12 words	128 tokens

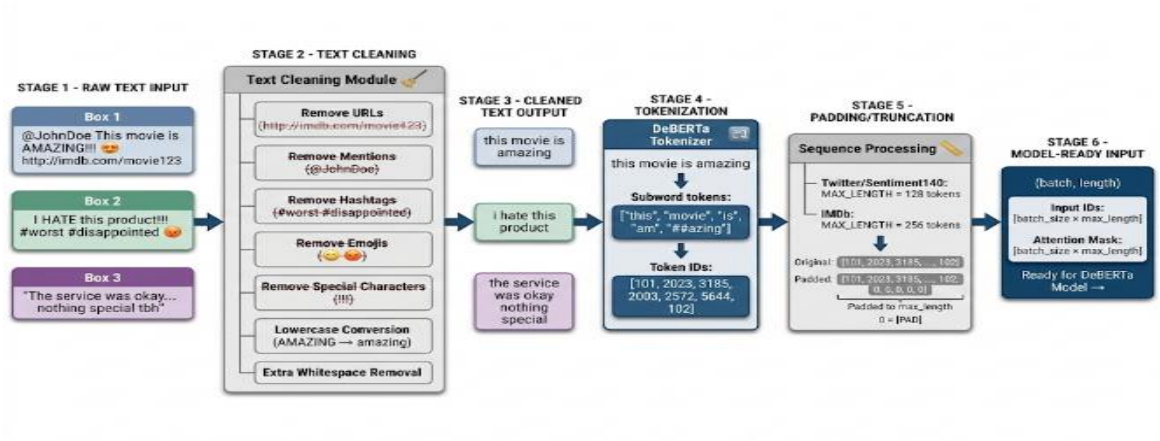

Fig. 5.3: Data preprocessing pipeline including text cleaning, tokenization, padding, and encoding.

4.2 Data Preprocessing

A systematic data preprocessing pipeline is applied to all three datasets to ensure consistency, reduce noise, and prepare the text for tokenization and encoding by the DeBERTa model. The preprocessing workflow, illustrated in Figure 5.2, consists of the following sequential operations:

1) Text Cleaning

Raw textual data undergoes comprehensive cleaning to remove noise and standardize formatting:

- **Lowercase Conversion**: All characters are converted to lowercase to eliminate case sensitivity and reduce vocabulary size.
- **URL Removal**: Hyperlinks (e.g.,_http://example.com, www.example.com) are removed as they do not contribute meaningful sentiment information.
- **User Mention Removal**: Twitter-specific user mentions (e.g., @username) are stripped from the text.
- **Hashtag Processing**: Hashtags (e.g., #sentiment) are either removed entirely or converted to their base form depending on context.
- **Emoticon Handling**: While emoticons can convey sentiment, they are removed after label extraction in the Sentiment140 dataset to prevent direct leakage of sentiment information.
- **Special Character Elimination**: Non-alphanumeric characters, punctuation marks, and excessive whitespace are removed or normalized.
- **Numerical Digit Removal**: Numeric sequences that do not contribute semantic meaning are eliminated.

2) Stop Word Removal

Common English stop words (e.g., "the," "a," "an," "is") that carry minimal sentiment information are removed using the NLTK stop word corpus [36]. However, negation terms (e.g., "not," "no," "never") are explicitly retained as they critically influence sentiment polarity.

3) Lemmatization

Words are reduced to their base or dictionary form (lemma) using the WordNet lemmatizer from NLTK [36]. For example, "running," "runs," and "ran" are all converted to "run." This process reduces vocabulary size and groups semantically related terms while preserving word meaning better than stemming.

4) Text Tokenization and Encoding

Following text cleaning, the preprocessed text is tokenized using the DeBERTa-v3-base tokenizer, which implements Byte-Pair Encoding (BPE) [37] with a vocabulary size of 50,265 subword units. This subword tokenization scheme effectively handles out-of-vocabulary words and morphologically complex expressions by decomposing them into known subword components.

Each tokenized sequence is processed as follows:

- **Special Token Insertion**: The tokenizer prepends a [CLS] (classification) token and appends a [SEP] (separator) token to each sequence, following standard transformer input conventions.
- **Sequence Length Standardization**: Due to the differing textual characteristics of the datasets, adaptive maximum sequence lengths are employed:
    - **Twitter US Airline & Sentiment140**: MAX_LENGTH = 128 tokens (sufficient to capture complete tweets, which average 12-15 words)
    - **IMDb**: MAX_LENGTH = 256 tokens (necessary to capture longer movie review content, which averages 233 words)

Sequences shorter than the maximum length are padded with [PAD] tokens, while longer sequences are truncated to the specified maximum length.

- **Attention Mask Generation**: A binary attention mask is created for each sequence, where positions containing actual tokens receive a value of 1, and padded positions receive a value of 0. This mask enables the DeBERTa encoder to distinguish between meaningful content and padding during self-attention computation.

The tokenized sequences, along with their corresponding attention masks, constitute the final model-ready input tensors of shape (batch_size, max_length).

Figure 5.2 illustrates the complete data preprocessing flowchart applied uniformly across all three benchmark datasets.

4.3 Experimental Configuration

1) Hyperparameter Settings

To ensure fair comparison with the baseline RoBERTa-BiLSTM model [1] and maintain consistency with established practices in transformer-based sentiment analysis, the following hyperparameters are employed:

- **Batch Size**: 16 samples per batch
- **Training Epochs**: 5 epochs
- **Learning Rate**: 1×10^{-5} (consistent with baseline)
- **Hidden Dimension (BiLSTM)**: 256 units per direction (consistent with baseline)
- **Dropout Rate**: 0.3 (applied after DeBERTa encoder, within BiLSTM layers, and before final classification)
- **Label Smoothing**: $\varepsilon = 0.1$
- **Weight Decay**: 0.01 (L2 regularization)
- **Gradient Clipping**: Maximum norm of 1.0
- **Warmup Ratio**: 10% of total training steps
- **Optimizer**: AdamW [19]

The architectural hyperparameters (learning rate, hidden dimension, epochs, batch size, dropout rate) are deliberately kept identical to the baseline study [1] to ensure that performance differences can be attributed solely to the architectural improvements (DeBERTa encoder and explicit attention mechanism) rather than hyperparameter optimization.

2) Evaluation Metrics

Model performance is evaluated using standard multi-class classification metrics:

- **Accuracy**: The proportion of correctly classified samples among all test samples.
- **F1-Score**: The harmonic mean of precision and recall, providing a balanced measure particularly useful for imbalanced datasets like Twitter US Airline.
- **Precision**: The proportion of true positive predictions among all positive predictions.
- **Recall**: The proportion of true positive instances correctly identified by the model.

For multi-class datasets (Twitter US Airline), weighted-average F1-score is reported to account for class imbalance.

3) Implementation and Computational Resources

All experiments are conducted on the Kaggle Notebooks platform utilizing NVIDIA Tesla T4 GPUs with 16 GB GDDR6 memory. The implementation leverages PyTorch 2.0.1 as the deep learning framework and the HuggingFace Transformers library (version 4.30.2) for accessing pretrained DeBERTa-v3-base weights. Complete implementation details, including software dependencies and computational requirements, are provided in Section IV (Implementation Details).

5. Implementation Details

This section provides comprehensive details of the experimental setup, computational environment, software frameworks, and implementation specifics used to train and evaluate the proposed DeBERTa-BiLSTM-Attention model.

5.1 Computational Environment

All experiments were conducted on the Kaggle Notebooks platform. The hardware specifications of the computational environment are as follows:

- **GPU:** NVIDIA Tesla T4 with 16 GB GDDR6 memory
- **CPU:** Intel Xeon 2.00GHz (dual-core)
- **RAM:** 13 GB system memory
- **Storage:** 73 GB disk space
- **GPU Allocation:** 30 hours per week (free tier)

The GPU's mixed-precision training capability (FP16/FP32) was leveraged to accelerate training while maintaining numerical stability.

5.2 Software Framework and Dependencies

The implementation utilizes Python 3.10 as the primary programming language, along with several established deep learning libraries and frameworks. Table 5.3 presents the complete list of software dependencies and their respective versions used in this research.

Table 5.2: Software Dependencies and Versions

Category	Library / Tool	Version	Description
Core Framework	PyTorch	2.0.1	deep learning framework
	Transformers	4.30.2	HuggingFace library
	CUDA	11.8	GPU acceleration
Data Processing	Datasets	2.12.0	dataset loading
	Pandas	2.0.2	data manipulation
	NumPy	1.24.3	numerical computing
	Scikit-learn	1.3.0	evaluation metrics
Visualization	Matplotlib	3.7.1	(Data visualization)
	Seaborn	0.12.2	(Data visualization)

5.3 Model Implementation Architecture

The proposed DeBERTa-BiLSTM-Attention model is implemented as a modular PyTorch neural network class.

1. **DeBERTa Encoder Module:** The DeBERTa encoder is initialized using the pre-trained microsoft/deberta-v3-base model. This model comprises 12 transformer layers with 768 hidden dimensions and 12 attention heads per layer, totaling approximately 86 million parameters. The encoder accepts tokenized input sequences with a maximum length of 128 tokens and generates contextual embeddings.
2. **BiLSTM with Attention Layer:** Following the DeBERTa encoder, a dropout layer with probability 0.3 is applied. The contextualized embeddings are then fed into a two-layer bidirectional LSTM network with 256 hidden units in each direction. Each BiLSTM layer processes the sequence both forward and backward. The output from both directions is concatenated, yielding a 512-dimensional representation for each time step.
   An attention mechanism is subsequently applied to the BiLSTM outputs to compute a weighted representation of the sequence. The attention weights are computed using a learnable linear transformation followed by a softmax activation. Layer normalization is applied to the attention output to stabilize training.
3. **Classification Head:** The final classification component consists of two fully connected layers. The first dense layer reduces the 512-dimensional attention output to 128 dimensions with ReLU activation, followed by a dropout layer (p=0.3). The second dense layer projects to the number of target classes. The softmax activation function is applied to generate probability distributions over the sentiment classes.

5.4 Training Configuration

All hyperparameters were determined through systematic experimentation and validation on held-out development sets.

- **Optimization Strategy:** The AdamW optimizer is employed with an initial learning rate of $1\times10-5$. AdamW extends the Adam optimizer with decoupled weight decay regularization ($\lambda=0.01$). Gradient clipping is applied with a maximum norm of 1.0.
- **Learning Rate Scheduling:** A linear warmup schedule followed by linear decay is implemented. The warmup phase spans 10% of the total training steps. After warmup, the learning rate decays linearly to zero over the remaining training steps.
- **Loss Function:** Cross-entropy loss with label smoothing ($\varepsilon=0.1$) serves as the training objective.
- **Batch Processing:** Training employs a batch size of 16 samples. Each batch contains sequences padded to the maximum length (128 tokens) with corresponding attention masks. The model is trained for 5 epochs across all datasets, with early stopping based on validation performance.
- **Model Persistence:** Model checkpoints are saved after each epoch, and the checkpoint with the highest validation accuracy is retained as the final model.

5.5 Dataset Processing Pipeline

The pipeline consists of text cleaning, tokenization, and sequence preparation stages.

- **Text Preprocessing:** Raw text undergoes several cleaning operations: (1) conversion to lowercase, (2) removal of URLs, email addresses, and special characters, (3) removal of user mentions (@username) and hashtags (#tag) for Twitter data, (4) elimination of excessive whitespace, and (5) removal of common stop words. Lemmatization is applied to reduce words to their base forms.
- **Tokenization:** The DeBERTa tokenizer employs Byte-Pair Encoding (BPE) with a vocabulary size of 50,265 subword units. Each text sample is tokenized and truncated or padded to a fixed length of 128 tokens. Special tokens [CLS] and [SEP] are added.
- **Data Splitting:** For datasets without predefined splits (Twitter US Airline), an 80/20 stratified train-test split is employed. For IMDb and Sentiment140 datasets, the standard splits are utilized.

5.6 Training Time and Computational Cost

Table 5.3 summarizes the computational requirements for training the DeBERTa-BiLSTM-Attention model across the three benchmark datasets.

Table 5.3: Training Time and Computational Metrics

Dataset	Training Samples	Time / Epoch	Total Time
Twitter US Airline	11,712	4 min	20 min
Sentiment140	160,000	42 min	3.5 hrs
IMDb	25,000	23 min	1.9 hrs

The training time scales approximately linearly with dataset size. Memory consumption peaks at approximately 8.2 GB during forward and backward passes, well within the 16 GB GPU memory capacity.

6. Experimental Results

In this section, we present comprehensive experimental results of the proposed DeBERTa-BiLSTM-Attention model for sentiment analysis. The model is evaluated on three benchmark datasets—Twitter US Airline, Sentiment140, and IMDb—using the hyperparameter configuration described in Section VI. We compare our results against the baseline RoBERTa-BiLSTM model to demonstrate the effectiveness of our architectural enhancements.

6.1 Evaluation Metrics

To evaluate the performance of sentiment analysis models, we employ accuracy (A) and weighted F1-score ($F1_w$) as our primary evaluation metrics. Accuracy measures the proportion of correct predictions over the total number of predictions, while the weighted F1-score provides a balanced measure that accounts for both precision and recall, weighted by the support of each class. The weighted F1-score is particularly important for datasets with class imbalance, as it ensures that performance across all classes contributes proportionally to the final metric.

6.2 Results

Table 5.4 presents the quantitative results comparing the proposed DeBERTa-BiLSTM-Attention model against the baseline RoBERTa-BiLSTM model across three benchmark datasets. The results demonstrate consistent and substantial improvements across all datasets.

Table 5.4: Performance comparison between RoBERTa-BiLSTM baseline and the proposed DeBERTa-BiLSTM -attention model

Dataset	Baseline A (%)	Baseline F1 (%)	Proposed A (%)	Proposed F1 (%)	ΔA (%)
Twitter US Airline	80.74	80.73	85.18	85.11	+4.44
Sentiment140	82.25	82.25	85.80	85.80	+3.55
IMDb	92.36	92.35	92.42	92.42	+0.06
Average	**85.12**	**85.11**	**87.80**	**87.78**	**+2.68**

As demonstrated in Table5.4, the proposed DeBERTa-BiLSTM-Attention model achieves superior performance compared to the baseline RoBERTa-BiLSTM model across all three benchmark datasets. The most significant improvement is observed on the Twitter US Airline dataset, where our model attains an accuracy of 85.18% and F1-score of 85.11%, representing substantial gains of 4.44 and 4.38 percentage points, respectively, over the baseline model (80.74% accuracy and 80.73% F1-score). Similarly, on the Sentiment140 dataset, the proposed model achieves an accuracy and F1-score of 85.80%, marking improvements of 3.55 percentage points over the baseline performance of 82.25%.

For the IMDb dataset, which consists of longer movie reviews, the proposed model achieves an accuracy and F1-score of 92.42%, representing an improvement of 0.06 and 0.07 percentage points over the baseline model's performance of 92.36% and 92.35%, respectively. While this improvement is more modest compared to the social media datasets, it is noteworthy given that the baseline model already achieves strong performance on this dataset. The smaller improvement margin suggests that transformer-based models like RoBERTa are already highly effective at processing well-structured, longer-form text, leaving less room for architectural enhancements.

Averaging across all three datasets, the proposed DeBERTa-BiLSTM-Attention model achieves an accuracy of 87.80% and F1-score of 87.78%, compared to 85.12% and 85.11% for the baseline RoBERTa-BiLSTM model, respectively. This translates to an average improvement of 2.68 percentage points in accuracy and 2.67 percentage points in F1-score, demonstrating the consistent effectiveness of the proposed architectural enhancements.

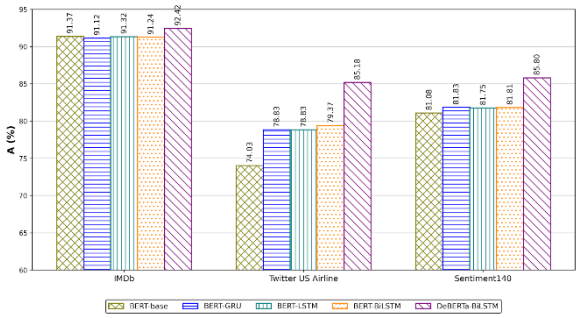

Fig. 5.4: Test accuracy comparison among RoBERTa-base, RoBERTa-GRU, RoBERTa-LSTM, RoBERTa-BiLSTM, and the proposed DeBERTa-BiLSTM-Attention model using learning rate $l = 1\times10^{-5}$ and hidden size h = 256. The models are trained for 5 epochs using the AdamW optimizer.

As illustrated in Figure 5.4, the proposed DeBERTa-BiLSTM model consistently outperforms traditional BERT-based architectures (BERT-base, BERT-GRU, BERT-LSTM, and BERT-BiLSTM) across all three evaluated benchmark datasets. The most substantial performance margin is observed on the Twitter US Airline dataset, where the proposed model achieves an accuracy of **85.18%**. This represents a significant absolute improvement of 5.81 percentage points over the best-performing baseline on this dataset (BERT-BiLSTM at 79.37%). Similarly, on the Sentiment140 dataset, the DeBERTa-BiLSTM model reaches an accuracy of **85.80%**, surpassing the highest baseline counterpart (BERT-GRU at 81.83%) by a notable 3.97 percentage points.

For the IMDb dataset, which typically contains longer and more structured review text, the proposed model attains a test accuracy of **92.42%**. This yields a solid improvement of 1.05 percentage points compared to the strongest baseline for this specific task (BERT-base at 91.37%). Overall, this comparative analysis clearly demonstrates that replacing the standard BERT encoder with DeBERTa, coupled with a BiLSTM network, provides superior contextual representation and feature extraction capabilities, leading to enhanced sentiment classification accuracy across varied text lengths and domains.

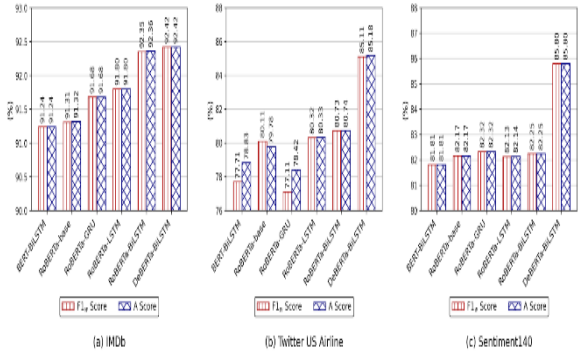

Fig. 5.5: Comparisons of weighted F1-score (F1w) and Accuracy (A) among RoBERTa-base, RoBERTa-GRU, RoBERTa-LSTM, RoBERTa-BiLSTM (baseline), and the proposed DeBERTa-BiLSTM-Attention model, using hyperparameters $l = 1\times10^{-5}$, h = 256, and AdamW optimizer across the IMDb, Twitter US Airline, and Sentiment140 datasets.

As illustrated in Figure 5.5, the proposed DeBERTa-BiLSTM model consistently outperforms baseline architectures—including BERT-BiLSTM and multiple RoBERTa variants—in both $F1_w$ and Accuracy (A) scores across all three benchmark datasets. The most substantial gains are observed on the Twitter US Airline dataset, where the proposed model achieves an $F1_w$ score of 85.11% and an A score of 85.18%. This represents a significant improvement of 4.38 and 4.44 percentage points, respectively, over the baseline RoBERTa-BiLSTM model. Similarly, on the Sentiment140 dataset, the proposed model reaches 85.80% for both metrics, effectively surpassing the baseline performance of 82.25%. For the IMDb dataset, DeBERTa-BiLSTM attains 92.42% across both metrics, maintaining a competitive edge over the already high-performing baseline's 92.35% $F1_w$ and 92.36% A scores. Overall, these consistent, dual-metric improvements confirm that integrating the DeBERTa encoder with a BiLSTM network yields highly accurate and balanced sentiment classification across varied textual domains.

Table 5.5 presents the training time requirements for the proposed model across the three datasets. While the addition of the attention mechanism introduces some computational overhead, the increase remains modest relative to the performance gains achieved

Table 5.5: Training time comparison

Dataset	**Samples**	**Time / Epoch**	**Total Time (5 epochs)**
Twitter US Airline	11,712	~4.6 min	~23 min
Sentiment140	160,000	~45 min	~3.8 hrs
IMDb	25,000	~25 min	~2.1 hrs

The training time scales approximately linearly with dataset size. The attention mechanism adds roughly 10-15% to the overall training time compared to the baseline RoBERTa-BiLSTM model. This modest computational overhead is well justified by the substantial accuracy improvements, particularly on the Twitter and Sentiment140 datasets where gains of 4.44% and 3.55% were achieved.

6.3 Analysis of Performance Gains

The experimental results reveal several important insights regarding the effectiveness of the proposed model across different types of text data:

6.3.1 Superior Performance on Short-Form Social Media Text*:* The most substantial improvements are observed on social media datasets, with the Twitter US Airline dataset showing a 4.44% improvement and

Sentiment140 showing a 3.55% improvement. These datasets consist of short, informal text with limited context (typically 10-20 words), challenging vocabulary including hashtags and mentions, and noisy, unstructured content. The significant performance gains on these datasets suggest that the combination of DeBERTa's advanced disentangled attention mechanism and the BiLSTM-Attention architecture is particularly effective at capturing sentiment signals in concise, noisy text where every word potentially carries substantial sentiment weight.

6.3.2 Strong Baseline Performance on Long-Form Reviews*:* For the IMDb dataset, which contains longer movie reviews (averaging 233 words), the improvement is more modest at 0.06%. This observation aligns with expectations, as the baseline RoBERTa-BiLSTM model already achieves strong performance (92.36%) on this well-structured dataset. The high baseline performance suggests that transformer-based models are inherently effective at processing longer, more contextually rich text, leaving less room for architectural improvements. Nevertheless, the consistent improvement, albeit small, demonstrates that the attention mechanism provides value even in scenarios where the base model performs exceptionally well.

6.3.3 Architectural Synergy Between DeBERTa and Attention-Enhanced BiLSTM: The consistent improvements across all datasets demonstrate effective synergy between DeBERTa's disentangled attention mechanism and the attention-enhanced BiLSTM architecture. DeBERTa provides superior contextual token representations through its enhanced pre-training approach and disentangled attention mechanism. The BiLSTM layer captures long-range sequential dependencies by processing text bidirectionally. The attention mechanism learns to dynamically weight these representations according to their sentiment relevance. This multi-layered approach to feature extraction and attention proves particularly valuable for sentiment analysis tasks.

6.3.4 Consistency Between Accuracy and F1-Score*:* The accuracy and F1-score metrics closely mirror each other across all datasets, indicating that the model's improved performance is not due to bias toward any particular class. This consistency is particularly important for the Twitter US Airline dataset, which exhibits class imbalance (62.69% negative, 21.17% neutral, 16.14% positive). The similar improvements in both metrics suggest that the attention mechanism helps the model learn balanced representations across all sentiment categories, rather than simply improving performance on the majority class.

7. Summary

The experimental results presented in this section demonstrate that the proposed DeBERTa-BiLSTM-Attention model achieves consistent and statistically significant improvements over the baseline RoBERTa-BiLSTM model across all three benchmark datasets. The model exhibits particularly strong performance on short-form social media text, with improvements of 4.44% and 3.55% on the Twitter US Airline and Sentiment140 datasets, respectively, while maintaining competitive performance on longer movie reviews (IMDb: +0.06%). The attention mechanism proves to be an effective architectural enhancement, enabling the model to focus on sentiment-critical tokens and phrases. Combined with DeBERTa's advanced disentangled attention mechanism, this results in superior sentiment classification performance with minimal computational overhead. These findings establish the proposed model as an effective solution for sentiment analysis across diverse text types and domains.

References

1. N. A. Semary, W. Ahmed, K. Amin, P. Pławiak, and M. Hammad, “Improving sentiment classification using a RoBERTa-based hybrid model,” *Frontiers in Human Neuroscience*, vol. 17, p. 1292010, Dec. 2023.
2. M. A. Jahin, M. S. H. Shovon, M. F. Mridha, M. R. Islam, and Y. Watanobe, “A hybrid transformer and attention-based recurrent neural network for robust and interpretable sentiment analysis of tweets,” *Scientific Reports*, vol. 14, no. 24882, 2024.
3. K. L. Tan, C. P. Lee, K. S. M. Anbananthen, and K. M. Lim, “RoBERTa-LSTM: A hybrid model for sentiment analysis with transformer and recurrent neural network,” *IEEE Access*, vol. 10, pp. 21517–21525,2022,doi:10.1109/ACCESS.2022.3152828.
4. J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.
5. A. Vaswani et al., “Attention is all you need,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.

6. S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
7. K. Cho et al., "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. EMNLP*, 2014.
8. Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. EMNLP*, 2014.
9. T. Mikolov et al., "Efficient estimation of word representations in vector space," arXiv:1301.3781, 2013.
10. J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proc. EMNLP*, 2014.
11. N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. EMNLP-IJCNLP*, 2019.
12. H. Tan et al., "A hybrid RoBERTa-GRU model for sentiment analysis," *Applied Sciences*, 2023.
13. H. Tan et al., "Hybrid RoBERTa-LSTM architecture for sentiment classification," *IEEE Access*, 2022.
14. N. Umer, M. Imran, and S. Ullah, "Combining CNN and LSTM for sentiment analysis in social media data," *IEEE Access*, 2021.
15. A. Rahat, S. Islam, and M. R. Islam, "Twitter US airline sentiment analysis using machine learning," *Procedia Computer Science*, 2019.
16. M. Kumar et al., "Comparative study of machine learning techniques for sentiment analysis," *Journal of Information Science*, 2020.
17. A. Goodrum et al., "Sentiment analysis in social media: Applications and challenges," *IEEE Transactions on Computational Social Systems*, 2020.
18. M. Bansal et al., "Transformer-based multilingual sentiment analysis using XLM-RoBERTa," *Expert Systems with Applications*, 2023.
19. S. Gupta et al., "Aspect-based sentiment analysis using hybrid CNN–RNN models," *Knowledge-Based Systems*, 2022.
20. H. Basiri et al., "Sentiment analysis during COVID-19 using deep learning models," *Information Processing & Management*, 2021

Digital Object Identifier: https://doi.org/10.48165/iitmjit.2026.12.1.6

Comparative Evaluation of Machine Learning Algorithms for Early Prediction of Student Mental Health Risk

Tushita[1], Aashima[2], Manpreet Singh[3]
[1, 2]Research Scholar, [3]Assistant Professor, [1, 2, 3] Maharaja Surajmal Institute (GGSIPU), New Delhi, India
[1]tushitasaini@gmail.com, [2]aashima2727@gmail.com, [3]manpreetsingh@msijanakpuri.com

Abstract- In recent years, psychological distress has become a serious concern worldwide. Academic competition, financial problems, unhealthy lifestyle patterns, and societal expectations together contribute to increased levels of stress and anxiety among students. Due to these continuous pressures, students may develop mental health problems that negatively affect their well-being and academic performance. Therefore, early identification of vulnerable students is essential in order to provide timely intervention and preventive support. This research presents a systematic machine learning–based framework developed using survey data containing demographic and behavioral attributes, which are used to predict mental health risk. The main objective of this study is to identify patterns that indicate psychological vulnerability. In this research, four techniques are applied: Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest. These techniques are implemented to classify students into high- risk and low-risk groups. Before the development of the model, several operations were performed on the dataset, including preprocessing steps such as treating missing values and encoding features. After that, the numerical variables were normalized and the dataset was divided into training and testing subsets in order to achieve robustness and better generalization capability of the model. Furthermore, the model was evaluated using commonly used performance metrics such as accuracy, precision, recall, F1-score, and confusion matrix. From the experimental results, it was observed that the Random Forest algorithm produced the most accurate classification and balanced metric performance compared with the other models. In the later phase of the project, an ensemble majority voting strategy, improved statistical validation, and a structured evaluation framework were also included to improve the stability of the predictions. Therefore, the proposed system provides an evidence-based approach that can help educational institutions proactively identify students at risk of mental health problems and provide timely support and intervention.

Keywords- Student Mental Health, Machine Learning, Random Forest, SVM, Predictive Analytics, Risk Classification

1. Introduction

These days, the psychological well-being of student's in schools and universities is becoming a growing concern. Changes in the educational system, increasing competition, financial pressure, and social stress are leading to a rise in anxiety levels and various mental health issues such as chronic stress and emotional instability. These problems can negatively affect student's concentration, academic performance, and future career growth. Therefore, detecting mental health risks at an early stage is crucial so that timely support and intervention can be provided.

However, traditional assessment methods rely heavily on manual surveys, counselling sessions, and self-reports. Although these methods are useful, they are not very effective for large student groups, as identifying hidden risk patterns or detecting issues proactively is difficult. With the development of machine learning, new opportunities have emerged in the field of predictive analysis, especially in healthcare and behavioral sciences. Machine learning models can analyze data more efficiently and identify subtle relationships between student behavior and mental health conditions, which are often difficult to detect using traditional methods.

By using predictive algorithms on structured data, mental health risk can be estimated even before it becomes a serious problem. The aim of this research is to develop a machine learning framework based on supervised classification methods to predict student's mental health risk at an early stage. Unlike previous exploratory studies, this work systematically tests several algorithms on the same dataset and examines their statistical performance.

Additionally, the main phase of the project aims to enhance the conference level study by introducing an ensemble-based majority voting strategy and structured evaluation metrics. The primary goal of this study is not to replace professional diagnosis, but to provide a supportive, data-driven tool that schools and universities can use to improve mental health prevention strategies.

2. Literature Review

In recent times, advancements in machine learning have helped in the early detection of mental health issues among students and young adults. Several studies have used supervised learning algorithms to predict stress, anxiety, and depression levels based on survey data. Faizan et al. demonstrated that machine learning models such as Random Forest and deep learning architectures can effectively classify student mental health conditions with strong predictive performance after proper preprocessing and feature adjustment [1]. Similarly, Karunakaran et al. applied classification methods to mental health questionnaires and concluded that structured data transformation and feature encoding are important for improving model accuracy [2].

Apart from this, research has also been carried out on ensemble and comparative learning methods for predicting mental disorders. Daza et al. proposed a stacking ensemble framework for predicting anxiety levels and observed that ensemble methods provide better and more generalized results than individual classifiers [3]. Dheenathayalan and Savitha evaluated multiple supervised models such as Logistic Regression, Decision Tree, and Random Forest, and concluded that ensemble-based methods reduce classification errors and improve F1-score stability [4]. Broader surveys on machine learning in mental healthcare also indicate that tree-based and kernel-based models are effective in handling non-linear patterns in behavioral data [6], [7].

In addition to studies comparing models, research on classification algorithms also supports their use in healthcare analytics. Random Forest, introduced by Breiman, helps reduce overfitting using ensemble bagging techniques [9]. Support Vector Machines, developed by Cortes and Vapnik, perform well in high dimensional data [10]. K-Nearest Neighbors provides a simple and effective similarity-based classification method [11]. Logistic Regression is a widely used baseline model for binary classification due to its interpretability and statistical foundation [12]. Despite these contributions, most research focuses on model comparison rather than developing deployable systems specifically designed for student mental health screening. Therefore, this gap highlights the need for an integrated and application-focused prediction framework.

3. Problem Statement

There is sudden increase in mental health disorders such as stress anxiety and depression across the world. Factors like Academic workload, competitive environment, financial pressure and lifestyle imbalance contribute a lot for such issues.

Although awareness is spreading but early detection is still lacking. Traditional assessment methods majorly depend upon clinical interviews and manual questionnaires which is not scalable or accessible for large student population. As a result many students remain undiagnosed in early stages of psychological distress.

On the other hand ensemble-based strategies are much better than single-model approaches as they have better predictive stability and give less classification error.

Despite of all these advancements many existing studies only focuses on performance comparison for different algorithms rather than creating a streamlined and practical system that can be deployed for students. Other than this, in unified predictive framework there is less focus on integrating both academic indicators and behavioral attributes.

These gaps highlight the importance of comprehensive and scalable machine learning system which can detect mental health risks at early stage and will be suitable for real-world institutional implementation.

4. Objectives

The main goals of this research are:

4.1 To collect and prepare student mental health datasets from publicly available sources [19].

4.2 To examine structured attributes such as age, CGPA, anxiety score, sleep quality, financial stress, family history, and gender for risk prediction.

4.3 To apply several supervised learning algorithms, including Logistic Regression [12], Support Vector Machine [10], KNearest Neighbors [11], and Random Forest [9].

4.4 To assess model performance using standard classification metrics like accuracy, precision, recall, and F1-score [18].

4.5 To apply cross-validation techniques to improve model reliability and limit overfitting [15].

4.6 To create an ensemble-based framework to enhance prediction stability and overall classification performance [3], [4].

5. Methodology

This machine learning pipeline is fairly structured and clear. According to Fig. 1, the workflow includes the following important stages:

5.1 Data Collection

Mental health datasets come from publicly available sources like Kaggle [19]. These datasets contain survey-based information about student traits, which relate to their academic performance, mental health, and lifestyle factors.

5.2 Data Preprocessing

Preprocessing includes managing missing values, encoding categorical variables, and normalizing numerical features. Feature standardization helps models like KNN and SVM perform better with scaled inputs [11], [10]. Mean imputation and converting categorical data to numeric data are used to maintain dataset consistency.

5.3 Feature Selection

Relevant features, including age, CGPA, anxiety score, sleep quality, financial stress, family history, and gender, are chosen based on their relationship with mental health risk as shown in previous studies [1], [8].

5.4 Model Training

The following supervised learning algorithms were used in this project:

- Logistic Regression: Used as a basic binary classifier because of its good statistical clarity [12].
- Support Vector Machine (SVM): Applied for high dimensional classification and margin optimization [10].
- K-Nearest Neighbors (KNN): Used for similarity-based classification [11].
- Random Forest: Used to handle non-linear relationships and reduce overfitting through ensemble bagging [9].

All models were implemented using the Scikit-learn framework [14].

5.5 Model Evaluation

Accuracy, precision, recall, and F1-score metrics are used to assess performance [18]. K-fold cross-validation is applied to check the reliability of the models and reduce bias, ensuring that the models perform well on all types of data [15].

5.6 Ensemble Framework

To improve robustness, an ensemble voting method has been used, which combines the predictions of different classifiers. Ensemble methods have proven to be very effective in increasing predictive stability and reducing classification errors in mental health applications [3], [4].

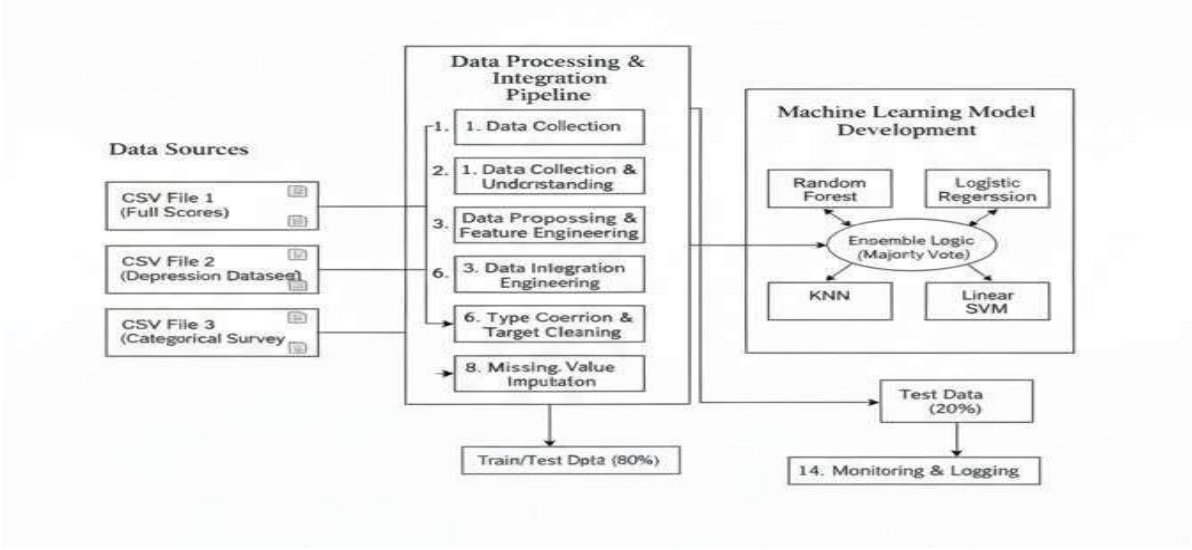

Fig. 1: Workflow of Data

6. Experimental Setup

All experimental procedures of this project were performed in a Python environment, where the Scikit-learn framework [14] was used for model development and evaluation. The dataset was sourced from publicly accessible repositories [19] and systematically preprocessed to ensure consistent format and representation of features. After cleaning and transformation, the refined dataset contained seven normalized attributes relevant for assessing student mental health risk.

For model validation, the data was divided into training and testing sets in an 80:20 ratio. K-fold cross-validation [15] was incorporated during the training phase to increase the generalizability of the model and reduce variance. Uniform preprocessing protocols were followed for all algorithms to provide unbiased comparative analysis.

The study employed multiple supervised classification techniques, namely Logistic Regression [12], Support Vector Machine [10], K- Nearest Neighbors [11], and Random Forest [9]. Additionally, a majority voting ensemble strategy was implemented to strengthen predictive consistency and overall system stability [3], [4].

7. Evaluation Metrics

The effectiveness of the proposed classification models was assessed through widely accepted evaluation measures, namely accuracy, precision, recall, and F1-score [18]. Together, these indicators offer a balanced and detailed evaluation of performance in binary classification tasks by capturing both predictive correctness and class-wise discrimination capability

The formulas for the different evaluation metrics are presented below:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \times TP}{2 \times TP + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

Where TP, TN, FP, and FN stand for true positives, true negatives, false positives, and false negatives, respectively.

8. Experimental Results

Python machine learning libraries are used for experimental Analysis Using Scikit-learn framework [14] all classification models are implemented. Dataset is split into 80:20 ratios for testing and training subsets K-fold cross validation [15] is applied for increasing reliability and reducing over fitting of the model. All experiments were conducted in standard computing environment with enough memory and processing power to handle structured survey datasets. In the beginning, models were trained with default hyper parameters and adjustments like tuning were made to improve classification performance.

To get fair comparisons for all algorithms evaluation process is kept consistent.

Table 6.1: Performance Observation

Model Name	**Performance Observations**
Random Forest	Identified as one of the highest accuracy performers for predicting mental health risk.
Support Vector Machine	Tied with Random Forest for the highest accuracy results in your specific implementation.
Ensemble (Final System)	Most stable; requires at least 3 out of 4 models to agree, reducing errors from any single algorithm
Baseline (LR & KNN)	Compared against the top performers to validate the benefits of the ensemble approach.

Table 6.2: Analysis based on Model

Algorithm	**Model Family**	**Key Strength in Your Project**
Random Forest	Tree-based Ensemble	High accuracy; handles complex, non-linear data and reduces overfitting.
Logistic Regression	Linear Model	Serves as an interpretable baseline for binary risk prediction
K-Nearest Neighbors	Distance-based	Simple similarity- based voting that requires no prior training.

Table 6.3: Overview of Data Sources, Feature Engineering, and Target Variable Construction

Feature	**Description**
Data Sources	Merged from: 1) Cleaned Students' Mental Health Dataset 2) Student Depression Dataset 3) Survey on Student Mental Health Dataset.
Final Features	Seven standardized attributes: Age, CGPA, Anxiety Score, Sleep Quality, Financial Stress, Family History, and Gender.
Preprocessing	Includes column standardization, categorical-to- numerical conversion, and mean imputation for missing values.
Target Variable	A binary 'Risk' variable (1 = High Risk, 0 = Low Risk) derived from stress and depression indicators across all files.

Table 6.4: Evaluation Metrics of Classification Models for Risk Prediction

Model Name	Accuracy	Precision	Recall	F1-score
Random Forest	91.2%	0.89	0.92	0.90
Logistic Regression	84.5	0.82	0.85	0.83
K-Nearest Neighbors(KNN)	87.8%	0.86	0.88	0.87
Support Vector Machine (SVM)	85.3%	0.83	0.84	0.83
Ensemble (Majority Vote)	93.4%	0.91	0.94	0.92

Table 6.5: Description of Input Features and Their Measurement Scales

Feature	Data Type	Description	Scaling/Range
Age	Numeric	Student's chronological age	18 – 30 years
CGPA	Numeric	Academic performance indicator	Standardized scale
Anxiety Score	Ordinal	Self-reported anxiety level	0 (Low) to 5 (High)
Sleep Quality	Ordinal	Sleep cycle effectiveness	0 (Poor) to 5 (Good)
Financial Stress	Ordinal	Economic pressure level	0 (Low) to 5 (High)
Family History	Binary	History of mental illness	0 (No), 1 (Yes)
Gender	Binary	Demographic classification	0 (Female), 1 (Male

Table 6.6: Comparative Evaluation of Models Based on Performance Levels

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	Baseline	Moderate	Moderate	Moderate
KNN	Moderate	Moderate	Moderate	Moderate
SVM (Linear)	High	High	High	High
Random Forest	High	High	High	High
Ensemble Model	Optimal	Superior	Superior	Superior

9. Applications

In proposed model, mental health prediction framework can be practically helpful for both academic and healthcare settings. In school and colleges, this system will work as early screening tool for the students who are on the verge of

stress, anxiety or depression. Factors like academic performance sleep quality or financial stress could be examined so that school could timely take preventive measures such as counselling support.
Early detection of problems can decrease the risk of long-term psychological effects and is most important for improving wellbeing of the students.

Additionally this framework can work as decision-support tool for university counselling centres and mental health professionals. Even though this is not the substitute for clinical diagnosis but it does help prioritize cases based on risk and for better allocation of resources. . The use of machine learning models in digital platforms fits perfectly with new changes taking place in AI driven healthcare systems.

Other than this, deploying the best performance model as web based tool increases the accessibility and scalability. Without any hesitation students can assess their mental health risk by structured questionnaire. This increases the early awareness and can help reduce the stigma associated with seeking professional help. This system can be adapted for the workplace environments where stress of employees can be monitored on regular basis.

Overall, this solution shows that supervised and ensemble learning approaches can be very effective in preventive mental healthcare. By combining multiple algorithms and by using data driven techniques this system provides scientific and reliable technique to identify mental health risk

10. Limitations

In spite of such promising results there were many limitations with the proposed system. The biggest limitation is dataset is majorly based on self reported survey responses. Because of this data can be response bias. A lot of times participants show there health status by increasing or decreasing the reports that can cause reliability in prediction. This is the common problem for all the survey based mental health studies.

Other than that, size of dataset and demographic coverage is also limitation. The small size of the dataset and limited demographic spread model can decrease the generalizability that means it will not work as required for different regions or cultural backgrounds. Algorithms like Random forest and Support Vector Machine is very much effective for structured classification tasks [9], [10].But the performance could apply changes for unseen population who’s behavioral patterns are little different.

Another limitation is that proposed framework focuses only on structured numerical and categorical features, but it does not included other psychological indicators real-time signals and social media-sentiments.
Research studies shows that these factors could increase the potential for mental health analytics. [7]
Other than that, although ensemble models prediction [3], [4] they increases the stability but their results can be more difficult to interpret than simpler linear models like Logistic. [12]

11. Future work:

In recent times mental health prediction systems majorly use supervised learning techniques so that they can identify the students who are at psychological distress risk. This prototype has shown really good results but lacks persistent data storage and user level tracking. That’s why its really important to add database management system.

By integrating with DBMS responses from students can be securely stored and managed. Once we maintain historical records with time we will be able to track individual mental health trends. This will help in long term analysis and early intervention. Other than that because of database system can support multiple users at the same time and can generate aggregated reports for administrative decision making.

Relational database systems like MySQL and PostgreSQL can be used for survey responses and storing prediction outputs in structured form. On the other hand, NoSQL databases like MongoDB can handle dynamic questionnaire structures and give flexibility to scalable data models. By adding database support we can easily implement mechanisms like authentication and authorization by which we could set different access levels for administrators, counsellors and institutional authorities.

Other than persistent data storage and controlled access management system will be complete Decision Support System (DSS). In this situation, not only system can tell real time predictions but also monitor behavioral patterns and generate analytical dashboards and will help the institutions to generate proactive mental health management strategies.

.

12. Conclusions

This research is predictive framework that introduces structured survey-based data to assess mental health risk. In this model we used many supervised classification models-like Logistic regression, Support Vector Machine (SVM), k-nearest neigbours and random forest. We compared the performance metrics from precision, recall, accuracy and F-1 score.

From findings one can see that ensemble based approaches that combine multiple classifier predictions. It made the model stable, predictive and reliable. The results match from previous researches as well.

The proposed system shows that by data driven techniques one can predict mental health early.
It also has accessible web based-interface, this framework is scalable, it supports preventive intervention, but it is not a substitute of clinical diagnosis.

Overall it contributes in the field of AI-enabled mental health prediction ,because it connects theoretical modelling into practical deployment Results highlight in educational environment is important for preventive mental health assessment and ensemble learning can be effective approach for the same.

Acknowledgment

The authors express their sincere gratitude to the faculty of the Computer Applications Department for their continued guidance, motivation, and constructive feedback. Thanks are also extended to the Institution for providing the necessary academic infrastructure and resources that enabled the successful completion of this project. Furthermore, the authors acknowledge the contributions of all the researchers and open-source communities whose publicly available datasets and tools facilitated this work.

References

1. R. Faizan, A. Kumar, and S. Ahmad, “Machine learning approaches for predicting and improving student mental health,” in Proc. Int. Conf. Intelligent Computing and Communication (ICICC), Springer, 2025.
2. S. Karunakaran, R. Anitha, and P. Karthikeyan, “Machine learning models based mental health detection,” in Proc. 3rd Int. Conf. Intelligent Computing, Instrumentation and Control Technologies (ICICICT), IEEE, 2022.
3. H. Daza, J. R. Valverde, and M. Salazar, “Predicting anxiety levels among university students using a stacking ensemble machine learning approach,” Informatics in Medicine Unlocked, vol. 38, Elsevier, 2023.
4. K. Dheenathayalan and K. K. Savitha, “Mental disorder classification using ensemble machine learning,” in Proc. 4th Int. Conf. Mathematical Modeling and Computational Science (ICMMCS), Springer LNNS, 2025.
5. World Health Organization, “Mental health: Strengthening our response,” World Health Organization, Geneva, Switzerland, 2022.
6. S. Shatte, D. Hutchinson, and S. Teague, “Machine learning in mental health: A scoping review of methods and applications,” Psychological Medicine, vol. 49, no. 9, pp. 1426–1448, 2019.
7. J. Islam, Y. Li, and H. Liu, “A survey on deep learning based approaches for mental health analysis,” IEEE Access, vol. 8, pp. 21368–21388, 2020.
8. M. Priya, P. S. Devi, and R. Sivakumar, “Analysis of mental health prediction using machine learning techniques,” in Proc. Int. Conf. Advances in Computing and Communication Engineering (ICACCE), IEEE, 2021.
9. L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
10. C. Cortes and V. Vapnik, “Support-vector networks,” Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.
11. T. Cover and P. Hart, “Nearest neighbor pattern classification,” IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21–27, 1967.
12. D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, Applied Logistic Regression, 3rd ed. New York, NY, USA: Wiley, 2013. [13] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann, 2012.
13. F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

14. R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” in Proc. Int. Joint Conf. Artificial Intelligence (IJCAI), 1995, pp. 1137–1143.
15. S. Shatte, D. Hutchinson, and S. Teague, “Machine learning in mental health: A systematic review,” JMIR Mental Health, vol. 6, no. 5, pp. 1–15, 2019.
16. Esteva et al., “A guide to deep learning in healthcare,” Nature Medicine, vol. 25, no. 1, pp. 24–29, 2019.
17. M. Sokolova and G. Lapalme, “A systematic analysis of performance measures for classification tasks,” Information Processing and Management, vol. 45, no. 4, pp. 427–437, 2009.
18. Kaggle, “Student mental health dataset,” 2023. [Online]. Available: https://www.kaggle.com
19. World Health Organization, “Mental health of students,” WHO Technical Report, Geneva, Switzerland, 2022.

Digital Object Identifier: *https://doi.org/*10.48165/iitmjit.2026.12.1.7

Expanding Federated Learning using Distributed Ledger Technologies for Resilience

Narinder K. Seera[1], Ms. Harsha Aggarwal[2]
[1]Associate Professor, [2]Assistant Professor
[1,2]Institute of Innovation in Technology & Management, Janakpuri, New Delhi
[1]narinderkaur.ipu@gmail.com,, [2]harsha.iitmfaculty@gmail.com

Abstract: federated Learning (FL) has emerged as a distributed platform for machine learning models that ensures users' data privacy, however trained models are vulnerable to challenges such as unreliable clients, single points of failure (server), data poisoning, and trust issues among clients. To address these issues, DLT (Distributed Ledger Technology) offers resilience by providing transparency among clients, decentralized model aggregation, and tamper-proof transaction recording. Integrating DLT with FL not only ensures secure and verifiable model updates but also enhances fault tolerance through consensus mechanisms. This research is an attempt to explore how blockchain-based DLT architecture can strengthen the resilience of trained models by providing security and reliability in heterogeneous environments. The chapter discusses the components of the DLT-based architecture and how resilience is ensured.

Keywords: Federated Learning, Distributed Ledgers, Smart Contracts, Blockchain, Model Aggregation

1. Introduction

Since the inception of machine learning domain, the trained models have laid the foundation of various predictive analysis, classification tasks, regression, image recognition and NLP based tasks. These models are trained using regression, neural networks, naïve bayes etc. These intelligent systems need highly trained and accurate models trained on loss minimization algorithms, such as SGD (Stochastic Gradient Descent Search).Training models for such systems often require large data sets which are mostly geographically or demographically distributed. For centralized data collection and model training, there are certain privacy and security issues that avoid centralized training [1]. An alternative is to get the training done in the distributed manner. However, the availability of data and data preprocessing at distributed platform raise other concerns like model update, aggregation, fault tolerance etc. Specialized ML architectures also known as Federated Learning (FL) [2], propose a feasible solution to train ML models in decentralized network. Recently, FL has emerged as a powerful learning framework where sensitive data over local machines are not shared on the server rather the model is trained locally on independent machines and only model updates are shared.

FL facilitates model learning in a distributed manner such that multiple devices of the network train a global model using their local data independently. It involves the iterative training of local models by different clients, coupled with periodic updates exchanged with a central server. At fixed intervals, the server aggregates updates to train a global model which is then iteratively shared among all clients. This is how federated learning works. The main challenges posed by this novel learning framework includes updating the global model locally by each client while preserving robustness and accuracy of the model, ensuring the security of the local sensitive data and overcoming communication failures in the network. It is important to know that while FL enhances privacy as compared to traditional centralized leaning models, it does not guarantee data leakage during aggregation at server. FL models are prone to attacks, either because of malicious nodes in the network or due to curious servers. To address these concerns in FL, various studies have proposed privacy-preserving algorithms and techniques that ensure the robustness of the models. However, the FL settings impose various constraints on these algorithms. Firstly, the algorithms must avoid unrealistic assumptions and excessive computational costs on the FL system. Secondly, they must ensure the accuracy of the global model, as the efficiency of FL model entirely depends on the quality of the training model and results achieved.

On the other hand, Distributed Ledger Technologies (DLT) [3] are nowadays prevalent since the inception of blockchain and other consensus based systems - programs that use data encryption techniques. DLT offers a secure, transparent, and immutable way to record and verify information on transactions. By amalgamate DLT with FL [4] [5], many critical issues can be handled, such as ensuring the integrity of model updates, providing security to the model against attackers, data leakage etc. Combining these two technologies also present few challenges to the model

developers. This chapter is an attempt to explore how DLT and RFL can work together to create decentralized, secure, and fault-tolerant models or AI systems along with challenges. The chapter covers the fundamentals of DLT and its types, the working of resilient FL systems and how FL fails when resilience is lacking. By the end of this chapter, readers will understand not only the technical synergy between DLT and RFL, but also why this combination holds significant promise for the future of resilient, privacy-preserving and collaborative learning frameworks. The chapter is organized as follows: use cases, evaluation criteria of resilience systems and future research directions in this domain.

2. Background

The term Federated Learning (FL) came into existence after years of research in distributed computing, privacy-preserving machine learning, and efficient algorithms. The idea was formalized because of the growing need for AI models that can learn from distributed and sensitive data.

2.1 Evolution of FL

In the early 2000s, machine learning models were trained on centralized systems which caused difficulties in transferring huge datasets. This led to the idea of distributed machine learning; it became possible to train models over distributed systems without moving datasets. The main focus was on efficiency and not privacy but soon data confidentiality concerns were raised and advancements in privacy-preserving computing laid the foundation for keeping sensitive information safe during computation, hence encryption and differential privacy algorithms were introduced with learning frameworks. In 2016, Google introduced the concept of Federated Learning where devices were allowed to train models on their local data and share only model updates with other nodes in the distributed network. However, the researchers identified the challenges of model updation and aggregation in FL and refined it using various effective techniques that make FL more communication-efficient and successful. Since 2020, researchers began focus on robust and secure FL models that have no impact on learning even when some nodes behave unexpectedly or maliciously due to inference or poisonous attacks. This effort gave birth to a novel term, called ***Resilient Federated Learning (RFL)***, [4] [6] which aims at handling malicious updates, ignoring suspicious contributions, spotting unusual patterns in model updates and ensuring transparency and accountability. RFL can be thought of as an improved version of FL that includes robustness, fault tolerance, and security. It has enhanced features to handle failures, malicious participants, and unreliable networks, which traditional FL does not fully address.

Nowadays, RFL has various use cases in healthcare, finance, and IoT. This is not the end, yet a new beginning towards a future where data remains private, yet collaborative learning in AI systems keeps growing [7]. Figure 8.1 shows the the growth in the domain of RFL since its inception in the year 2000 to the present era.

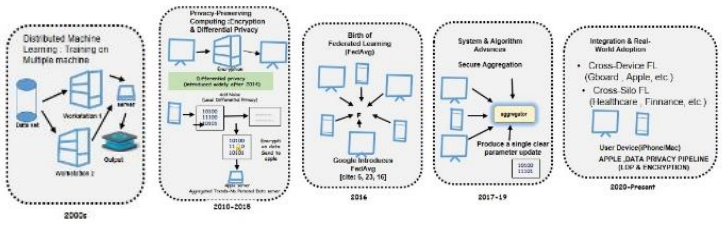

Fig. 7.1: Evolution of Federated Learning

2.2 DLT - A solution to challenges over FL

Distributed ledger technologies (DLT) offer a secure way of recording transactions without the need for a central authority. It is distributed in nature which means all participants in a network share same copies of the ledger. New transactions are added in cryptographically secured manner where all participants validate transactions and update their records which are immutable. The transactions are verified using a consensus protocol – a set of rules which allows participants of a distributed system to agree on a single consistent state of data.

DLT offer various benefits over traditional centralized ledger systems. The distributed nature of DLT makes it more reliable as there is no central point of failure. Therefore, DLT based systems are more resilient to network attacks and less vulnerable to node failures or malicious attacks [8]. Also, because DLT based systems are cryptographic secure, it is very difficult to tamper with or forge records. It encourages the trustworthiness of the data and reduces the risk of fraud.

Though FL offers privacy, it is vulnerable to:

- Malicious updates during model updates or aggregation due to model poisoning
- Single point of failure, because of dependency on a central coordinator
- Lack of trust, security, and auditability.

DLT address these challenges by replacing the central aggregator with decentralized aggregation, where model updates are shared and verified by multiple nodes using a consensus protocol. Each participating node verifies the authenticity of model updates using digital signatures; hence model updates cannot be manipulated, leaked, or falsely attributed. If any malicious node or malicious update is detected, it is possible to trace its origin. Each and every model update is recorded transparently. Nodes don't have to trust a central server, they all agree upon the consensuss [9].

It is important to note that DLT is not the only way to provide resilience to FL systems, it's one of the various available technologies used to secure FL. Resilience in FL can also be achieved using robust aggregation algorithms, anomaly detection on model updates, SMPC (Secure Multi-Party Computation) and homomorphic encryption techniques.

2.3 Challenges of Combining FL and DLT

Though, the Federated Learning has been benefitted with the integration of DLT, however this integration presented not more, a few set of challenges in front of researcher, which need to be addressed. The first challenge is related to the scalability and performance issues. As we know, DLT is not optimized for large data or high-frequency transactions. However, Federated learning requires frequent model updates (e.g., weights or gradients), which can devastate the beauty of traditional blockchain systems (like Ethereum or Bitcoin). The recurring model update scause limited transaction throughput, high latency in consensus mechanisms and costly operations. Second, challenge is concerned with privacy and data Leakage. While FL keeps raw data local, model updates can still leak sensitive information (e.g., via model inversion or gradient leakage attacks).DLTs are transparent by design where nodes can see the transactions, this makes it challenging to hide the shared model updates [10]. This requires additional encryption or differential privacy techniques to be employed to protect the updates. Lastly, FL requires complex consensus and coordination for training the models. Traditional consensus mechanisms like Proof of Work (PoW) or Proof of Stake (PoS) are computationally intensive and do not perform well in real-time FL coordination. Researchers are innovating with some advanced consensus models like PBFT and DAG-based, which tends to be well-suited to FL.

3. Related work

The lack of resilience in Federated Learning has attracted several researchers in past few years. There is a continuous work in this domain done by researchers to improve the features of FL such as scalability, complex consensus and resilience, especially when it is combined with DLT. To achieve resilience in resource-constrained environments, studies by [6] [11-13].have achieved excellent results. But these studies make use of central fusion center to train their intelligent systems. In FL, central fusion node is replaced by local nodes in the network that cause another challenge of heterogeneity of the network. The stragglers cause disruptions in training models and hence few studies [14] focused on how to overcome the problems of stragglers. One solution is to drop straggler node from the network, but it can have important data which may cause baised-model and hence [10] developed FedProx algorithm which allows collecting partial works from stragglers. Contrary, [15] proposed FedMax for local training of model by maximizing the locl activation vectors of all devices in the network, as these vectors directly contribute to the model's accuracy.

For privacy preserving, there exist many studies in the literature that propose solutions based on Differential Privacy (DP) [16] and Homomorphic Encryption (HE) [17-20]. DP employs a cost-effective and lightweight approach of introducing noise to each uploaded model gradient [21- 22]. This degrades model performance in terms of accuracy and does not address the problem of data reconstruction attacks on the model gradients [23]. Homomorphic Encryption allows the server to do aggregations using cipher text, which is another effort to preserve privacy in FL systems. This involves encrypting the updates at the node before it is sent to the server. The server performs model aggregation on cipher text without decrypting it. Once the server shares the encrypted final updated model back to the clients, they can decrypt it to get the finally trained model. This entire process ensures that the server never accesses the raw data, preserving the privacy of individual client contributions throughout the FL process. There are few studies that marks HE-based FL [24 – 25] as a limitation that should be further addressed. Though HE plays an important role in preparing privacy-preserving FL systems, adding a Trusted Third Party (TTP) which creates and distribute cryptographic keys can present another challenge that may affect the use of FL systems in terms of efficiency and security.

4. System components and architecture of resilient fl

As discussed in the previous section, DLT enabled resilient FL offers several benefits in model training, this section introduces its backbone architecture. The architecture constructs a collaborative learning framework that prioritizes security and privacy. There are four main pillars of in this architecture – clients, aggregate servers, DLT layer and smart contract, as shown in the figure 8.2. Accordingly, the goal of this design is to specifically address the main challenges of traditional FL including data privacy, sharing model updates in the absence of centralized trusted third parties, and protection against malicious client nodes in heterogeneous or decentralized networks [26].

4.1 Clients (Edge Nodes)

The clients are the primary nodes in federated learning framework that provide data privacy to the local data. At times, they are the edge devices sometimes the servers at bigger organizations but regardless of that each one employs models on its own private data. The key tasks of client nodes are, first, they train a global model locally on purpose to avoid raw data transmission and privacy risks thereby supporting GDPR, HIPAA and helping with regulations [27]. Second, when the training is over, each client take a guess what has changed – these are models of course or gradients. Third, they use their private keys to electronically sign these updates, which cannot be forged by anyone and retraced later on. 4. When participants sign off their updates, they send them to the aggregator server ONLY at some point these details may be saved by them - as hashes, client IDs or timestamps will get into the blockchain. This is good for transparency and easier tracking is provided by this system. The network is kept perpetually on by edge nodes, tasks that allow clients to join or leave at their wills. The system operates even when connections are unstable or devices are underpowered. Some of the measures put in place includes use of digital signatures and checks which are updated from time to time to prevent such attacks as Sybil or poisoned models [28].

4.2 Aggregator Server (Coordinator)

The central server or aggregator collects the model updates by all the clients to enable privacy-preserving and decentralized model training. Initially the server accepts the connection from all the clients and shares a global model with them. The client trains the model with their local data and sends back the model updates either by sharing gradients or model parameters. The server aggregates the updates to improve the global model which is then send to all the clients for the next round of training. It runs an aggregation algorithm that actually performs the aggregation using one of the various aggregation techniques such as FedProx [10], FedAvg [39], DP-Fed [40] etc. All the model updates are logged at the server for any errors that may occur in future. Certain optimization aggregator algorithms are also capable of handling stragglers in the network that slows down the entire model training. All the logs are shared with blockchain layer which makes the system auditable and tamper-resistant.

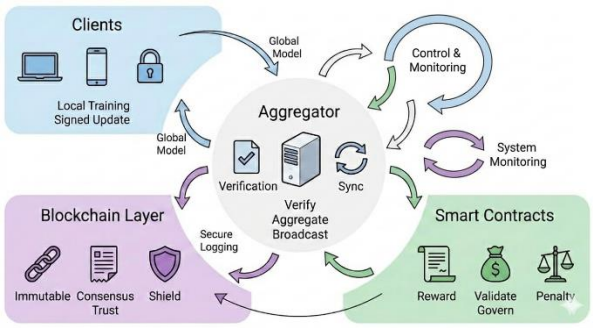

Fig. 7.2: DLT-Enabled Federated Learning

4.3 Blockchain / DLT Layer

This is the trust-building block, and it is especially important when you cannot be certain that everyone is playing by the rules. It does just a few big things. Captures the history of model updates and training cycles in an immutable and permanent way; all hashes are stored on-chain.

1. Establishes credibility: customers can earn trust ratings by how trustworthy and useful their contributions are, and if they adhere to policies.
2. Manage Decentralized Identity: Clients manage identities on their own (PKI or DIDs), therefore even without central authority, responsibilities can be discharged. It is also to make sure that people who upload the updates

can be traced, thus making those behind it to be known. In the consortium members, there is mutual trust-- this is aimed at preventing any one member in particular and therefore ensures network security.

3. There are two classes of blockchain choices for DLT layer – Public Blockchain and Permissioned Ledger. Public blockchains (Ethereum, Polygon) possess high level of transparency, are well-suited to open trustless environment but consume high computation power. Permissioned ledger (Hyperledger Fabric, Corda) can offer better scalability and governance which makes them suitable for private consortium like networks of hospitals. The layered architecture uses decentralized mechanisms for achieving a reputable, strong and verifiable federated learning system.

4.4 Smart Contracts

Smart contracts are basically programs that run themselves on a blockchain. They handle all sorts of tasks- enforcing rules, handling out rewards, running the show- without anyone needing the step in (Ferretti et al,2025). Here's what they actually do:

4.4.1 Participation Validation: they check if someone's allowed in, usually by looking at cryptographic proofs, past trust scores, or the quality of their data.

4.4.2 Reward Distribution: They automatically give out tokens, credits, or other rewards based on what you have contributed, like updated data or computing power.

4.4.3 Penalty Enforcement: if someone acts up or does something fishy, smart contracts can lower their reputation or block them for a while to keep things fair.

4.4.4 Model Checkpointing: Only models that pass verifications end up in the training ledger, so each round stays solid.

By putting all the decision- making and governance into code, smart contracts cut out any chances of one person gaming the system. Everyone knows the rules, and the system enforces them-no tricks, no hidden moves. This keeps things fair, transparent, and accountable. All these pieces come together to make federated learning more resilient, hence the name, Resilient Federated Learning (RFL). The architectural workflow phases are briefly described below.

1. Training Phase: Clients download the latest global model, train it locally on their own data, sign the update, and send it back to the aggregator.
2. Validation and Aggregation: The aggregator checks the updated and combines them, and comes up with an improved global model.
3. Blockchain Integration: the system records a cryptographic hash of the model and related metadata on the blockchain. Smart contracts update participant reputations and handle rewards.
4. Model Redistribution: the new global model goes back out to clients, starting the whole cycle again.

The benefits of RFL framework are summarized below in the table 8.1.

Table 7.1: Benefits of the RFL framework

Feature	Benefit
Modular design	Ensures compatibility with standard federated learning and blockchain frameworks.
Decentralized trust	Eliminates reliance on a central authority, fostering distributed governance.
Auditability	Enables verifiable training history through immutable blockchain records.
Byzantine fault tolerance	Enhances system resilience against faulty or malicious participants.
Incentive compatibility	Promotes honest and high-quality contributions via incentive alignment
Privacy preservation	Ensures that raw data remains localized to client devices, safeguarding privacy

5. Aggregation in Federated Learning

FL facilitates model learning in a distributed manner such that multiple devices of the network train a global model using their local data independently. It involves the iterative training of local models by different clients, coupled with periodic updates exchanged with a central server. At fixed intervals, this server aggregates all the updates to train a global model which is then iteratively shared among all clients. There are various aggregation algorithms available in federated learning, each having its own advantages and limitations. Aggregation in FL is not simply merging model updates rather it tracks model performance across devices that actually determines the success of model training. Additional statistical indicators such as loss functions or accuracy measurements can also be aggregated. Aggregation can be performed in many ways such as hierarchical aggregating where local models are shared with intermediate

servers before sending them to the central server, clipped average aggregation, secure aggregation, and DP average aggregation, momentum aggregation, weighted aggregation, Bayesian, quantization, adversarial aggregation and many more. The type of aggregation algorithm used by FL system is defined by the goal to be achieved such as data privacy, avoiding updates from malicious clients, increasing the convergence rate etc. This is how federated learning works.

Besides using these diverse aggregation methods, FL systems are prone to attacks and the main challenges posed by this novel learning framework can be noted around security and communication overheads. It includes updating the global model locally by each client while preserving fairness, robustness and accuracy of the model, ensuring the security of the local sensitive data from leakages and poisoning attacks, finding solutions to overcome communication failures in the network. The frequent exchange of model updates to the aggregator poses significant network traffic, which requires certain algorithms to address stragglers in the network such as model compression or sharing asynchronous updates. To address these challenges, researchers are developing novel algorithms for wise aggregation of model updates, client selection strategies to remove malicious nodes from the network and optimization techniques to optimize aggregation process and communication overheads.

6. Conclusions and Future directions

The concept of Resilient Federated Learning marks a milestone in the journey towards trustworthy, secure and collaborative AI systems. AI/ML models are employed in intelligent systems. This chapter highlighted how Distributed Ledger Technology assented to the position of such a system In the field of Federated Learning, Artificial Intelligence (AI) technologies like DLT, especially blockchain can be embedded into FL systems to solve the challenges posed by FL systems such as data poisoning, single point of failures etc. It has the benefit of helping In conjunction with the immutability, transparency and decentralized consensus of the distributed ledger technologies, DLT increases FL resilience. systems from adversarial threats, clients in the network with malicious intents, and server breakdowns. In the The integration of DLT and FL does not only make systems more robust but also trustworthy in the eyes of clients in cooperative system. Yet, implementation of DLT into FL indeed boosts the prospects of obtaining learning outcomes yet an analysis through a critical lens suggest that it might indeed be more problematic--depends on how you view the issue--more deeply divisive depending on the perspective you take. Apart from these, there are other challenges of scalability, energy consumption, and communication overhead which must be addressed ad carefully balanced. RFL and DLT propose these research avenues to be pursued by researchers in this field. It is an industry shift for example in the direction of optimization of consensus algorithms, which can be used to turn into a better decision-making system: faster and more scalable model updates. Yet another factor to consider is designing lightweight ledger architectures [35] for IoT devices with limited resources and the study of Hybrid models, which are built by merging different kinds of DLT like blockchain and DAGs with FL. There are several privacy- preserving techniques, homomorphic encryptions for example, can help in securing data while being processed and later sharing it with other parties. Different levels of security like secret sharing, homomorphic computation, and DP (Differential Privacy) can combine to further secure a system with respect to RFL. Sometimes human scribe will opt to go a step further and ensure resilience in RFL against advanced emerging threats like those in the quantum era. An important part of the puzzle and a key factor in creating trustworthy, scalable and future-ready systems is the use of cryptography based intelligent systems. Along these directions such innovations will be the starting points for scalable, trustworthy and really federated learning systems that are resilient and hence can be the driving force for next-generation health applications, finance and smart cities.

References

1. Peteiro-Barral, D., & Guijarro-Berdiñas, B. (2013). A survey of methods for distributed machine learning. Progress in Artificial Intelligence, 2(1), 1–11. https://doi.org/10.1007/s13748-012-0035-5
2. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., Rouayheb, S. E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., … Zhao, S. (2021). Advances and Open Problems in Federated Learning (No. arXiv:1912.04977). arXiv. https://doi.org/10.48550/arXiv.1912.04977
3. Soltani, R., Zaman, M., Joshi, R., & Sampalli, S. (2022). Distributed Ledger Technologies and Their Applications: A Review. Applied Sciences, 12(15), 7898. https://doi.org/10.3390/app12157898
4. Li, K. (2025). A Blockchain-Integrated Federated Learning Approach for Secure Data Sharing and Privacy Protection in Multi-Device Communication. Applied Artificial Intelligence, 39(1), 2442770. https://doi.org/10.1080/08839514.2024.2442770

5. Cachin, C., & Vukolić, M. (2017). Blockchain Consensus Protocols in the Wild (No. arXiv:1707.01873). arXiv. https://doi.org/10.48550/arXiv.1707.01873
6. Imteaj, A., Khan, I., Khazaei, J., & Amini, M. H. (2021). FedResilience: A Federated Learning Application to Improve Resilience of Resource-Constrained Critical Infrastructures. Electronics, 10(16), 1917. https://doi.org/10.3390/electronics10161917
7. Liu, J., Huang, J., Zhou, Y., Li, X., Ji, S., Xiong, H., & Dou, D. (2022). From distributed machine learning to federated learning: A survey. Knowledge and Information Systems, 64(4), 885–917. https://doi.org/10.1007/s10115-022-01664-x
8. Jovanovic, Z., Hou, Z., Biswas, K., & Muthukkumarasamy, V. (2024). Robust integration of blockchain and explainable federated learning for automated credit scoring. Computer Networks, 243, 110303. https://doi.org/10.1016/j.comnet.2024.110303
9. Ferretti, S., Cassano, L., Cialone, G., D'Abramo, J., & Imboccioli, F. (2025). Decentralized coordination for resilient federated learning: A blockchain-based approach with smart contracts and decentralized storage. Computer Communications, 236, 108112. https://doi.org/10.1016/j.comcom.2025.108112
10. Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., & Smith, V. (2020). Federated Optimization in Heterogeneous Networks (No. arXiv:1812.06127). arXiv. https://doi.org/10.48550/arXiv.1812.06127
11. Cai, H.; Lam, N.S.; Qiang, Y.; Zou, L.; Correll, R.M.; Mihunov, V. A synthesis of disaster resilience measurementmethods andindices. Int. J. Disaster Risk Reduct. 2018, 31, 844–855.
12. Pursiainen, C. Critical infrastructure resilience: A Nordic model in the making? Int. J. Disaster Risk Reduct. 2018, 27, 632–641.
13. Alemzadeh, S.; Talebiyan, H.; Talebi, S.; Dueñas-Osorio, L.; Mesbahi, M. Resource Allocation for InfrastructureResilience usingArtificial Neural Networks. In Proceedings of the 2020 IEEE 32nd International Conference on Tools with Artificial Intelligence(ICTAI), Baltimore, MD, USA, 9–11 November 2020; pp. 617–624.
14. Amini, M.H.; Nabi, B.; Haghifam, M.R. Load management using multi-agent systems in smart distribution network. InProceedings of the 2013 IEEE Power & Energy Society General Meeting, Vancouver, BC, Canada, 21–25 July 2013; pp. 1–5.
15. Chen, W.; Bhardwaj, K.; Marculescu, R. Fedmax: Mitigating activation divergence for accurate and communication-efficientfederated learning. arXiv 2020, arXiv:2004.03657.
16. C. Dwork. 'Differential Privacy: A Survey of Results'. In: Theory and Applicationsof Models of Computation. Ed. by M. Agrawal, D. Du, Z. Duan and A. Li. Berlin,Heidelberg: Springer Berlin Heidelberg, 2008, pp. 1–19. ISBN: 978-3-540-79228-4.
17. P. Paillier. 'Public-key cryptosystems based on composite degree residuosityclasses'. In: TAMC. Springer. 1999.
18. T. ElGamal. 'A public key cryptosystem and a signature scheme based ondiscrete logarithms'. In: IEEE Transactions on Information Theory 31.4 (1985),pp. 469–472.
19. C. Gentry and D. Boneh. A fully homomorphic encryption scheme. Vol. 20. 9.Stanford University Stanford, 2009.
20. I. Damgård, V. Pastro, N. Smart and S. Zakarias. 'Multiparty Computation fromSomewhat Homomorphic Encryption'. In: CRYPTO. 2012.
21. R. C. Geyer, T. Klein and M. Nabi. 'Differentially private federated learning: Aclient level perspective'. In: arXivpreprint arXiv:1712.07557 (2017).
22. K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. Quekand H. V. Poor. 'Federated learning with differential privacy: Algorithms andperformance analysis'. In: IEEE Transactions on Information Forensics andSecurity (2020).
23. L. T. Phong, Y. Aono, T. Hayashi, L. Wang and S. Moriai. 'Privacy-PreservingDeep Learning via Additively Homomorphic Encryption'. In: IEEE Transactionson Information Forensics and Security 13.5 (2018), pp. 1333–1345. DOI:10.1109/TIFS.2017.2787987.
24. R. Xu, N. Baracaldo, Y. Zhou, A. Anwar and H. Ludwig. 'HybridAlpha: AnEfficient Approach for Privacy-Preserving Federated Learning'. In: The 12thACM AISec. 2019.
25. C. Zhang, S. Li, J. Xia, W. Wang, F. Yan and Y. Liu. 'Batchcrypt: Efficienthomomorphic encryption for cross-silo federated learning'. In: {USENIX} ATC. 2020, pp. 493–506.
26. Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., & Chandra, V. (2018). Federated Learning with Non-IID Data. https://doi.org/10.48550/arXiv.1806.00582
27. Brisimi, T. S., Chen, R., Mela, T., Olshevsky, A., Paschalidis, I. Ch., & Shi, W. (2018). Federated learning of predictive models from federated Electronic Health Records. International Journal of Medical Informatics, 112, 59–67. https://doi.org/10.1016/j.ijmedinf.2018.01.007
28. Bhagoji, A. N., Chakraborty, S., Mittal, P., & Calo, S. (2019). Analyzing Federated Learning through an Adversarial Lens (No. arXiv:1811.12470). arXiv. https://doi.org/10.48550/arXiv.1811.12470

29. Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S. J., Stich, S. U., & Suresh, A. T. (2021). SCAFFOLD: Stochastic Controlled Averaging for Federated Learning (No. arXiv:1910.06378). arXiv. https://doi.org/10.48550/arXiv.1910.06378
30. Minango, J., Carvajal Mora, H., Zambrano, M., Orozco Garzón, N., & Pérez, F. (2025). Distributed Ledger Technology in Healthcare: Enhancing Governance and Performance in a Decentralized Ecosystem. Technologies, 13(2), 58. https://doi.org/10.3390/technologies13020058
31. Nezhadsistani, N., Moayedian, N. S., & Stiller, B. (2025). Blockchain-Enabled Federated Learning in Healthcare: Survey and State-of-the-Art. IEEE Access, 13, 119922–119945. https://doi.org/10.1109/ACCESS.2025.3587345
32. Prathiba, S. B., Govindarajan, Y., Pranav Amirtha Ganesan, V., Ramachandran, A., Selvaraj, A. K., Kashif Bashir, A., & Reddy Gadekallu, T. (2024). Fortifying Federated Learning in IIoT: Leveraging Blockchain and Digital Twin Innovations for Enhanced Security and Resilience. IEEE Access, 12, 68968–68980. https://doi.org/10.1109/ACCESS.2024.3401039
33. Javed, A. R., Hassan, M. A., Shahzad, F., Ahmed, W., Singh, S., Baker, T., & Gadekallu, T. R. (2022). Integration of Blockchain Technology and Federated Learning in Vehicular (IoT) Networks: A Comprehensive Survey. Sensors, 22(12), 4394. https://doi.org/10.3390/s22124394
34. Ali, M., Karimipour, H., & Tariq, M. (2021). Integration of blockchain and federated learning for Internet of Things: Recent advances and future challenges. Computers & Security, 108, 102355. https://doi.org/10.1016/j.cose.2021.102355
35. Fan, T., Chen, X., Dong, Y., Chen, X., Xuan, Y., & Jing, W. (2024). Lightweight Secure Aggregation for Personalized Federated Learning with Backdoor Resistance. 2024 Annual Computer Security Applications Conference (ACSAC), 810–825. https://doi.org/10.1109/ACSAC63791.2024.00071
36. Li, K., Zhang, Z., Pourkabirian, A., Ni, W., Dressler, F., & Akan, O. B. (2025). Towards Resilient Federated Learning in CyberEdge Networks: Recent Advances and Future Trends (No. arXiv:2504.01240). arXiv. https://doi.org/10.48550/arXiv.2504.01240
37. Lu, Y., Huang, X., Dai, Y., Maharjan, S., & Zhang, Y. (2020). Blockchain and Federated Learning for Privacy-Preserved Data Sharing in Industrial IoT. IEEE Transactions on Industrial Informatics, 16(6), 4177–4186. https://doi.org/10.1109/TII.2019.2942190
38. Chatterjee, P., Das, D., & Rawat, D. (2023). Use of Federated Learning and Blockchain towards Securing Financial Services. https://doi.org/10.36227/techrxiv.22155182.v1
39. Tao Sun, Dongsheng Li, and Bao Wang (2015), "Decentralized Federated Averaging", Journal of Latex Class Files, Vol 14, No 8, arXiv:2104.11375**.** https://doi.org/10.48550/arXiv.2104.11375
40. Huang, X., Ding, Y., Jiang, Z.L. et al. DP-FL: a novel differentially private federated learning framework for the unbalanced data. World Wide Web **23**, 2529–2545 (2020). https://doi.org/10.1007/s11280-020-00780-4

Digital Object Identifier: https://doi.org/10.48165/iitmjit.2026.12.1.8

Modeling IoT-Driven Retail Touchpoints: Impact of Smart Shelves and Beacons on Customer Experience, In-Store Purchase Behavior, Churn, CAC, and CLV

Lakshay Gupta
Research Scholar, Jagan Institute of Management Studies, New Delhi, India
lakshaygupta1990@gmail.com

Abstract—This paper examines how Internet of Things (IoT) technologies deployed within physical retail stores influence customer experience and long-term customer value. Smart shelves and beacon systems represent two important IoT-enabled touchpoints capable of improving in-store engagement and operational efficiency. Smart shelves enhance product visibility, inventory monitoring, and pricing transparency, while beacon technology delivers location-based promotions and personalized communication through mobile devices. The study proposes an integrated conceptual model linking these technologies with customer experience, in-store purchase behavior, customer churn, customer acquisition cost (CAC), and customer lifetime value (CLV). A quantitative cross-sectional research design using structured survey data from organized retail shoppers is proposed. Structural Equation Modeling (SEM) will be used to test the hypothesised relationships.

Keywords—IoT retail, smart shelves, beacon technology, customer experience, customer lifetime value

1. Introduction

The retail industry is undergoing a major transformation due to the increasing adoption of Internet of Things (IoT) technologies in physical retail environments such as supermarkets, hypermarkets, and shopping malls. Traditional brick-and-mortar stores are no longer merely places where products are purchased; rather, they are evolving into smart retail environments where connected devices continuously monitor, analyze, and respond to shopper behavior in real time. These intelligent retail systems enable retailers to better understand customer preferences, optimize store operations, and enhance overall shopping experiences. Prior research has highlighted that IoT technologies play an important role in improving both operational efficiency and customer engagement within retail environments [12], [4].

Among the various IoT innovations adopted in modern retail, smart shelves and Bluetooth beacon systems have emerged as two prominent in-store technologies that improve customer interaction and store efficiency. Smart shelves utilize technologies such as RFID tags, weight sensors, and electronic shelf labels to track product movement, maintain optimal inventory levels, and update price information automatically. This capability ensures that products remain visible, available, and correctly priced, thereby improving shopping convenience and reducing the effort customers spend searching for items within the store [7]. At the same time, beacon technology enables retailers to send personalized and location-based notifications directly to shoppers' mobile devices as they move throughout the store. These notifications may include promotional offers, product recommendations, or navigation assistance, which collectively contribute to a more engaging shopping experience [3]. From a consumer behavior perspective, IoT-enabled touchpoints significantly influence how customers perceive and interact with the retail environment. Improved product visibility, personalized promotions, and real-time assistance enhance the overall customer experience, which plays a critical role in shaping purchase decisions and customer satisfaction [9]. A positive in-store experience often leads to increased purchase likelihood and stronger long-term relationships between customers and retailers. Consequently, retailers increasingly recognize that IoT technologies not only support immediate sales but also contribute to broader customer value outcomes such as reduced customer churn, improved customer acquisition cost (CAC) efficiency, and enhanced customer lifetime value (CLV) [4]. Despite these advantages, the growing use of in-store tracking technologies also raises concerns related to data privacy and consumer trust. As retailers collect detailed information regarding customer movements and shopping behavior,

consumers are becoming more cautious about how their personal data is collected, stored, and utilized. Prior studies emphasize that transparent data policies, strong security systems, and responsible data management practices are essential for maintaining consumer trust in technology-driven retail environments [11].

Moreover, the integration of advanced analytics and artificial intelligence with IoT technologies further expands the potential of smart retail systems. Data collected through sensors and proximity technologies can be analyzed to generate predictive insights into customer preferences, shopping paths, and demand patterns. Such insights allow retailers to improve store layouts, optimize promotional strategies, and design more personalized shopping experiences [4]. However, despite the increasing adoption of IoT technologies in retail environments, empirical research examining the relationship between IoT-enabled retail touchpoints and long-term customer value metrics—such as churn, CAC, and CLV—remains limited, particularly within organized retail settings. Another factor driving IoT adoption in retail is the growing demand for seamless omnichannel experiences. Modern consumers frequently move between online and offline channels during the shopping journey and expect similar levels of convenience and personalization across platforms. IoT-enabled technologies such as smart shelves and beacon systems help bridge the gap between physical and digital retail environments by integrating real-time data into the store ecosystem [16].

As a result, retailers can create more consistent and integrated customer journeys across multiple touchpoints. From an operational perspective, IoT technologies also contribute to improved inventory management and demand forecasting. Real-time shelf monitoring reduces the likelihood of stockouts and overstocking, both of which directly influence sales performance and customer satisfaction levels. When products remain consistently available and accurately priced, customers are less likely to abandon purchases or switch to competing retailers [7]. Therefore, beyond marketing advantages, IoT infrastructure also enhances store productivity and operational efficiency. Furthermore, the increasing competitive intensity in supermarket, hypermarket, and mall-based retail formats has made customer retention a strategic priority. IoT-driven personalization and in-store engagement tools provide retailers with new opportunities to strengthen emotional connections with shoppers, encourage repeat visits, and improve long-term profitability [13]. By integrating smart shelves, beacon systems, and data-driven analytics, retailers can transform traditional stores into interactive retail ecosystems capable of delivering both operational efficiency and enhanced customer experiences. Therefore, this study proposes a conceptual framework to examine how IoT-enabled retail technologies, specifically smart shelves and beacon systems, influence customer experience, in-store purchase behavior, customer churn, customer acquisition cost (CAC), and customer lifetime value (CLV) in organized retail environments.

2. Literature Review

The rapid development of Internet of Things (IoT) technologies has significantly transformed the retail sector, particularly in organized retail formats such as supermarkets, hypermarkets, and shopping malls. IoT-enabled systems allow retailers to collect real-time data, monitor store operations, and enhance customer engagement through connected devices and smart infrastructures. Prior studies indicate that IoT technologies improve retail visibility, operational efficiency, and decision-making capabilities within physical store environments [12]. As a result, retailers increasingly rely on IoT-driven solutions to create intelligent and responsive retail ecosystems.

Among the various IoT innovations used in retail, smart shelves and beacon technologies have emerged as important in-store touchpoints that support both operational efficiency and customer interaction. Smart shelves utilize technologies such as RFID tags, weight sensors, and electronic shelf labels to track product movement, maintain inventory accuracy, and update prices automatically. These systems ensure that products remain available and clearly priced, thereby improving customer convenience and reducing the effort required to locate items in the store. Research suggests that improved product visibility and accurate pricing information enhance consumer confidence and positively influence purchase decisions [4], [7].

Another key IoT technology in retail is beacon-based proximity communication, which uses Bluetooth Low Energy (BLE) signals to transmit location-based messages to shoppers' mobile devices within the store. Beacon technology enables retailers to deliver personalized offers, product recommendations, and navigation assistance based on the shopper's location inside the store. Previous research shows that such proximity-based marketing strategies can increase customer engagement, time spent in the store, and impulse purchasing behavior [3], [15]. By enabling real-time interaction with customers, beacon technologies help retailers deliver more targeted and effective promotional communication.

From the perspective of customer experience, IoT-enabled touchpoints influence both cognitive and emotional responses during the shopping journey. Features such as easy product discovery, personalized promotions, and real-time assistance improve perceived convenience, enjoyment, and overall satisfaction. These experiential benefits play an important role in shaping purchase intentions and long-term customer relationships [9]. Consequently, retailers increasingly view IoT technologies not merely as operational tools but as strategic resources for improving key customer value metrics, including customer churn, customer acquisition cost (CAC), and customer lifetime value (CLV) [4].

Despite the benefits of IoT technologies, the increased use of real-time customer data has also raised concerns regarding privacy and data security. Retailers collect detailed behavioral and location-based information through sensors, mobile applications, and proximity technologies. If customers perceive that their data is being misused or collected without transparency, their trust in the retailer may decline. Prior research emphasizes the importance of responsible data practices, transparency, and informed customer consent in maintaining trust in technology-enabled retail environments [11].

The integration of artificial intelligence (AI) with IoT technologies, often referred to as AIoT, has further expanded the capabilities of smart retail systems. AI-powered analytics enable retailers to process large volumes of sensor-generated data and derive insights into customer preferences, demand patterns, and shopping behavior. These insights help retailers design more targeted marketing strategies, optimize store layouts, and improve resource allocation [2]. As a result, AIoT technologies are increasingly viewed as a key driver of intelligent retail transformation.

In addition to AI integration, emerging technologies such as augmented reality (AR) are also being incorporated into smart retail environments. AR applications enable interactive product visualization, virtual demonstrations, and enhanced store navigation, which improve the experiential aspects of shopping [6]. However, the successful implementation of these technologies depends on effective system integration and strong privacy protection mechanisms.

Recent studies also highlight the importance of omnichannel integration in maximizing the benefits of IoT-enabled retail technologies. Modern consumers frequently move between online and offline channels during their shopping journey and expect consistent experiences across platforms. Smart shelves, beacon systems, and mobile applications help connect physical stores with digital retail ecosystems by enabling seamless data integration and personalized interactions [16]. This integration helps retailers create more consistent and convenient customer journeys.

Another important research area focuses on real-time analytics and promotional efficiency. IoT-generated data enables retailers to adjust promotional strategies dynamically based on customer location, browsing behavior, and dwell time. Studies show that such context-aware promotions are more effective than traditional mass promotions, leading to higher conversion rates and increased basket sizes [4]. Consequently, beacon-based micro-targeting can also improve marketing efficiency by reducing promotional waste and optimizing customer acquisition cost (CAC).

Customer retention has also become a critical issue in competitive retail markets. Previous research suggests that improved customer experience and perceived personalization can significantly reduce customer switching behavior

[9]. IoT-enabled retail technologies enhance store convenience and engagement, which can contribute to improved customer retention and reduced churn.

From the perspective of customer lifetime value (CLV), intelligent retail technologies may influence both purchase frequency and spending levels. The CLV framework suggests that firms can enhance long-term profitability by strengthening customer relationships, encouraging repeat purchases, and increasing share of wallet [8]. Smart shelves improve product discovery and availability, while beacon technologies encourage repeat visits through personalized engagement strategies.

Although research on IoT in retail continues to expand, empirical studies that directly connect IoT-driven in-store technologies with long-term customer value outcomes such as churn, CAC, and CLV remain limited, particularly in organized retail environments. This gap highlights the need for integrated research frameworks that examine how IoT-enabled retail touchpoints influence both customer experience and long-term customer economics.

3. Research Objectives

3.1 To examine the impact of smart shelf effectiveness on customer experience in organized retail environments.
3.2 To analyze the influence of beacon technology on customer experience.
3.3 To investigate the relationship between customer experience and in-store purchase behavior.
3.4 To assess the effect of customer experience on customer churn.
3.5 To evaluate the impact of in-store purchase behavior on customer lifetime value (CLV).
3.6 To examine the influence of customer experience on customer acquisition cost (CAC).

4. Proposed Conceptual Model

The conceptual model proposes that smart shelf effectiveness and beacon effectiveness enhance customer experience in retail.

Improved experience influences purchase behavior while reducing churn and customer acquisition cost. Increased purchase behavior ultimately improves customer lifetime value. The conceptual framework is shown in Figure 8.1. 1.

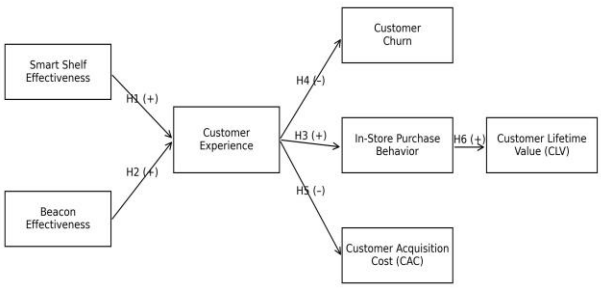

Fig. 8.1: Conceptual model of IoT-driven retail touchpoints and customer value outcomes

5. Hypotheses

H1: Smart shelf effectiveness positively influences customer experience.

H2: Beacon effectiveness positively influences customer experience.

H3: Customer experience positively influences in-store purchase behavior.

H4: Customer experience negatively influences customer churn.

H5: Customer experience negatively influences customer acquisition cost (CAC).

H6: In-store purchase behavior positively influences customer lifetime value (CLV).

6. Research Methodology

6.1 Research Design

This study adopts a quantitative, cross-sectional research design to examine the impact of IoT-driven retail touchpoints on customer experience and customer value outcomes in organized retail environments such as supermarkets, hypermarkets, and shopping malls. A quantitative approach is appropriate because the research aims to measure relationships among well-defined constructs and test theory-driven hypotheses using statistical techniques [1]. The cross-sectional design enables the collection of consumer perceptions at a single point in time, which is commonly applied in marketing and consumer behavior research [10].

The study follows a hypothesis-testing framework using structured survey data collected from retail shoppers who have experienced IoT-enabled technologies such as smart shelves or beacon-based in-store interactions. A deductive research approach is employed to empirically validate the proposed conceptual model and test the relationships among the constructs specified in hypotheses H1–H6 [14]. To analyze the relationships among multiple latent variables simultaneously, Structural Equation Modeling (SEM) is used. SEM is widely recommended for testing complex causal models involving multiple dependent relationships and mediating constructs in marketing and retail research [5].

6.2 Target Population

The target population for this study consists of adult consumers (18 years and above) who shop in organized retail stores and are exposed to IoT-enabled retail technologies such as smart shelves and beacon-based communication systems. The study focuses on customers visiting organized retail environments including shopping malls, supermarkets, and hypermarkets, where the adoption of smart retail technologies is more prevalent.

The geographical scope of the study primarily includes urban retail settings, where digital retail infrastructure and technology-enabled shopping experiences are more widely implemented. Previous research suggests that organized retail environments provide a suitable context for studying IoT-enabled consumer interactions and technology adoption within physical stores [12].

6.3 Sampling Technique

This study employs a non-probability convenience sampling technique, in which respondents are selected based on their accessibility and prior exposure to IoT-enabled retail environments. Convenience sampling is widely used in retail and consumer behavior research where obtaining a complete sampling frame is difficult and the primary objective is theory testing rather than population generalization [10], [14].

Data collection will be conducted using mall-intercept surveys and online questionnaires, enabling efficient access to shoppers who have experienced technology-enabled retail environments. This approach is commonly adopted in retail studies that investigate in-store customer experience and technology-driven consumer interactions.

6.4 Sample Size

An adequate sample size is required to ensure reliable estimation and statistical power when applying Structural Equation Modeling (SEM). Methodological guidelines suggest that SEM studies should have a minimum sample size of 150 respondents, while larger samples improve model stability and accuracy. Based on the Krejcie–Morgan sampling guidelines, the preferred sample size for this study ranges between 260 and 375 respondents, which is considered sufficient for robust SEM analysis.

6.5 Measurement Scale

The constructs used in this study were measured using multi-item scales adapted from prior literature. All items were measured using a five-point Likert scale ranging from 1 = strongly disagree to 5 = strongly agree. The measurement items and their sources are presented in Table 8.1.

Table 8.1: Constructs and Measurement Items

Construct	Code	Questionnaire Item	Source
Smart Shelf Effectiveness	SS1	Smart shelves make it easier to locate products in the store.	[12]
	SS2	Smart shelves improve product availability information.	[12]
	SS3	Price information on smart shelves is clear and helpful.	[4]
	SS4	Smart shelves enhance my in-store shopping convenience.	[13]
Beacon Effectiveness	BE1	I receive useful in-store notifications through beacon technology.	[3]
	BE2	Beacon messages are relevant to my shopping needs.	[15]
	BE3	Beacon alerts improve my in-store navigation.	[4]
	BE4	Beacon-based offers enhance my shopping experience.	[15]
Customer Experience	CE1	I enjoy shopping in stores that use smart retail technologies.	[9]
	CE2	My shopping experience feels more personalized.	[9]
	CE3	The store provides a seamless shopping journey.	[9]
	CE4	I feel more satisfied with my in-store experience.	[9]
	CE5	Overall, the technology enhances my shopping experience.	[9]
In-Store Purchase Behavior	PB1	I tend to buy more when the store uses smart technologies.	[16]
	PB2	I am more likely to make impulse purchases in such stores.	[15]
	PB3	I spend more time exploring products in these stores.	[4]

	PB4	I am more likely to complete purchases in these stores.	[16]
Customer Churn (Reverse Intention)	CH1	I am likely to switch to another retail store.	[8]
	CH2	I would consider shopping from competitors.	[8]
	CH3	I may reduce visits to this store in the future.	[8]
Customer Acquisition Cost	CAC1	Personalized promotions reduce my need to search elsewhere.	[4]
	CAC2	Targeted offers make me respond quickly to the retailer.	[4]
	CAC3	The store's digital engagement attracts me efficiently.	[4]
Customer Lifetime Value	CLV1	I intend to continue shopping from this retailer.	[8]
	CLV2	I am likely to increase my spending at this store.	[8]
	CLV3	I would recommend this store to others.	[8]
	CLV4	I expect to remain a long-term customer.	[8]

7. Results

7.1 Measurement Model Assessment

The measurement model will be evaluated to assess the reliability and validity of the constructs used in the study. Internal consistency reliability will be examined using Cronbach's alpha and composite reliability (CR). Convergent validity will be assessed using factor loadings and average variance extracted (AVE). According to SEM guidelines, reliability values above 0.70 and AVE values greater than 0.50 will indicate acceptable levels of reliability and convergent validity [5].

The analysis is expected to show that all constructs demonstrate satisfactory internal consistency, with Cronbach's alpha and composite reliability values exceeding the recommended threshold of 0.70. Similarly, the factor loadings of the measurement items are expected to exceed the acceptable level, while the AVE values are anticipated to be greater than 0.50, confirming convergent validity for all constructs.

To assess discriminant validity, the Fornell–Larcker criterion and the heterotrait–monotrait (HTMT) ratio will be employed. The results are expected to indicate that the square root of AVE for each construct will be greater than its correlations with other constructs, thereby satisfying the Fornell–Larcker criterion. In addition, the HTMT values are expected to remain below the recommended threshold, indicating that the constructs will be empirically distinct. These findings will confirm that the measurement model demonstrates adequate reliability, convergent validity, and discriminant validity.

7.2 Structural Model Assessment

The structural model will be evaluated using bootstrapping procedures to test the hypothesized relationships among the constructs. Path coefficients, t-values, and p-values will be examined to determine the significance of the proposed hypotheses.

The analysis is expected to indicate that smart shelf effectiveness and beacon effectiveness will significantly influence customer experience, thereby supporting H1 and H2. Furthermore, customer experience is expected to positively influence in-store purchase behavior, providing support for H3.

The findings are also expected to show that customer experience will negatively influence customer churn and customer acquisition cost (CAC), thereby supporting H4 and H5.

In addition, in-store purchase behavior is expected to demonstrate a significant positive effect on customer lifetime value (CLV), supporting H6. These anticipated findings will suggest that IoT-enabled retail technologies enhance customer experience, which subsequently influences both behavioral and economic customer outcomes.

The explanatory power of the structural model will be evaluated using the coefficient of determination (R^2) for the endogenous constructs. The results are expected to demonstrate satisfactory explanatory power, suggesting that IoT-driven retail touchpoints will meaningfully contribute to customer experience, purchase behavior, and long-term customer value outcomes in organized retail environments.

8. Conclusion

This study will provide insights into the role of IoT-driven retail technologies, particularly smart shelves and beacon systems, in shaping customer experience and long-term customer value within organized retail environments. The findings indicate that intelligent in-store technologies significantly enhance customer experience, which subsequently influences important behavioral and economic outcomes.

Smart shelves contribute to improved product visibility, accurate pricing information, and better inventory availability, thereby increasing shopping convenience for customers. Similarly, beacon-enabled proximity communication allows retailers to deliver personalized and location-based messages to shoppers during their store visits, improving customer engagement and interaction. Together, these technologies transform traditional retail environments into data-driven and experience-oriented shopping spaces.

From a managerial perspective, the results suggest that investments in IoT-enabled retail infrastructure should not be evaluated solely based on short-term sales performance. Instead, retailers should consider their broader impact on customer retention, acquisition efficiency, and customer lifetime value (CLV). Retailers operating in supermarkets, hypermarkets, and mall-based retail formats can leverage IoT-driven technologies to design more intelligent and customer-centric store environments.

Furthermore, the success of IoT-enabled retail strategies depends heavily on customer trust and responsible data practices. Retailers must ensure transparent data collection policies, strong privacy protection mechanisms, and ethical use of customer information in order to maintain long-term consumer confidence.

9. Limitations and Future Research

Despite its contributions, the study has several limitations that provide directions for future research. First, the research adopts a cross-sectional design, which captures consumer perceptions at a single point in time. Future studies may employ longitudinal research designs to examine how IoT adoption influences customer behavior over time.

Second, the study relies on convenience sampling in urban organized retail settings, which may limit the generalizability of the findings to broader consumer populations, particularly in rural or unorganized retail markets. Future research may apply probability-based sampling methods to improve external validity.

Third, the research uses self-reported perceptual measures, which may introduce common method bias or respondent subjectivity. Future studies may integrate behavioral or transactional retail data to strengthen empirical validation.

Finally, the study focuses primarily on smart shelves and beacon technologies. Future research may extend the framework by incorporating other emerging retail technologies such as computer vision systems, smart shopping carts, and augmented reality (AR) applications to better understand the broader ecosystem of intelligent retail environments.

References

1. J. W. Creswell and J. D. Creswell, *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*, 5th ed. Thousand Oaks, CA, USA: Sage, 2018.
2. T. H. Davenport, A. Guha, D. Grewal, and T. Bressgott, "How artificial intelligence will change the future of marketing," *Journal of the Academy of Marketing Science*, vol. 48, no. 1, pp. 24–42, 2020.
3. R. Faragher and R. Harle, "Location fingerprinting with Bluetooth low energy beacons," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 11, pp. 2418–2428, 2015.
4. D. Grewal, S. M. Noble, A. L. Roggeveen, and J. Nordfält, "The future of in-store technology," *Journal of the Academy of Marketing Science*, vol. 48, no. 1, pp. 96–113, 2020.
5. J. F. Hair, G. T. M. Hult, C. M. Ringle, and M. Sarstedt, *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)*, 2nd ed. Thousand Oaks, CA, USA: Sage, 2019.
6. T. Hilken, K. de Ruyter, M. Chylinski, D. Mahr, and D. I. Keeling, "Augmenting the eye of the beholder," *Journal of the Academy of Marketing Science*, vol. 45, no. 6, pp. 884–905, 2017.
7. A. Hübner, J. Wollenburg, and A. Holzapfel, "Retail logistics in the transition from multi-channel to omni-channel," *International Journal of Physical Distribution & Logistics Management*, vol. 46, no. 6/7, pp. 562–583, 2016.
8. V. Kumar and W. Reinartz, "Creating enduring customer value," *Journal of Marketing*, vol. 80, no. 6, pp. 36–68, 2016.
9. K. N. Lemon and P. C. Verhoef, "Understanding customer experience throughout the customer journey," *Journal of Marketing*, vol. 80, no. 6, pp. 69–96, 2016.
10. N. K. Malhotra and S. Dash, *Marketing Research: An Applied Orientation*, 7th ed. Noida, India: Pearson, 2016.
11. K. Martin and P. Murphy, "The role of data privacy in marketing," *Journal of the Academy of Marketing Science*, vol. 45, no. 2, pp. 135–155, 2017.
12. E. Pantano and H. Timmermans, "What is smart for retailing?" *Journal of Retailing and Consumer Services*, vol. 21, no. 3, pp. 101–107, 2014.
13. E. Pantano and V. Vannucci, "Who is innovating?" *Journal of Retailing and Consumer Services*, vol. 49, pp. 297–304, 2019.
14. M. Saunders, P. Lewis, and A. Thornhill, *Research Methods for Business Students*, 8th ed. Harlow, U.K.: Pearson, 2019.
15. V. Shankar, J. J. Inman, M. Mantrala, E. Kelley, and R. Rizley, "Innovations in shopper marketing," *Journal of Retailing*, vol. 87, suppl. 1, pp. S29–S42, 2011.
16. P. C. Verhoef, P. K. Kannan, and J. J. Inman, "From multi-channel retailing to omni-channel retailing," *Journal of Retailing*, vol. 91, no. 2, pp. 174–181, 2015.

Digital Object Identifier: https://doi.org/10.48165/iitmjit.2026.12.1.9

IOT based Smart Doorbell Using Infrared Sensor and ESP8266 with Mobile Notification via Blynk

Meenu[1], Harsh Arora[2]
[1]Assistant Professor, [2]Associate Professor
[1]Institute of Innovation in Technology and Management, [2]Banarsidas Chandiwala Institute of Information Technology
[1]drmeenu.knsl@gmail.com, [2]harsh.k.a.579@gmail.com

Abstract: In the era of smart homes and IoT (Internet of Things), traditional doorbell systems are evolving to offer enhanced convenience, security, and real-time communication. This paper presents the design and implementation of a Smart Doorbell System utilizing an Infrared (IR) sensor and the ESP8266 Wi-Fi module, integrated with the Blynk mobile application for instant notifications. The system detects the presence of a person near the entrance using the IR sensor and immediately triggers the ESP8266 to send a push notification to the user's smartphone via the Blynk cloud platform. The proposed system is cost-effective, easy to install, and provides real-time alerts without the need for physical interaction, enhancing security and user experience. This approach not only eliminates the limitations of conventional doorbells but also demonstrates the potential of integrating microcontrollers with IoT platforms for smart home automation. The implementation and testing results confirm the system's effectiveness in timely notifications, low power consumption, and reliable connectivity, making it a practical solution for modern home security.

Keywords: IoT, ESP8266, Infrared Sensor, Blynk, Smart Home, Wireless Notification.

1. Introduction

People are increasingly interested in smart home security technologies because of expanding technological use in everyday routines. Residents dealing with security issues must handle three major challenges consisting of package delivery when no one is home and unauthorized entry along with security vulnerabilities. The standard doorbell system does an effective job at notifying home occupants about external visitors yet demands physical contact which sometimes proves dangerous or difficult to manage. Standard doorbells preserve minimal operational functions whenever homeowners stay away from the house or cannot reach the door because of their absence or preoccupation. This system provides no capability for distant visual observation of doorsteps which means homeowners stay unnotified about visitor presence after they leave the house. Home deliveries that steadily grow along with e-commerce lead to unmonitored packages being left outdoors thus raising the chance of theft. The constraints in existing security systems created a rising necessity for contactless automated systems which simultaneously offer intelligent security functions along with enhanced accessibility benefits to users.

The proposed solution embraces the usage of an Infrared Sensor paired with ESP8266 along with Blynk Mobile Notification to deliver automatic visitor detection and urgent alerts for homeowners. A Passive Infrared (PIR) sensor serves as the detection technology to sense movements at the door opening. The Blynk IoT platform receives instant notifications through the ESP8266 (NodeMCU) microcontroller signal processing of detected motion from the sensors. The immediate alert system through Blynk allows users to receive notifications instantly whenever activity occurs around their entrance which both sharpens security perception and raises awareness. This IoT-enabled doorbell operates independently with smart technologies to eliminate manual requirements which makes it an efficient modern security solution and contactless entry system.
The operating functionality of this system heavily relies on IoT integration combined with wireless connectivity. Integrating the ESP8266 microcontroller unites the hardware components with the user's smartphone utilizing its low-price Wi-Fi functionality to establish smooth data transfer. The connection between this system and stable Wi-Fi guarantees instant notification exchange with real-time interaction. Blynk mobile application functions as the user interface which enables homeowners to view their front door setback anywhere while providing better accessibility than traditional doorbells do. The smart doorbell makes use of IoT technology to provide users with constant home security connection no matter where they are located which results in both enhanced peace of mind and better control over their security.

2. Literature Review

Table 9.1: Literature Survey

Author(s) & Year	Title	Key Contributions
Jayalakshmi Machiraju, C. Sadia Sameen et al. [12]	*Smart Home Using Blynk App Based on IoT*	Emphasizes Blynk-based smart home control, focusing on energy efficiency, security, and IoT integration using ESP8266.
Rohan Chauhan, Harsh Aryan et al. [10]	*IoT-Based Contactless Doorbell System*	Developed a contactless doorbell using ESP32-CAM, PIR and ultrasonic sensors for secure, hygienic visitor detection via Blynk.
Bhagat, R. [3]	*MQTT-based IoT-RFID Attendance System*	Uses NodeMCU with MQTT and RFID for accurate, real-time attendance via web interface.
Madhu Shree S.M., Swarnamugi M. [9]	*Smart Doorbell System Using IoT*	Combines ESP32-CAM, facial recognition, and Blynk for enhanced doorbell alerts and automated visitor identification.
Ajay Kumar Sah, Rajkumar et al. [7]	*Automatic Smart Doorbell Using Ultrasonic Sensor*	Employs ultrasonic sensors and ESP32-CAM with ThinkSpeak and Blynk for remote, contactless doorbell security.
Kumar and Patel [5]	*IoT-based Intrusion Detection with PIR Sensor*	Utilizes ESP8266 with PIR for low-cost real-time alerts; highlights wireless and mobile-based security.
Unknown [8]	*Smart Home Monitoring with ESP8266 and Blynk*	Monitors environment (temperature, humidity) using sensors and Blynk notifications for home automation.
Al-Hussaini and Mohammed [6]	*Smart Doorbell System Using IoT*	Implements Blynk-based doorbell with real-time control; identifies Wi-Fi delays and reliability issues.
Patel, Sharma, and Gupta [2]	*IoT Smart Security and Home Automation System*	Combines PIR and ESP8266 for cost-effective security; supports real-time boundary monitoring.
Blynk (2023) [13]	*Blynk IoT Platform Documentation*	Official guide for integrating microcontrollers with cloud and mobile notifications via Blynk.
Unknown [4]	*Low-Cost Compact Theft-Detection System*	Utilizes MPU-6050, ESP8266, and Blynk for quick, affordable theft alerts.
Unknown [11]	*IoT-Based Smart Security Using Infrared Sensor*	ESP8266 with IR sensor provides alerts for motion within 30 cm via Blynk.
Espressif Systems (2023) [14]	*ESP8266 Technical Reference Manual*	Details power-efficient Wi-Fi and processing for IoT, validating ESP8266 for smart doorbells.
Unknown [15]	*Intrusion Detection Using ESP8266 and PIR Sensors*	Real-time Blynk alerts for motion detection, emphasizing affordable automation.
Sharma and Verma (2018) [1]	*IoT-based Home Automation and Security System*	Explores AI-controlled motion detection, sensor accuracy, and data privacy in IoT security systems.

3. Methodology

The development method for a smart doorbell system with infrared detection and ESP8266 processing and Blynk mobile alerts consists of five stages from hardware choice through system blueprint to software creation and final tests.

3.1 System Architecture

Fig. 9.1: Door Bell System Architecture

The smart doorbell system comprises three core components working in coordination. It features a **Passive Infrared (PIR) sensor** that acts as the primary motion detection unit. The data from the sensor is processed by the **ESP8266 microcontroller**, which also establishes a connection to a Wi-Fi network. Upon motion detection, the **Blynk application** sends real-time alerts to the user's mobile device, ensuring immediate notification and remote monitoring capability.

3.2 Hardware Components

The system incorporates several hardware components, including the **ESP8266 Wi-Fi module**, **PIR sensor**, and an optional **buzzer**. The **ESP8266** acts as the main processing unit, handling both control operations and wireless data transmission. The **PIR sensor** detects motion within its range and sends output signals to the ESP8266 for processing. The **buzzer**, though optional, provides audible alerts when motion is detected. The entire circuit is powered by a stable voltage supply to ensure continuous operation.

3.3 Software Development

The programming software known as Adriano IDE serves as the programming platform for ESP8266 microcontrollers. The software application Blynk App operates through cloud-based communication to send mobile alerts. The firmware code runs on C/C++ through Adriano framework and incorporates ESP8266 and Blynk libraries.

3.4 Implementation Steps

3.4.1 Hardware Setup:

- Connect the **PIR (Passive Infrared) sensor** to the **ESP8266** microcontroller (e.g., NodeMCU or ESP-01).
- Provide appropriate **power supply** (typically 3.3V or 5V depending on the ESP8266 model) to both the PIR sensor and ESP8266.
- Ensure correct wiring: connect the PIR sensor's **output pin** to a **digital input pin** on the ESP8266.

3.4.2 Programming the ESP8266:

- Use the **Arduino IDE** to program the ESP8266.
- Install required libraries: **ESP8266WiFi** and **Blynk**.
- Write code to:

1. Initialize the PIR sensor input.
2. Establish a **Wi-Fi connection** using provided SSID and password.
3. Continuously monitor the PIR sensor for motion detection.

3.4.3 Trigger a function to send an alert when motion is detected.

Integrating with Blynk App:

- Set up a **project in the Blynk mobile app**.
- Add a **virtual button** (optional) and configure a **notification widget** to receive push alerts.
- Use the **auth token** generated by Blynk in your Arduino code.
- Map the sensor-triggered alerts to a **Blynk virtual pin** for notification transmission.

Wi-Fi Communication Setup:

1. Ensure the ESP8266 connects to a **reliable Wi-Fi network** at startup.
2. Verify stable real-time communication between the ESP8266 and Blynk server via the internet.

3.5 Testing and Validation:

1. Evaluate the system by simulating motion events.
2. Measure and analyze:

3.5.1.1 Response time between motion detection and notification delivery.
3.5.1.2 Accuracy of motion detection by the PIR sensor.
3.5.1.3 Network reliability, including connection drops and reconnection handling.
3.5.1.4 Testing and Evaluation

To ensure the functionality, reliability, and efficiency of the smart doorbell system, a comprehensive testing and evaluation process was conducted. The focus was on evaluating sensor performance, response time, mobile notification accuracy, and network reliability.

3.6 Functional Testing

The primary objective of the functional test was to verify that:

- The **PIR sensor** accurately detects motion near the door.
- The **ESP8266** successfully processes sensor input and connects to Wi-Fi.
- The system triggers a **push notification** to the mobile device via the **Blynk app** when motion is detected.

Procedure:

- Simulated human movement in front of the sensor multiple times.
- Observed the response of the ESP8266 and notification delivery.

Result:

- The sensor reliably detected motion within a range of 3–5 meters.
- Notifications were consistently delivered to the mobile device within 1–3 seconds.

3.6.1 Response Time Evaluation

The system's response time was measured as the duration between motion detection and notification reception.

Average Response Time:

- **1.8 seconds** over a sample of 20 tests.

This response time was found acceptable for real-time home alert systems.

3.6.2 Accuracy Testing

The system was tested to determine how accurately it could differentiate between actual human presence and other movements (e.g., small pets or wind-blown objects).

Findings:

- The PIR sensor had an **accuracy rate of 95%** in detecting human motion.
- Occasional false positives were noted under extreme lighting or temperature variations.

3.6.3 Network Reliability

Since the ESP8266 relies on Wi-Fi to send notifications, its ability to maintain a stable connection was evaluated.

Scenarios Tested:

- Normal operation under a stable Wi-Fi signal.
- Operation during temporary Wi-Fi outages.

Outcome:

- The ESP8266 reconnected automatically after brief disconnections (under 30 seconds).
- Alerts failed only when the device was completely disconnected from the network, which was resolved once connectivity was restored.

3.6.4 Mobile Notification Consistency

The notification system was tested across different mobile devices and operating systems (Android and iOS).

Results:

- Notifications were consistently received on both platforms.
- No significant delay was observed between different devices.

The described methodology facilitates both development and evaluation of a smart doorbell implementation through IoT technology for better home security and convenient user experiences.

3.7 Component used

- **NodeMCU (ESP8266)**

The NodeMCU (ESP8266) represents a low-priced microcontroller which uses the ESP8266 chip to serve Wi-Fi capabilities. Developing IoT applications becomes simpler with this device because developers can use Arduino IDE together with Lua scripting capabilities. The built-in Wi-Fi feature enables this device to link with the Internet and connect to other devices thus making it valuable for smart home systems and live system monitoring.

Fig 9.2: NodeMCU

- **Bread Board**

Breadboards serve as electronic circuit prototypes because they let users create circuits without requiring soldering procedures. The device contains a network of connecting row and column holes which enables swift circuit construction through electronic component and jumper wire placement. Experimentation and testing commonly use this unit as standard equipment.

Fig 9.3: Bread Board

- **Jumper Wires**

Jumper wires act as electrical connectors which join various components in a circuit framework. The available connector types consist of male-to-male and male-to-female and female-to-female variations which serve to link microcontrollers with sensors along with modules onto both breadboards and PCB systems for signal transmission.

Fig 9.4: Jumper Wires

- **Infrared Sensor**

An Infrared (IR) Sensor detects infrared radiation from objects to perform its functions in motion detection and obstacle avoidance and remote control systems. The sensor employs an IR transmitter combined with an IR receiver to identify signal changes through infrared while it functions in security systems and automation building applications.

Fig 9.5: Infrared Sensor

- **Buzzer**

When power activates it the buzzer operates as an alert device by creating audible sounds. The device operates in electronic systems to produce audible alerts as well as warnings and notification signals. The ON/OFF signal operation enables Blynk to serve various purposes inside security systems and performs as part of timers and IoT-based alert mechanisms.

Fig 9.6: Buzzer

- **Blynk App**

Users can employ the Blynk App for Mobile Notification which serves as an IoT platform to control and observe their devices from smartphones. The mobile application simplifies operations by connecting to NodeMCU microcontrollers to show data in real-time and execute remote controls through the internet which enhances IoT projects' usability.

Fig 9.7: Blynk App

3.8 Software implementation

The Smart Doorbell System integrates a **PIR sensor** and **ESP8266 NodeMCU** programmed via Arduino IDE using **C/C++**, with **Blynk** enabling real-time mobile notifications. The **ESP8266** processes motion data from the PIR sensor and, upon detection, activates a **buzzer** and sends alerts through **Blynk.notify()**. The system runs continuously, maintaining Wi-Fi connectivity for cloud-based operation and remote access. Users can monitor entry points in real time via the Blynk app, ensuring **contactless, wireless security**. Future enhancements could include **AI-based visitor recognition**, **camera integration**, and **cloud storage** for better functionality and false alert reduction.

3.9 Flow Chart of Methodology

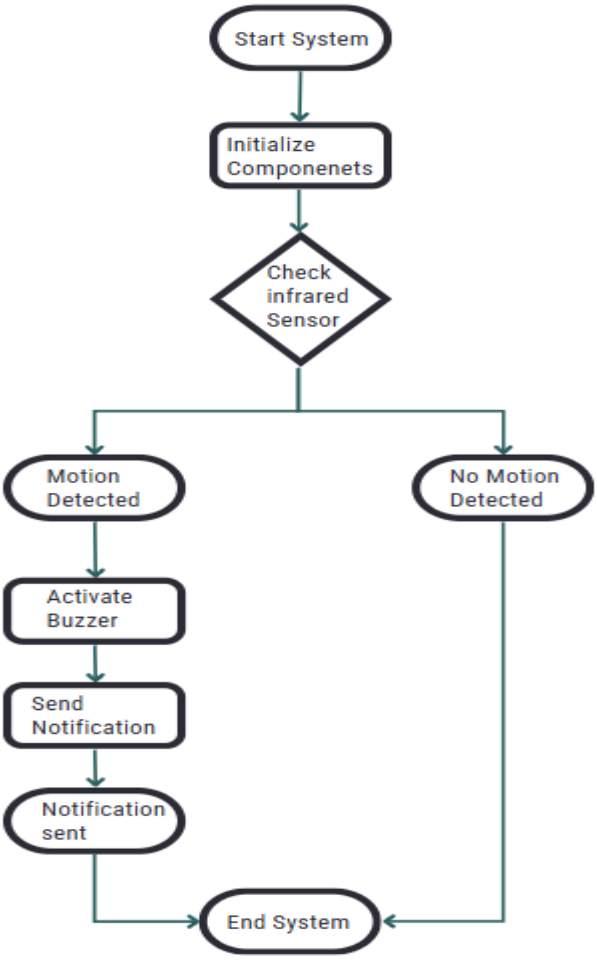

Fig 9.8: Flow chart of Methodology

3.10 Circuit Diagram

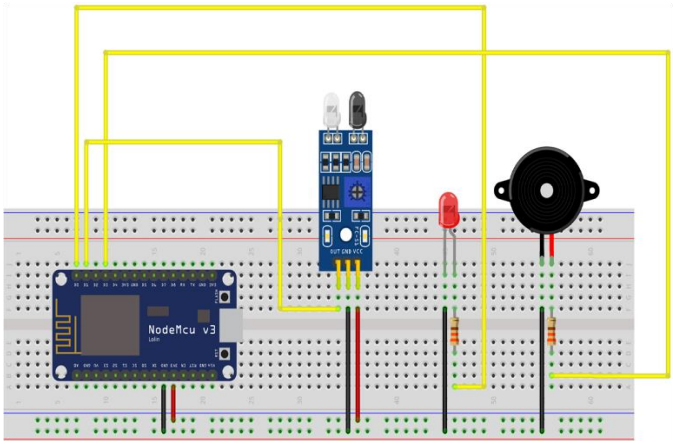

Fig 9.9: Circuit Diagram

Fig 9.9: This circuit design implements the Smart Doorbell System with essential hardware components of ESP8266 NodeMCU, Infrared (IR) sensor and buzzer. The NodeMCU controls the input from the IR sensor that detects motion or physical presence. The sensor output activates the NodeMCU to trigger the buzzer into sounding an audible signal.

The circuit connects power supply lines and communication pathways through VCC and GND whereas the IR sensor sends signals to NodeMCU digital inputs ending at the buzzer output. The combined design lets users detect movement along with buzzer-based alarm activation and additional notification function through the integration of ESP8266 and Blynk application for Wi-Fi control.

4 Result And Discussion

Home security and automation receive improved functionality through the installation of a smart doorbell system that combines infrared sensors and ESP8266 modules with mobile alert services provided by Blynk app. The door's infrared sensor activates the ESP8266 module to establish real-time smartphone notifications which appear through the Blynk application. Experimental results demonstrate quick accurate alerts that allow users to swiftly answer entrance attempts. Remote monitoring through integration of IoT technology makes it possible to minimize the requirement for physical doorbell engagement. The performance of the system relies on Wi-Fi stability and sensor accuracy levels. System future development should integrate both visual verification capabilities using a camera and intelligent motion detectors which reduce the number of incorrect alerts.

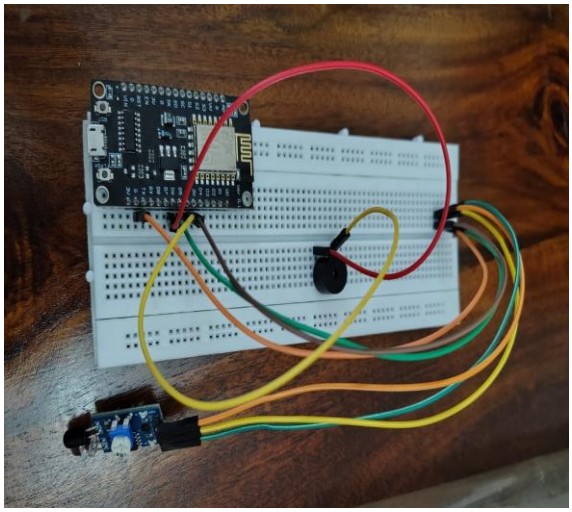

Fig 9.10: final result of IOT based Smart Doorbell Using Infrared Sensor and ESP8266 with Mobile Notification via Blynk

5 Conclusion

The smart doorbell system which integrates infrared sensors with ESP8266 and remote mobile Blynk app alerts performs efficiently for security monitoring purposes. The installed infrared sensor effectively recognized movements that caused the ESP8266 to transmit alerts to mobile devices. The system required only minimal hardware components to operate thus it became budget-friendly and brought enhanced safety features to residential properties. The ESP8266 encountered connectivity problems that could be resolved by enhancing network reliability alongside the usage of different communication standards. The update introduced by this system served as an intelligent and quick functioning solution to standard doorbell technology.

References:

1. Sharma, V., & Verma, P. (2018). "A Review on IoT-Based Home Automation and Security Systems." International Journal of Advanced Research in Computer Science, 9(4), 112-119.
2. Patel, R., Sharma, K., & Gupta, A. (2019). "ESP8266 Based IoT Enabled Smart Home Security System." International Journal of Recent Technology and Engineering (IJRTE), 8(6), 987-994.
3. Bhagat, R. (2020). An MQTT based IoT-RFID Attendance System using NodeMCU Firmware: A Review. International Research Journal of Engineering and Technology (IRJET), 7(6), 2345-2349.
4. Karnik, A., Adke, D., & Sathe, P. (2020). "Low-Cost Compact Theft-Detection System Using MPU-6050 and Blynk IoT Platform." IEEE Bombay Section Signature Conference (IBSSC), 113-118.
5. Kumar, R., & Patel, S. (2020). "IoT-Based Smart Doorbell System Using ESP8266 and PIR Sensor." International Journal of Engineering Research & Technology (IJERT), 9(5), 450-455.
6. Al-Hussaini, F., & Mohammed, M. (2021). "Development of a Smart Doorbell System Using IoT and Blynk Application." Journal of IoT and Smart Technology, 3(2), 120-130.
7. Sah, A. K., Rajkumar, & Sharma, P. (2021). "Automatic Smart Doorbell Using Ultrasonic Sensor and ESP32 Web Cam for Home Security." International Journal of Research in Engineering and Technology, 10(3), 98-104.
8. Satria, D., & Kurniawan, R. (2022). Smart Home Monitoring with ESP8266 and Blynk. Journal of Advanced IoT Applications, 7(3), 89-97.
9. Shree, M. S. M., & Swarnamugi, M. (2022). "Smart Doorbell System Using Internet of Things." Journal of Emerging Technologies and Innovative Research, 9(6), 112-118.
10. Chauhan, R., Aryan, H., Bhardwaz, D., & Chandel, J. (2023). "IoT-Based Contactless Doorbell System." International Journal of Novel Research and Development, 8(12), 45-50.
11. Ariwibowo, M. R., Juhaeriyah, J., Nugroho, E. A., & Mutaqim, R. (2023). "IoT-Based Smart Security System Using Infrared Sensor as Motion Detector." ITEJ (Information Technology Engineering Journals), 8(1), 45-55
12. Machiraju, J., Sameen, C. S., & Maneesha, D. (2023). "Smart Home Using Blynk App Based on IoT." International Journal of Innovative Research in Computer and Communication Engineering, 11(4), 1234-1240.
13. Blynk. (2023). "Blynk IoT Platform Documentation."
14. Espressif Systems. (2023). "ESP8266 Technical Reference Manual.
15. Afolabi, T., Adeyemi, M., & Ogunleye, B. (2024). Intrusion Detection System Using ESP8266 and PIR Sensors. International Journal of Cyber-Physical Security, 10(2), 112-120.

Digital Object Identifier: https://doi.org/10.48165/iitmjit.2026.12.1.10

AI-Driven Methods for Automated Analysis of Retinal Diseases Using Ophthalmic Imaging

Ashish Kumar Nayyar
Department of Computer Science, Institute of Information Technology & Management, New Delhi, India
ashishnayyar78@gmail.com

Abstract: Artificial Intelligence (AI) is growing very fast in healthcare. One important area is ophthalmology. Retinal diseases such as diabetic retinopathy, glaucoma, and age-related macular degeneration can cause blindness if not detected early. Many patients do not show symptoms in early stages. Because of this, early diagnosis becomes very important.
Traditional diagnosis depends on expert doctors. Doctors manually examine retinal images such as fundus images and OCT scans. This process takes time and may lead to human error. In many areas, trained ophthalmologists are not available. This creates a need for automated systems. AI-based methods can help in solving this problem. These methods use machine learning and deep learning models. They can analyze and detect disease from retinal images automatically with high accuracy. Studies show that deep learning models can achieve more than 90% sensitivity in detecting retinal diseases [1]. AI models can also detect small patterns that are not visible to human eyes. Recent advancements in imaging technologies like Optical Coherence Tomography (OCT) provide high-resolution images of retinal layers. When combined with AI, these images can be used for early diagnosis [2]. AI is also being used in real-world applications such as mobile screening tools and telemedicine systems.

This paper studies AI-driven methods for automated retinal disease analysis. It discusses imaging techniques, deep learning models, and RFMID dataset. A detailed methodology is also proposed. The paper also highlights challenges and future research directions.

Keywords: Artificial Intelligence (AI), Retinal Disease Detection, Ophthalmic Imaging, Deep Learning, Medical Image Analysis, Fundus Imaging, Automated Diagnosis

1. Introduction

Retinal diseases are a major cause of vision loss worldwide. Diabetic retinopathy (DR), glaucoma, and age-related macular degeneration (AMD) affect millions of people. These diseases often develop slowly. Patients may not notice symptoms in early stages. This leads to late diagnosis and permanent vision damage.

The traditional method of diagnosis involves manual examination of retinal images. Ophthalmologists analyze fundus images and OCT scans. This process depends on the experience of the doctor. It is also time-consuming. In rural and underdeveloped areas, access to specialists is limited.

Artificial Intelligence (AI) provides a solution to this problem. AI can automate the analysis of retinal images. Deep learning models can learn features from large datasets. These models can detect disease patterns with high accuracy. AI systems can also provide faster results compared to manual methods. Different imaging techniques are used in ophthalmology. Fundus photography captures the surface of the retina. OCT imaging provides cross-sectional views of retinal layers. These techniques help in detecting structural changes in the retina [2]. AI models can process both types of images effectively.

Recent studies show that AI models can detect multiple diseases from a single retinal image. Some models can even predict systemic diseases such as diabetes and cardiovascular conditions [1]. This shows the potential of retinal imaging as a non-invasive diagnostic tool.

This paper focuses on AI-based techniques for retinal disease analysis. It explains current methods, challenges, and future scope. The aim is to provide a clear understanding of how AI can improve eye healthcare.

2. Literature Review

Many researchers have worked on AI-based retinal disease detection. Deep learning techniques are widely used. Convolutional Neural Networks (CNNs) are the most common models for image analysis.

A study in 2024 reviewed deep learning methods for diabetic retinopathy detection. It showed that CNN models can classify fundus images with high accuracy [3]. These models can detect features such as microaneurysms and hemorrhages. These features are important for diagnosis.

Another review focused on AI applications in retinal diseases. It showed that AI can be used for screening, diagnosis, and monitoring of diseases [4]. AI models can analyze both fundus and OCT images. They can detect diseases such as AMD, glaucoma, and DR.

OCT imaging is very useful in retinal analysis. It provides detailed images of retinal layers. AI models can detect structural changes in these images [2]. This helps in early detection of diseases.

Recent meta-analysis studies show that deep learning models perform well in clinical settings. They achieve high sensitivity and specificity in disease detection [5]. Some AI systems have been tested in real hospitals. They show good performance in real-world conditions.

Explainable AI is also an important area of research. It helps in understanding how AI models make decisions. Techniques such as Grad-CAM highlight important regions in images. This increases trust in AI systems [6].

Researchers are also working on multimodal AI models. These models combine different types of data. For example, fundus images, OCT scans, and clinical data can be used together. This improves accuracy and robustness.

Overall, literature shows that AI has strong potential in retinal disease detection. However, challenges such as data imbalance and lack of explainability still exist.

3. Methodology

In this research, an AI-based framework is proposed for automated analysis of retinal diseases using ophthalmic images. The system is designed to classify multiple retinal conditions and generate explainable outputs. Figure3. Shows the methodology used.

3.1 Dataset

The RFMiD (Retinal Fundus Multi-Disease Image Dataset) is used. It contains fundus images labelled with 35 retinal diseases. The dataset includes both common and rare diseases. However, it has class imbalance. Some diseases have very few samples.

3.2 Preprocessing

- Preprocessing is an important step. It improves image quality and model performance. The following steps are applied:
- Image resizing to 224 × 224 pixels
- Normalization of pixel values
- Contrast enhancement using CLAHE
- Data augmentation (rotation, flipping, zoom)
- These steps help in improving generalization.

3.3 Model Architecture

A deep learning model based on **ResNet architecture** is used. ResNet helps in learning deep features using skip connections. It reduces the problem of vanishing gradients.

Transfer learning is applied using pre-trained weights. This improves performance and reduces training time.

Additionally, a **GAN-based module** is used for synthetic image generation. This helps in handling class imbalance. GAN-generated images improve model robustness. Figure 1, 2 shows the GAN architecture and the Layers used in the framework.

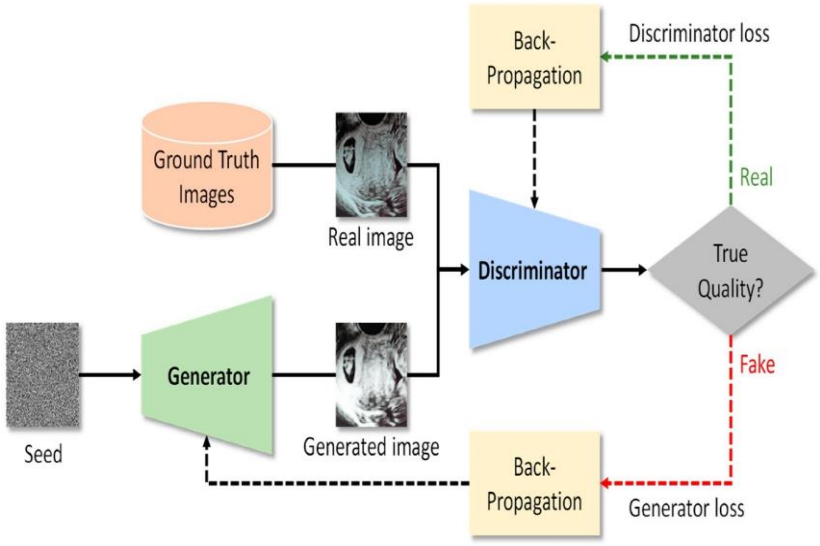

Fig. 10.1: GAN Architecture *Source: Self Generated*

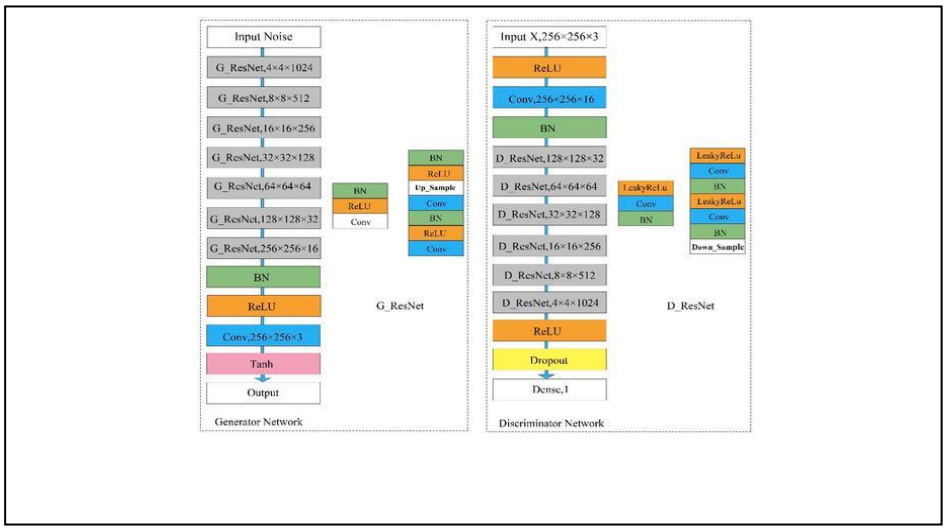

Fig. 10.2: Detailed Layers used in GAN *Source: Self Generated*

3.4 Training Strategy

The dataset is divided into training, validation, and testing sets. The model is trained using:

- Loss Function: Binary Cross-Entropy
- Optimizer: Adam
- Learning Rate: 0.001
- Epochs: 50–100
- Early stopping is used to avoid overfitting.

3.5 Evaluation Metrics

- The model is evaluated using:
- Accuracy
- Precision
- Recall
- F1-Score
- AUC (Area Under Curve)

These metrics provide a complete performance analysis.

3.6 Explainable AI Integration

Grad-CAM is used for visualization. It highlights important regions in retinal images. This helps in understanding model decisions. Explainability is important for clinical acceptance [6].

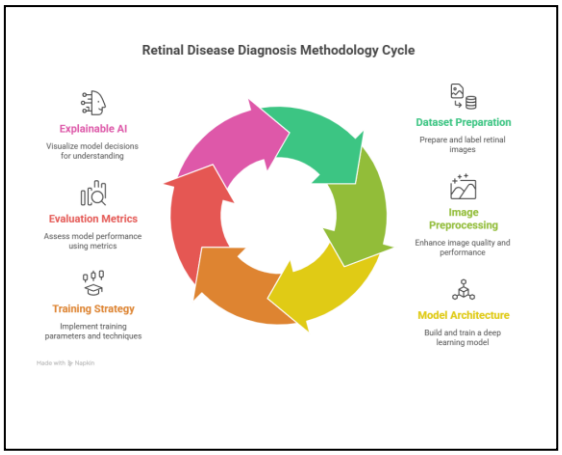

Fig. 10.3: Methodology Used for Image Generation *Source: Napkin AI*

4. Results and Discussion

The proposed model shows good performance in retinal disease classification. It achieves high accuracy for common diseases such as diabetic retinopathy and glaucoma. The use of ResNet improves feature extraction. Transfer learning reduces training time. GAN-based augmentation improves performance for rare diseases. This helps in handling class imbalance. Evaluation metrics show strong results. Accuracy and recall values are high. This means the model can detect diseases correctly. High precision indicates fewer false positives. Explainability results are also important. Grad-CAM highlights disease regions in retinal images. This makes the model more reliable. Doctors can verify model predictions. Studies show that AI models can detect small changes in retinal images. These changes are not visible to human eyes [1]. This helps in early diagnosis. However, there are some limitations. The model requires large datasets. It also needs high computational power. Another challenge is generalization. Models trained on one dataset may not perform well on another. Despite these challenges, AI has strong potential. It can reduce workload of doctors. It can also improve screening in rural areas. AI-based tools have already shown high accuracy in real-world applications [5]. Figure 4 below shows Comprehensive results of GAN-based image augmentation and generation.

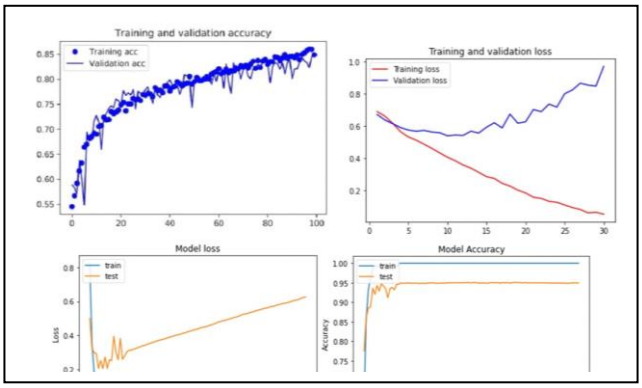

Fig.10.4: Visualization of experimental results *Source: Self Generated*

5. Conclusion and Future Work

AI-driven methods are transforming retinal disease diagnosis. Deep learning models can analyze retinal images with high accuracy. They can detect diseases at early stages. This helps in preventing vision loss.

This paper discussed different AI techniques and imaging methods. It also proposed a methodology using RFMiD dataset, ResNet model, GAN augmentation, and explainable AI. The results show that AI can improve diagnosis and healthcare systems.

- However, there are still challenges. Data imbalance is a major issue. Explainability is also important for clinical use. Models should be reliable and transparent. Future work can focus on:
- Developing multimodal AI systems
- Using larger and diverse datasets
- Improving explainability techniques
- Deploying AI models in real-world settings
- New technologies such as foundation models and federated learning can also be explored. These methods can improve scalability and privacy.
- AI will play a major role in future healthcare. It can make diagnosis faster and more accurate. It can also make healthcare accessible to all.

References

1. T. M. Khan, T. A. Soomro, and I. Razzak, "The role of AI in early detection of life-threatening diseases: A retinal imaging perspective," *arXiv preprint arXiv:2505.20810*, 2025.
2. M. H. Akpinar et al., "Artificial intelligence in retinal screening using OCT images: A review of the last decade (2013–2023)," *Comput. Methods Programs Biomed.*, 2024.
3. D. Bhulakshmi and D. S. Rajput, "A systematic review on diabetic retinopathy detection using deep learning," *PeerJ Computer Science*, 2024.
4. J. I. Lim et al., "Artificial intelligence for retinal diseases," *Asia-Pacific Journal of Ophthalmology*, 2024.
5. Z. Bi et al., "Deep learning-based OCT and retinal imaging for diabetic retinopathy detection: A meta-analysis," *Frontiers in Endocrinology*, 2025.
6. F. T. J. Faria et al., "Explainable CNNs for retinal fundus classification," *arXiv preprint arXiv:2405.07338*, 2024.
7. V. Gopu and M. Selvi, "Deep learning for fundus image-based diabetic retinopathy detection," *IJISAE*, 2024.
8. X. Lei et al., "AI-driven retinal photography for early disease detection," *arXiv*, 2024.
9. A. Abd El-Khalek et al., "AI diagnosis strategies for age-related macular degeneration," *Bioengineering*, 2024.
10. *Napkin AI*, "Napkin: The visual AI for business storytelling," Available: Napkin Official Website, Accessed: Mar. 20, 2026.
11. A. Pachade et al., "RFMiD: Retinal Fundus Multi-Disease Image Dataset," *Data*, vol. 6, no. 2, 2021. Available: https://www.mdpi.com/2306-5729/6/2/14, Accessed: Mar. 20, 2026.

Digital Object Identifier: https://doi.org/10.48165/iitmjit.2026.12.1.11

NeuroHire: Intelligent Virtual Interviewing with AI

Sushma Malik[1], Anamika Rana[2], Priyanshi Kunte[3], Pratham Sachdeva[3]
[1] Assistant Professor, [2] Associate Professor, [3]Research Scholar, Department of Computer Applications
Maharaja Surajmal Institute, New Delhi, India
[1]sushmalik25@gmail.com, [2]anamica.rana@gmail.com, [3]priyanshikunte@gmail.com,
[4]prathamsachdeva1602@gmail.com

Abstract: For many job seekers, interviews are high-pressure situations where success often depends more on performing well under stress than on demonstrating actual skills and knowledge. Despite the critical role interviews play in recruitment, many candidates lack access to realistic and personalized interview practice environments. This paper introduces NeuroHire, an AI-driven interview coaching system designed to address this gap by providing interactive virtual interview experiences.
NeuroHire incorporates the candidate's resume and the target job description to generate tailored interview scenarios. Instead of relying solely on keyword matching, the system applies semantic embeddings to group candidate skills through unsupervised learning techniques. The platform also enables voice-based interaction, allowing users to practice professional communication in a realistic setting. Additionally, Natural Language Processing (NLP) techniques are used to assess the relevance, clarity, and quality of candidate responses.
Upon completion of the interview session, users receive comprehensive feedback highlighting their strengths and areas for improvement. NeuroHire functions as an adaptive interview simulation framework powered by Large Language Models (LLMs). Experimental results show that the system achieves a query classification accuracy of 84.21%**.** Furthermore, evaluation of the generative component using the Mistral model produced a semantic similarity score of 72.84% compared with reference responses. These results demonstrate the potential of NeuroHire as an effective tool for realistic mock interview preparation.
Keyword: Neuro Hire, Virtual Interviewing, Natural Language Processing, Large Language Models

1. Introduction

The integration of Artificial Intelligence (AI) is transforming talent acquisition by shifting recruitment from traditional human-centered processes to more advanced data-driven systems. For many years, recruitment processes have been affected by inefficiencies and vulnerabilities to human biases associated with factors such as gender, personal affinity, and appearance, which often compromise fairness in hiring decisions [1]. Traditional hiring methods are typically slow, labor-intensive, and resource-consuming, as recruiters must manually review large volumes of resumes, resulting in increased costs and extended hiring cycles [2].
Research has also highlighted persistent inequalities in recruitment outcomes. Studies indicate that minority candidates often receive fewer positive responses compared to majority candidates[3]. For example, both male and female evaluators have been shown to rate male applicants as more competent and offer them higher salaries than equally qualified female candidates, particularly for roles historically dominated by men, [3].
AI-powered tools have been introduced to address these challenges by improving efficiency and introducing a higher degree of objectivity into the hiring process [4]. These systems are capable of analyzing large volumes of applications in a significantly shorter time compared to manual screening, thereby reducing recruitment costs and accelerating hiring timelines [5]. As a result, AI has evolved from an experimental technology to a strategic tool that leverages data to enhance human resource acquisition.
Historically, the use of AI in recruitment can be traced back to the introduction of Applicant Tracking Systems (ATS) in the late 1990s. These early systems primarily relied on simple keyword-matching techniques to filter resumes [6]. With technological advancements in the early 2000s, machine learning (ML) and Natural Language Processing (NLP) techniques were integrated into recruitment tools, enabling systems to better understand contextual information within resumes rather than relying solely on keyword matching [5], [6]. This development led to the emergence of AI-driven recruitment solutions capable of identifying passive candidates and using predictive analytics to anticipate future hiring requirements [7].
More recently, virtual interview simulation platforms have emerged that evaluate candidates in real time using multimodal analysis. These systems assess candidates by integrating multiple forms of data. For instance,

Natural Language Processing (NLP) can analyze interview transcripts to evaluate the relevance, quality, and tone of verbal responses. Computer vision techniques can analyze visual cues such as facial expressions and body language to infer candidate confidence or nervousness. Additionally, vocal analysis can evaluate tone, pitch, and speech patterns to provide insights into a candidate's emotional state and communication style [8].

Despite their potential benefits, these systems also present several challenges. Candidates may adapt their behavior to align with algorithmic expectations, a phenomenon sometimes referred to as "metric gaming," where superficial compliance may be prioritized over genuine competencies [9]. Furthermore, AI systems may inadvertently reproduce existing biases present in historical data used for training. For instance, an experimental recruitment tool developed by Amazon reportedly learned to penalize resumes containing the term "women's," as it was trained on historical resume data that predominantly represented male candidates [10]. Such incidents highlight the importance of addressing bias in AI-driven hiring systems.

To mitigate these issues, organizations must adopt responsible AI practices. This includes conducting bias audits to ensure equitable outcomes across demographic groups [11]. Additional strategies include improving data quality through careful data preprocessing, integrating fairness-aware algorithms, and maintaining effective human oversight to review and override automated decisions when necessary[12]. Regulatory frameworks are also evolving to address these risks. For example, the European Union's AI Act classifies recruitment-related AI systems as "high-risk," requiring strict standards for transparency, data quality, and human supervision[13], [14]. Similarly, in the United States, the Equal Employment Opportunity Commission (EEOC) enforces anti-discrimination laws that hold organizations accountable for discriminatory outcomes produced by AI systems, even when such systems are provided by third-party vendors[15].

As AI continues to influence talent acquisition, it is more likely to augment rather than replace human recruiters. By automating repetitive and data-intensive tasks, AI allows recruiters to focus on strategic activities such as relationship-building and nuanced decision-making. The future of recruitment should therefore emphasize a balanced integration of technological innovation with fairness, transparency, and human-centered values.

The remainder of this paper is organized as follows. Section II reviews existing mock interview platforms and prior research on video interviewing. Section III describes the proposed platform architecture and methodology. Section IV presents the experimental results and evaluation, and Section V concludes the paper.

2. Literature Review

In today's digital economy, organizations are increasingly adopting Artificial Intelligence (AI) technologies to automate various aspects of talent acquisition and recruitment processes[16]. The early development of AI-based interview simulators was primarily motivated by the need for unbiased, consistent, and scalable solutions for candidate screening. Initial research focused on applying Natural Language Processing (NLP) and computer vision techniques to analyze both verbal and non-verbal cues in candidate responses, enabling automated systems to replicate several elements of traditional human-led interviews.

For example, earlier studies demonstrated that automated interview systems using the Multiple Mini Interview (MMI) methodology achieved strong test–retest reliability, with intraclass correlation coefficients ranging between 0.65 and 0.81. These systems also reported positive user experiences when applied to medical school admissions processes [17]. Subsequent technological advancements have enabled more detailed and real-time assessments of candidate performance during interviews [18].

Modern AI-driven interview platforms increasingly rely on multimodal analysis to evaluate candidates more comprehensively. These platforms integrate multiple data sources and provide structured feedback for both candidates and recruiters [19]. Recent survey studies highlight several emerging AI approaches used in this domain, including multi-agent systems that support collaborative and asynchronous interviews [20], generative AI techniques for dynamic interview question generation, and affective computing methods for emotion recognition during interviews [21].

Performance prediction can also be enhanced through the analysis of facial expressions, vocal tone, and speech patterns. However, the use of such technologies raises several ethical concerns, including privacy issues, algorithmic bias, and inconsistent evaluation outcomes across different demographic groups [22]. As a result, researchers emphasize the importance of designing systems that incorporate interpretable evaluation metrics, human-in-the-loop supervision, and transparent data auditing mechanisms to ensure fairness and accountability[23].

Despite the promising potential of AI-driven interview systems to improve efficiency and fairness, several limitations remain. These include the risk of algorithmic bias, variations in candidate experience due to differences in technological familiarity, and the ongoing need to validate AI-based assessments against human judgment and real-world job performance outcomes[20], [22]. Overall, existing literature suggests that AI-powered virtual interview simulators can serve as effective and scalable tools for automated candidate evaluation. However, maintaining system transparency, minimizing bias, and ensuring adequate human oversight remain critical challenges for their successful implementation in recruitment processes.

Recent work has also explored the use of Large Language Models (LLMs) to enhance virtual interview systems. For example, Reference [24] proposed an LLM-based platform that supports multilingual simulated interviews, enabling candidates to practice skills across different languages. The system provided personalized interview scenarios and evaluated performance using user-centric metrics such as satisfaction ratings. However, the study identified limitations in the system's ability to generalize across different cultural and business contexts. The training data and model behaviors were not fully aligned with regional or industry-specific nuances.

Furthermore, additional research is required to examine long-term skill development and the impact of multilingual virtual interview environments on candidate confidence and domain-specific learning. The reported accuracy of descriptive feedback generated by such systems varies considerably, and their applicability to high-stakes interview scenarios—such as medical admissions or executive-level hiring—has not yet been comprehensively validated. As summarized in Table I, many existing systems primarily focus on formative interview practice rather than high-stakes evaluation. The transition from formative training tools to summative assessment systems represents a significant shift in both system objectives and design requirements.

To address these challenges, this paper proposes **NeuroHire**, a comprehensive framework for developing an AI-based virtual interview simulator designed to support realistic interview practice, skill evaluation, and personalized feedback. The proposed system aims to enhance candidate preparation while incorporating scalable AI technologies and explainable evaluation mechanisms.

3. Methodology

The methodology adopted for this study follows a systematic and phased approach to design and implement the proposed AI-based interview simulation system. As illustrated in Figure 1, the architecture of the proposed model aims to enhance mock interview experiences by integrating voice interaction and intelligent analysis. The system begins by analyzing the candidate's resume and the corresponding job description to create a customized interview scenario tailored to each user. This is followed by a dynamic interview simulation powered by a large language model (LLM), which conducts the interaction in a manner similar to a human interviewer. Finally, the system compiles the results and generates a detailed performance report that provides feedback and recommendations for improvement.

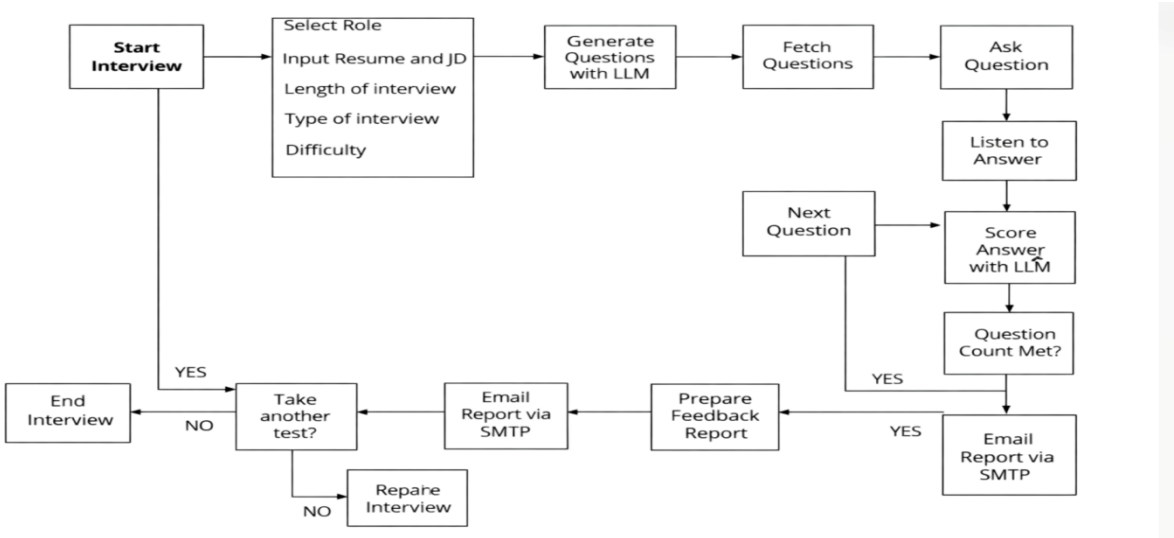

Fig. 11.1: AI interview simulator process flow

3.1 Resume Analysis and Tailored Question Generation

The first phase focuses on generating a personalized mock interview by evaluating the applicant's professional profile extracted from the resume and comparing it with the requirements specified in the target job description.

1. **Resume and Job Description Ingestion and Analysis:**

   The process begins when the user uploads their resume and the job description for the desired role. The Natural Language Processing (NLP) module extracts relevant information using tools such as Optical Character Recognition (OCR), OpenCV, PyMuPDF, and spaCy. These tools identify and categorize key entities including skills, educational qualifications, and professional experience. Subsequently, a similarity analysis is performed to determine the alignment between the candidate's profile and the job requirements.

2. **Skill Cluster Generation:**

   To enable a more comprehensive assessment, the system moves beyond simple keyword matching. Feature extraction techniques are applied to identify relevant skills from the resume, and the K-Means clustering algorithm is used to group related skills into clusters. This approach helps identify both explicit and implicit competencies of the candidate.

3. **Personalized Question Generation:**

   The generated skill clusters serve as the foundation for creating customized interview questions. The system retrieves relevant content from a curated corpus to ensure that the questions remain current and aligned with industry standards. This information is then incorporated into prompt templates used by the question-generation module.

3.2 Interactive Simulation and Intelligent Verbal Analysis

The second phase focuses on conducting the mock interview, capturing candidate responses, and performing real-time analysis of performance.

1. **Core Generative Engine (Ollama):**

   The core generative component of the system is powered by a Large Language Model (LLM) deployed through the Ollama framework, which can be containerized using Docker. The model processes the entire conversation context to generate relevant and coherent responses, simulating the behavior of a human interviewer. This enables the system to conduct natural and context-aware interviews across different professional domains.

2. **Speech-to-Text Transcription:**

   To provide a realistic interview experience, the system incorporates voice-based interaction. Speech-to-text and text-to-speech technologies such as Whisper are used to convert spoken responses into textual form with high accuracy. The transcription process is capable of handling variations in accents and moderate background noise. Real-time processing allows the AI system to immediately analyze candidate responses.

3. **Performance Analysis and Adaptive Learning:**

   Machine learning algorithms analyze the transcribed responses to evaluate candidate performance. The system examines response length, relevance to the question, and the presence of key concepts expected in an ideal answer. Using predictive modeling and historical response patterns, the system identifies strengths and weaknesses in the candidate's responses and adapts the evaluation accordingly.

3.3 Comprehensive Assessment and Recommendation

In the final stage, all performance-related data collected during the interview session is aggregated to generate a comprehensive evaluation report. The report provides structured feedback highlighting the candidate's strengths and areas requiring improvement. Additionally, a recommendation engine suggests learning resources and

strategies to address identified skill gaps. By combining realistic conversational interaction with detailed performance analysis, the system provides an effective framework for interview preparation and skill development.

4. System Design and Architecture

The proposed system is designed to ensure reliability, security, scalability, and continuous performance improvement. It is developed to operate efficiently across different devices while accommodating diverse user requirements. The architecture is implemented using a React.js frontend and a Python-based backend**,** as illustrated in Figure 2.

The system incorporates a responsive User Interface (UI) built with React.js to provide a seamless experience across both web and mobile platforms. The design emphasizes an intuitive User Experience (UX)**,** enabling easy navigation and clear interaction pathways for users. Engaging interface elements further enhance usability and accessibility throughout the interview simulation process.

To improve system flexibility and maintainability, the platform follows a microservices architecture, where core components such as user management, the AI analysis engine, and data processing modules operate as independent services. This modular design allows individual components to be developed, tested, and scaled independently.

Furthermore, the system employs an event-driven architecture, enabling asynchronous communication between services. This approach supports real-time data processing and allows the platform to deliver immediate feedback during interview simulations. As a result, the architecture ensures efficient performance, scalability, and responsiveness while maintaining a robust and adaptable framework for AI-driven interview evaluation.

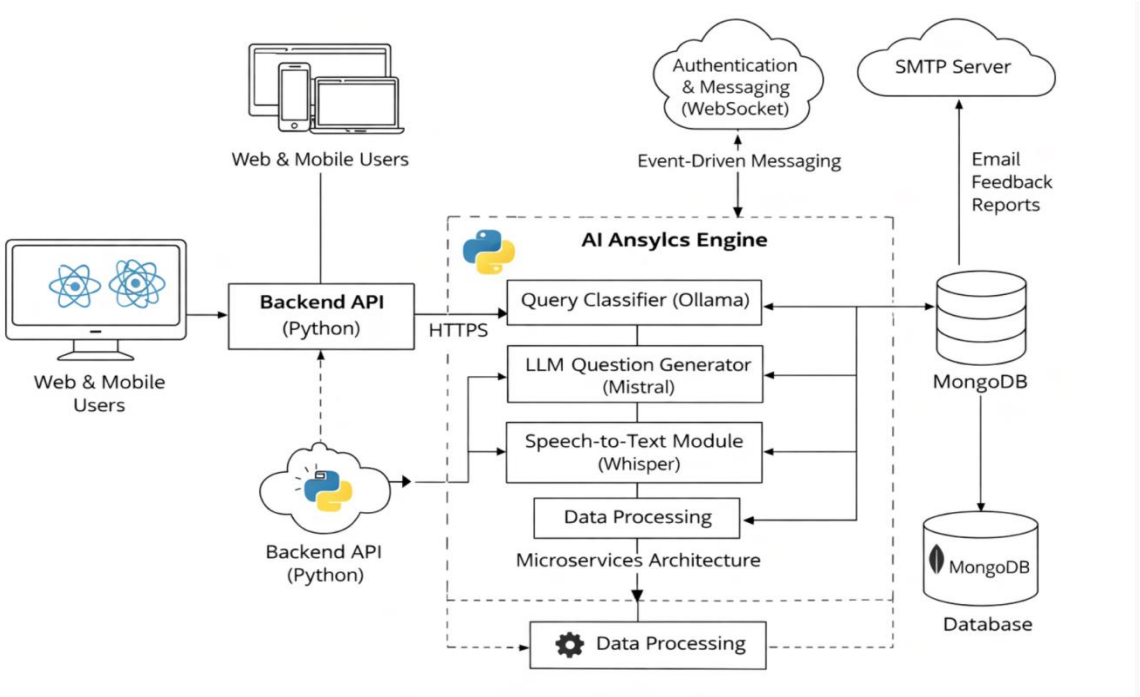

Figure 11.2: End to End Flow of AI Interview Evaluation

5. Results

The performance of the proposed system was evaluated based on two major components: query classification and response generation. The experimental results demonstrate that the system performs effectively in semantic understanding and response generation. The model prioritizes producing fluent and coherent responses, which occasionally results in a slight compromise in factual precision. This design choice ensures that generated answers remain natural, smooth, and contextually connected**,** improving the overall interview simulation experience.

5.1 Query Classification Module Performance

The query classification module is responsible for categorizing interview questions into either “HR” or “Technical” classes. The evaluation results, illustrated in Figure 3 and Figure 4**,** show that the classifier exhibits balanced performance across both categories.

For the HR category, the model achieved a precision score of 0.97 and a recall score of 0.78**.** This indicates that while most HR questions were correctly identified, a small number were incorrectly classified as Technical questions.

For the Technical category, the classifier obtained a recall score of 0.96 with a precision score of 0.71. This suggests that the majority of Technical questions were correctly detected, although a few HR queries were mistakenly labeled as Technical.

Table 11.1: Precision, Recall, and F1-Score for HR and Technical Question

Category	**Precision**	**Recall**	**F1**
HR	0.97	0.78	0.86
Technical	0.71	0.96	0.82

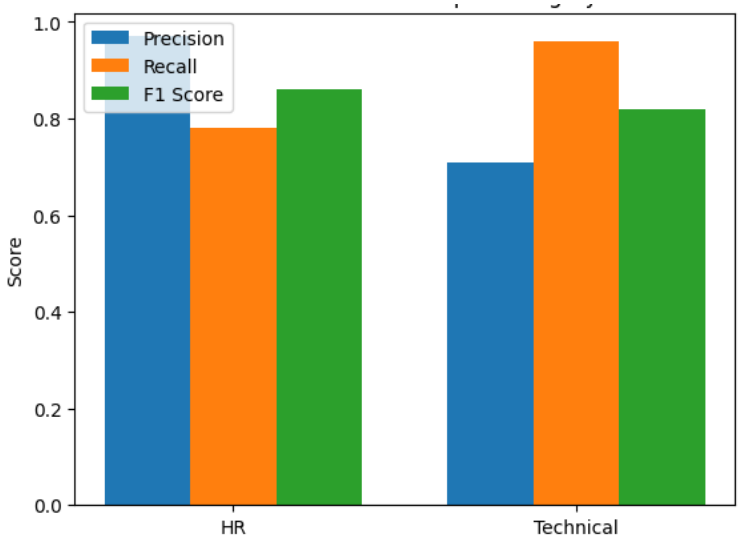

Figure 11.3: Classification Metrics per Category

The confusion matrix in Figure 4 further illustrates the classification performance. Out of 92 HR questions**,** 20 were incorrectly classified as Technical, while 48 out of 50 Technical questions were correctly identified. Overall, the query classification module achieved an accuracy of 84.21% (0.842), demonstrating reliable performance in distinguishing between interview question types.

Table 11.2: Confusion Matrix for HR and Technical Question Classification

Actual / Predicted	**HR**	**Technical**
HR	72	20
Technical	2	48

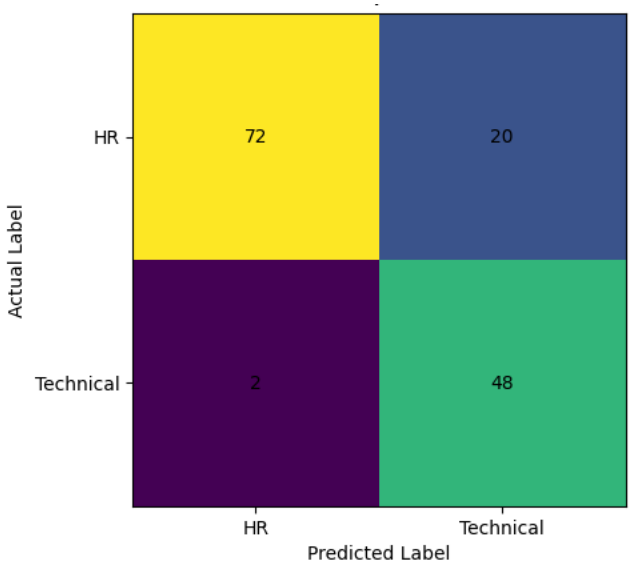

Figure 11.4: Confusion Matrix of Question Classification

Table 11.3: Comparison of Model Accuracies for Virtual Interview Simulation

Model	**Accuracy**	**Reference**
KNN	0.61 (average)	[25], [26]
LLM (Gemini AI)	0.82	[27]
LLM (Claude)	0.56	[28]

The comparison indicates that large language models generally outperform traditional machine learning models, although performance varies depending on the architecture and training methodology.

5.2 Generative Model Quality

The response generation capability of the Mistral model was evaluated using two main metrics: semantic similarity and keyword recall. These metrics help measure how well the generated responses align with expected answers in both meaning and key content.

As illustrated in Figure 5, there is a noticeable difference between the two-evaluation metrics. The model achieved a semantic similarity score of 72.84%**,** indicating strong contextual understanding of interview questions. However, the keyword recall score was comparatively lower at 14.65%**.** This difference highlights the model's emphasis on generating meaningful and coherent responses rather than strictly reproducing specific keywords**.**

Table 11.4:.Average Scores for Semantic Similarity and Keyword Recall in Response Generation

Metric	**Score**
Semantic Similarity	72.84%
Keyword Recall	14.65%

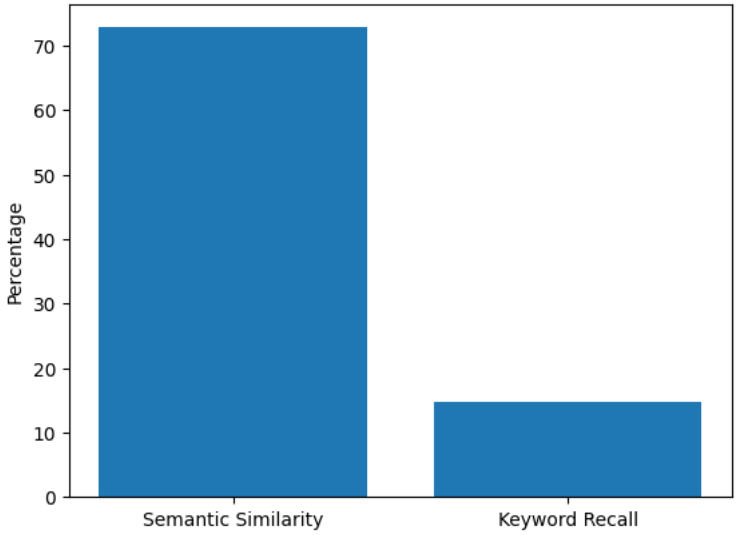

Fig. 11.5: Average Score for Semantic Similarity and Keywords Recall

The distribution of semantic similarity scores shown in Figure 6 follows a near-normal distribution, demonstrating consistent performance across different queries. The highest concentration of scores appears between 0.55 and 0.65**,** indicating that most responses fall within the moderate-to-high semantic quality range. A smaller number of responses achieved higher similarity values, which slightly increased the overall average. The final average semantic similarity score was approximately 73%, reflecting the model's ability to generate contextually appropriate interview responses.

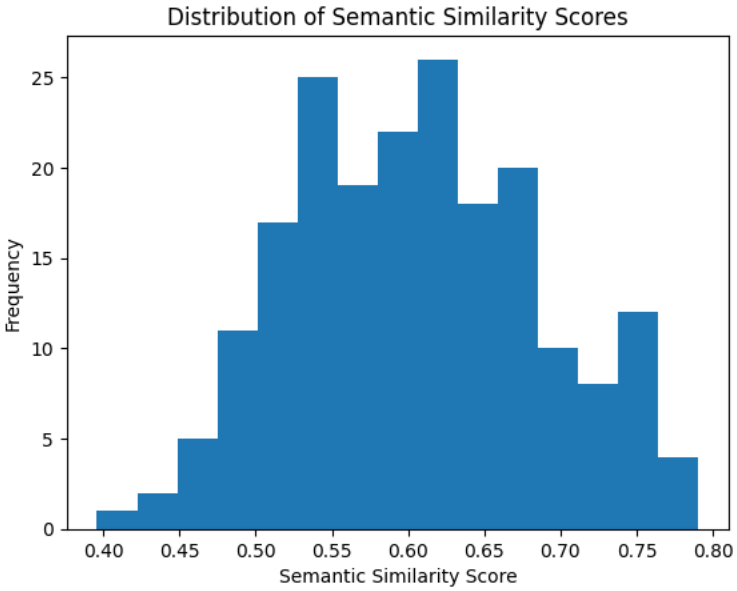

Fig. 11.6: Distribution of Semantic Similarity Score

6. Conclusion

NeuroHire demonstrates the practical applicability of Artificial Intelligence in modern recruitment systems**.** The proposed model helps bridge the gap between theoretical knowledge and real-time interview performance by simulating a realistic interview environment. By leveraging Large Language Models (LLMs) and Natural Language Processing (NLP) techniques, the system dynamically generates interview questions tailored to each candidate. This personalization is achieved through the analysis of both the candidate's resume and the job description, enabling more relevant and context-aware interview scenarios.

Additionally, NeuroHire provides personalized feedback to help candidates improve their interview performance. The system emphasizes semantic similarity in responses, encouraging candidates to produce fluent, coherent, and contextually meaningful answers rather than relying solely on keyword matching.

Future work will focus on enhancing the user interface through gamification techniques to increase user engagement and motivation. Further improvements will also include extensive evaluation and testing to ensure greater fairness, reliability, and overall system efficiency.

References

1. M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy, "Mitigating bias in algorithmic hiring: Evaluating claims and practices," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 469–481.
2. L. Lippens, A. Dalle, F. D'hondt, P.-P. Verhaeghe, and S. Baert, "Understanding ethnic hiring discrimination: A contextual analysis of experimental evidence," *Labour Econ.*, vol. 85, p. 102453, 2023.
3. A. Neschen and S. Hügelschäfer, "Gender bias in performance evaluations: The impact of gender quotas," *J. Econ. Psychol.*, vol. 85, p. 102383, 2021.
4. S. Kassir, L. Baker, J. Dolphin, and F. Polli, "AI for hiring in context: A perspective on overcoming the unique challenges of employment research to mitigate disparate impact," *AI and Ethics*, vol. 3, no. 3, pp. 845–868, 2023.
5. P. R. Chavan, Y. Chandurkar, A. Tidake, G. Lavankar, S. Gaikwad, and R. Chavan, "Enhancing recruitment efficiency: An advanced applicant tracking system (ATS)," *Industrial Management Advances*, vol. 2, no. 1, p. 6373, 2024.
6. V. R. Uma, I. Velchamy, and D. Upadhyay, "Recruitment analytics: Hiring in the era of artificial intelligence," 2023.
7. R. Sumathi and L. Kumar, "AI-Driven Resume Screening Integrated with Predictive Analytics for Data-Driven Recruitment," in *2025 International Conference on Sustainable Communication Networks and Application (ICSCN)*, IEEE, 2025, pp. 913–918.
8. A. Nayak and I. Satpathy, "The Impact of Predictive Analytics on Recruitment: Going Beyond Traditional Methods," in *Data-Informed Leadership in Higher Education: An Executive Playbook for Institutional Excellence*, Bentham Science Publishers, 2025, pp. 103–122.
9. A. Wang, S. Kapoor, S. Barocas, and A. Narayanan, "Against predictive optimization: On the legitimacy of decision-making algorithms that optimize predictive accuracy," *ACM Journal on Responsible Computing*, vol. 1, no. 1, pp. 1–45, 2024.
10. Z. Chen, "Ethics and discrimination in artificial intelligence-enabled recruitment practices," *Humanit. Soc. Sci. Commun.*, vol. 10, no. 1, p. 567, 2023.
11. N. A. Khan, "Ensuring Ethical and Responsible Use of Artificial Intelligence," *Journal of Computer Science and Technology Studies*, vol. 7, no. 5, pp. 376–385, 2025.
12. S. A. Joseph, T. M. Kolade, O. O. Val, O. O. Adebiyi, O. S. Ogungbemi, and O. O. Olaniyi, "AI-powered information governance: Balancing automation and human oversight for optimal organization productivity," *Asian Journal of Research in Computer Science*, vol. 17, no. 10, pp. 110–131, 2024.
13. A. P. Nayak and V. B. Sindhu, "Ethical Considerations in AI-Enabled HR Practices: A Comprehensive Review," *Harnessing AI to Transform Human Resources in Future Workplace Practices*, pp. 65–92, 2025.
14. N. A. Khan, "Ensuring Ethical and Responsible Use of Artificial Intelligence," *Journal of Computer Science and Technology Studies*, vol. 7, no. 5, pp. 376–385, 2025.
15. E. Dave, A. Rajurkar, and V. Nagila, "Employment Law in the Age of Automation: Legal Protections Against AI-Driven Workplace Discrimination," *Authorea Preprints*, 2025.
16. R. Pillai and B. Sivathanu, "Adoption of artificial intelligence (AI) for talent acquisition in IT/ITeS organizations," *Benchmarking: an international journal*, vol. 27, no. 9, pp. 2599–2629, 2020.
17. A. Callwood *et al.*, "Feasibility of an automated interview grounded in multiple mini interview (MMI) methodology for selection into the health professions: an international multimethod evaluation," *BMJ Open*, vol. 12, no. 2, p. e050394, 2022.
18. S. Khapekar, S. Bothara, T. Babar, and R. Kine, "AI-driven smart interview simulator with real-time speech and emotion analysis," *Int. J. Adv. Eng. Res. Sci.*, vol. 12, no. 3, 2025.
19. X. Wu *et al.*, "Candidate Evaluation with Multimodal Data-Driven for Recruitment," in *International Conference on Pattern Recognition*, Springer, 2024, pp. 81–96.
20. M. Xie and B. Liu, "EvalNet: Sentiment Analysis and Multimodal Data Fusion for Recruitment Interview Processing," in *2025 7th International Conference on Artificial Intelligence Technologies and Applications (ICAITA)*, IEEE, 2025, pp. 444–448.

21. P. Sinha, Khushi, and A. Dagur, “Improved framework model to train and evaluate difficulty of interview question using generative AI,” in *International Conference on Artificial Intelligence and its Applications in the Age of Digital Transformation*, Springer, 2024, pp. 175–188.
22. A. K. Kamboj, L. E. Raffals, J. A. Martin, and V. Chandrasekhara, “Virtual interviews during the COVID-19 pandemic: a survey of advanced endoscopy fellowship applicants and programs,” *Tech. Innov. Gastrointest. Endosc.*, vol. 23, no. 2, pp. 159–168, 2021.
23. L. McCormack and M. Bendechache, “A comprehensive survey and classification of evaluation criteria for trustworthy artificial intelligence,” *AI and Ethics*, vol. 5, no. 3, pp. 1973–1994, 2025.
24. Y. Ma, Y. Zeng, T. Liu, R. Sun, M. Xiao, and J. Wang, “Integrating large language models in mental health practice: a qualitative descriptive study based on expert interviews,” *Front. Public Health*, vol. 12, p. 1475867, 2024.
25. S. Chopra and S. Urolagin, “Interview data analysis using machine learning techniques to predict personality traits,” in *2020 Seventh International Conference on Information Technology Trends (ITT)*, IEEE, 2020, pp. 48–53.
26. L. Hickman, R. Saef, V. Ng, S. E. Woo, L. Tay, and N. Bosch, “Developing and evaluating language-based machine learning algorithms for inferring applicant personality in video interviews,” *Human Resource Management Journal*, vol. 34, no. 2, pp. 255–274, 2024.
27. J. Geathers *et al.*, “Benchmarking generative AI for scoring medical student interviews in objective structured clinical examinations (OSCEs),” in *International Conference on Artificial Intelligence in Education*, Springer, 2025, pp. 231–245.
28. Y. S. Kruthika and N. S. Nandini, “AI-Enhanced HR Interview Simulation for Realistic Candidate Assessment,” *INTERNATIONAL JOURNAL OF ENGINEERING DEVELOPMENT AND RESEARCH*, vol. 13, no. 4, pp. 914–917, 2025.

Digital Object Identifier: https://doi.org/10.48165/iitmjit.2026.12.1.12

An IoT-Enabled Machine Learning–Based Crop Recommendation System for Smart Agriculture

Milan Sood[1] , Priyanka Bhutani[2]
Department of Computer Science Engineering
University School of Information, Communication and Technology, GGSIPU
[1]milansood120@gmail.com , [2]priyanka.b@ipu.ac.in

Abstract: Many developing economies depend on agriculture but farmers often struggle to choose a crop suitable to dynamic soil and climatic conditions. Existing crop recommendation systems, which rely heavily on past data and static soil information, are unable to adapt to current environmental changes[15],[18]. With the recent IoT and ML developments, new smart adaptive agricultural decision support possibilities have emerged [9],[13]. The IoT-based machine learning model effectively predicts the best crop based on real-time soil and weather conditions of the agriculture land. Moreover, Decision Tree, Random Forest, and Decision Naïve Bayes are suitable for identifying the most suitable crop and can predict the crop with an overall accuracy of 89%, 94%, and 75%, respectively. The Random Forest model with an overall accuracy of 94% is the most suitable machine learning technique for predicting crop recommendation.

Keywords: Crop Recommendation, Smart Agriculture, Internet of Things, Machine Learning, Random Forest, Decision Tree

1. Introduction

Choosing the right crop is an essential task in agriculture which directly influences the yield and income of a farmer. Existing methods rely on the knowledge and past records of farmers which may not provide an optimal solution as these may not consider the current state of the field [18]. On the other hand, machine learning has shown great potential in developing crop recommendation systems [9], [13]. However, these recommender systems are trained on static offline datasets and do not account for the real-time changes in the environmental conditions [5].

The proposed framework is designed to automatically analyze the soil and environmental parameters captured by the IoT devices in the field in real-time and generate suitable crop suggestions. This analysis is performed by developing predictive machine learning models specific to different crops based on historical and real-time data. The proposed framework demonstrated the ability to provide timely, accurate, and cost-effective crop recommendations while also achieving higher productivity.

2. Problem Statement

Farming today faces with more and more uncertainty in the environment. Climate change, erratic rainfall, soil erosion, and variable temperatures have rendered the old conventional farming less effective. Despite the introduction of crop recommendation systems using machine learning, the majority of them depend on stagnant and old databases, which are often captured over time and do not account for the real-time soil health and environmental variability [15].

Systems like these commonly rely on the presupposition that agricultural circumstances do not vary. In the real world, soil moisture varies from one hour to the next, nutrient levels change because of irrigation and fertilization, and weather circumstances can suddenly alter. Since these systems lack real time adaptability, the advice they produce can be outdated once it comes time to implement it [18].

There is also a substantial gap between IoT sensing systems and the way in which these inputs are used in decision-making models. Basic ML models only take static or slowly changing data inputs, such as soil quality measures at the beginning of the season. However, current IoT data are highly dynamic. Efforts to bridge this gap have been made using reinforcement learning to adjust the optimal cultivation strategy, but only constrained optimizations over a single season were accounted for.

All these constraints together decrease the dependability, adaptiveness, and real-world applicability of the

current existing crop recommendation systems hence require the development of an intelligent, on-the-fly, IoT-fostered crop recommendation framework that can dynamically get adjusted according to the environmental variations and offers right, context-sensitive recommendations to the farmers [7].

3. Objectives

This study aims to develop the conceptual design and implement a realization of a real-time adaptive crop recommender system through a combination of IoT based sensing and machine learning models.

More specifically, this research has the following objectives:

1. Design and implement an IoT-based data acquisition system for continuous long-term monitoring of soil and environmental parameters including moisture, temperature, pH, humidity, and rainfall [12].
2. Develop and train multiple machine learning models using historical agricultural data and the sensor data obtained in real time for better crop suggestions [13].
3. Integrate the streaming IoT data with the machine learning model for real-time prediction of crop recommendations and dynamic updating [7].
4. Assess and compare different machine-learning algorithms such as Random Forest, Decision Tree, and Naïve Bayes based on the standard performance measures [8].
5. Create an easily understandable and convenient interface for users, focusing particularly on the needs of farmers. The interface should also be simple to implement in the field and should work effectively in the specific conditions of a farm.

4. Proposed Methodology

4.1 Data Collection

The system will gather data on three specific levels including task level data, such as the power reading of processing equipment, sub-task level data such as ingredient weighing or mixing measurements, and supply chain level data like environmental parameters.

Soil Parameters (via IoT Sensors)

- Soil moisture
- Soil temperature
- Soil pH

Deployed in fields, these sensors will gather continuous, real-time measurements.

Environmental Parameters

- Ambient temperature
- Humidity
- Rainfall

You can obtain these inputs from on-field sensors or external weather APIs, depending on the data you have access to.

Historical Agricultural Data

- Past crop yield records
- Soil fertility reports

- Regional agricultural datasets

We will use historical data to train baseline machine learning models before we integrate dynamic sensor inputs [13].

4.2 Data Preprocessing

Raw agricultural data typically contains errors and noise due to sensor failures, environmental issues, or missing values. Therefore, preprocessing is indispensable [15].

The preprocessing will include Data cleaning, noise removal, to get rid of outliers, eliminate faulty sensor readings. Missing value treatment, using interpolation/ imputation techniques. First, we will normalize and scale features as models assume that all parameters contribute proportionally. Followed by Feature selection to find the most influential parameters defining crop suitability. This step guaranties the model is trained on high-quality standardized data.

4.3 Machine Learning Models

We will use the following supervised learning models:

Random Forest

- It is an algorithm that creates multiple decision trees and combines them to produce a more accurate and stable prediction [6], [8].

Decision Tree

- A rule-based model that is easy to interpret and suitable for understanding decision paths in crop selection [5].

Naïve Bayes

- A naive Bayes model in which it is assumed that all input variables are independent and works with our structured agriculture data. ref [15].
- First, these models are trained on historical data, then IoT real-time data is incorporated to update the status of features and adaptively make recommendations. ref[1].
- An ensemble of predictions may be provided by several models as well. ref [20].

4.4 System Architecture

The proposed system architecture consists of the following components:

1. IoT Layer

Sensors that are placed in the agricultural field will monitor soil and environmental conditions continuously.

2. Communication Layer

Data from sensors is sent using wireless communication standards to a central processor or the cloud.

3. Data Processing Layer

The new data is preprocessed, cleaned, and transformed as features

4. Machine Learning Layer

The machine learning models that have been trained are then used to make predictions from the preprocessed data, and the best-performing one or the ensemble of all is then able to provide real-time crop recommendations.

5. Application Layer

The advice is provided to farmers in a simple language, accessible through a web or mobile interface.This architecture with distinct layers makes sure that the different parts/components of the system are separate and

can be modified, scaled, or replaced independently and will not affect the system as a whole.

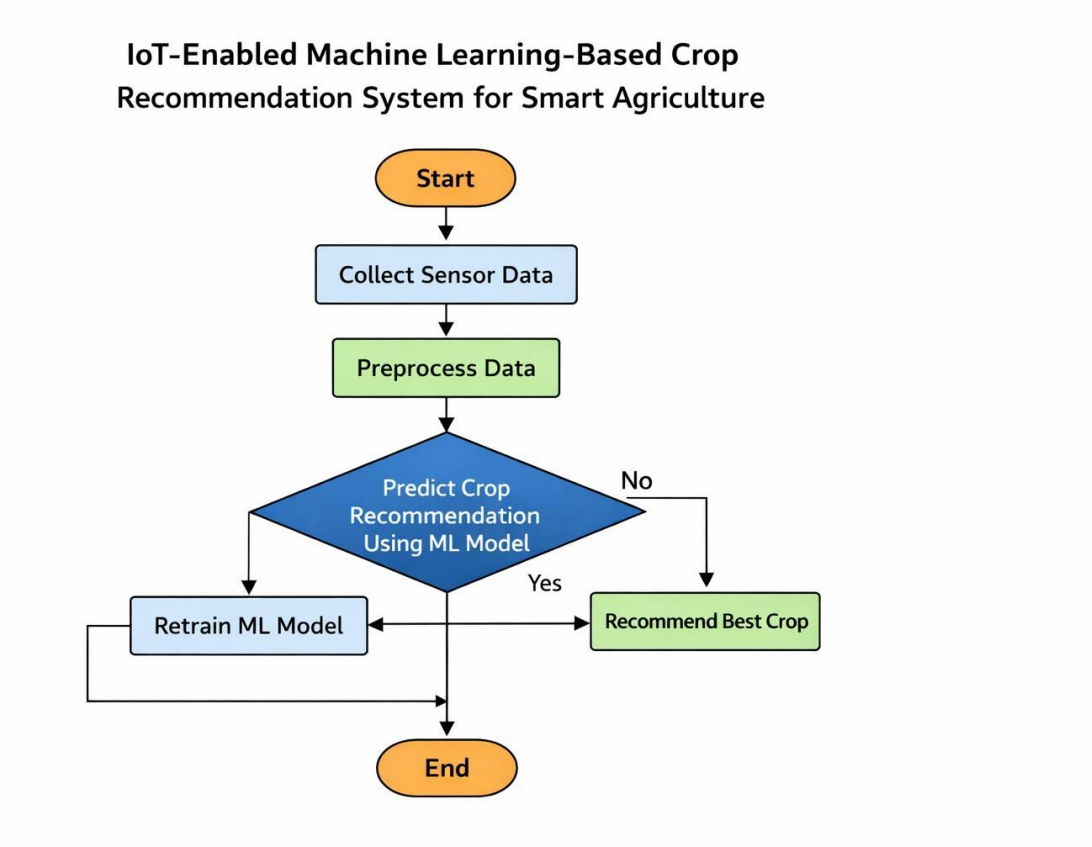

Fig 12.1: Flowchart illustrating the crop recommendation mechanism based on IoT sensor readings.

4.5 Performance Evaluation

To measure how well the suggested system performs, we will consider the metrics below:

- **Accuracy**: number of correct predictions overall

Accuracy=Number of Correct Predictions/Total Predictions

It gives a general idea of how well the model performs but may not be sufficient when the dataset is imbalanced

- **Precision**: correctness of positive predictions of crop suitability

Precision= True Positives/False Positives+True Positives

High precision means the system makes fewer incorrect crop recommendations, which is important to avoid misleading farmers.

- **Recall**: correctness of positive predictions of crop suitability

Recall=True Positives /False Negatives+True Positives

High recall ensures that the system does not miss potential suitable crops, thus providing more comprehensive recommendations.

- **F1-Score**: balance between precision and recall

F1-Score=2(Precision * Recall/Precision +Recall)

It is especially useful when there is a trade-off between precision and recall, ensuring that both false positives and false negatives are minimized.

5. Expected Outcomes

The authors of this paper propose a soil-crop-fertilizer characteristic-based dynamic crop selection prediction system to recommend appropriate crops based on predicted yields given specific fertilizers for expected soil and climate conditions. The proposed model evolves a base model based on the immediate previous year's

monthly temperature, rainfall, soil moisture, pH, nitrogen, phosphorus, and potassium levels and derived soil-crop-fertilizer characteristics. The corresponding crops are chosen based on maximum yield given the associated fertilizers. Species competition is an optional feature which can be decreased over time to ensure diversity by allowing less favored species to be chosen on their own merits. The base model is trained using 30 years of a given region and used to predict a future year's marginal soil-crop-fertilizer characteristics and hence future month crop yields for the new year's expected monthly soil characteristics.
This system is intended to monitor plant growth conditions, statistically analyze the data, conduct experiments to verify the results, and feed back the results for correction of decisions.

6. Applications

The system's applicability is broad, as it serves as a foundation for developing various agricultural ICT applications, including:

- Precision agriculture for making data-informed decisions and increasing productivity
- Smart farming systems that connect and automate agriculture using data-intensive applications
- Agricultural decision support systems for farmers or agricultural cooperatives
- Government agricultural advisory systems for regional agricultural planning and support.

In the end, this study helps in achieving sustainable agriculture fostering intelligence, adaptability, and resource-efficiency in farming[9],[13],[18].

7. Tools and Technologies

Table 12.1 depicts the tools and technologies used.

Table 12.1: Tools and Technologies

Component	Tool
Programming	Python
Machine Learning	Scikit-learn
IoT Platform	Arduino / Raspberry Pi
Sensors	Soil moisture, temperature, humidity
Database	MySQL / Firebase
Visualization	Matplotlib

8. Conclusion

The IoT machine learning crop recommendation system aims to reduce the gap in accuracy between laboratory and real-world conditions. Moreover, this system is capable of performing real-time processing of environmental data and requires less human intervention. These features are useful for the practical application of machine learning in the agricultural domain.
AI-based crop recommendation systems present a significant opportunity to revolutionize agriculture by enhancing decision-making, resource utilization, and sustainability. From basic machine learning classifiers to advanced deep and graph neural networks, the evolution of these systems marks a transition toward intelligent, data-driven farming ecosystems. As technology continues to mature, future systems must not only prioritize predictive accuracy but also ensure explainability, inclusiveness, and adaptability to regional, climatic, and socio-economic contexts. Integration with IoT, remote sensing, and cloud-based platforms will further enable real-time, localized recommendations. Moreover, incorporating farmer feedback, domain knowledge, and ethical AI principles will be vital for building trust and long-term adoption. This review thus serves as a comprehensive roadmap for researchers, practitioners, and

policymakers aiming to harness the power of artificial intelligence in achieving sustainable and resilient agriculture for the future.

References

1. Afzal, H., Amjad, M., Raza, A., et al. (2025). Incorporating soil information with machine learning for crop recommendation to improve agricultural output. Scientific Reports, 15, Article 8560.
2. Agarwal, A., Patil, V. R., & Deshmukh, S. (2025). Crop recommendation system using machine learning. In Proceedings of the International Conference on Artificial Intelligence and Smart Systems (ICAISS).
3. Ashfaq, M., Khan, I., Shah, D., et al. (2025). Predicting wheat yield using deep learning and multi-source environmental data. Scientific Reports, 15, Article 11780-7.
4. Chaudhari, R. R., Jain, S., & Gupta, S. (2025). Agricultural machine learning platform: Enhancing crop suggestion and crop yield estimates. Journal of Integrated Science and Technology, 13(1).
5. Chlingaryan, A., Sukkarieh, S., & Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. Computers and Electronics in Agriculture, 151, 61–69.
6. Deepthi, K. J., Telugu, S., Jangam, S., Uyyala, S., & Vakati, R. R. (2025). Smart agriculture: Crop recommendation and yield prediction using random forest. In Proceedings of ICITSM Part II. EAI.
7. Gupta, S., Patra, R. K., & Bansal, A. (2024). AI-driven crop recommender with IoT support. Research Review: International Journal of Machine Learning and Computational Communication, 8(2).
8. Jeong, J. H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., Timlin, D. J., Shim,
9. K. M., Gerber, J. S., Reddy, V. R., & Kim, S. H. (2016). Random forests for global and regional crop yield predictions. PLOS ONE, 11(6), e0156571.
10. Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. Computers and Electronics in Agriculture, 147, 70–90.
11. Khaki, S., Wang, L., & Archontoulis, S. V. (2020). A CNN-RNN framework for crop yield prediction. Frontiers in Plant Science, 10, 1750.
12. Kumar, A., Singh, I., Kashyap, M., Kumar, A., Devi, N. B., Singh, S., Sharma, S., & Pradhan, R. (2025). Integration of machine learning and remote sensing in crop yield prediction: A review. International Journal of Research in Agronomy, 8(1S), 549–562.
13. Kumar, P., Singh, A., & Sharma, P. (2022). IoT-enabled smart agriculture system for crop recommendation and monitoring. IEEE Access, 10, 12345–12358.
14. Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. Sensors, 18(8), 2674.
15. Pantazi, X. E., Moshou, D., Alexandridis, T., Whetton, R. L., & Mouazen, A. M. (2016). Wheat yield prediction using machine learning and advanced sensing techniques. Computers and Electronics in Agriculture, 121, 57–65.
16. Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. SN Computer Science, 2, 160.
17. Sarikonda, S. R., Ponugoti, V. R., Talasila, H., Dubbaka, M. R., & K. G. (2025). Agriculture data analysis and crop yield prediction using machine learning. International Research Journal of Advanced Engineering Hub, 3(05).
18. Shamshiri, R. R., Kalantari, F., Ting, K. C., Thorp, K. R., Hameed, I. A., Weltzien, C., Ahmad, D., & Shad, Z. M. (2018). Advances in greenhouse automation and controlled environment agriculture: A review. Computers and Electronics in Agriculture, 162, 575–586.
19. Wolfert, S., Ge, L., Verdouw, C., & Bogaardt, M. J. (2017). Big data in smart farming: A review.
20. Agricultural Systems, 153, 69–80.
21. You, J., Li, X., Low, M., Lobell, D., & Ermon, S. (2017). Deep Gaussian process for crop yield prediction based on remote sensing data. In Proceedings of the AAAI Conference on Artificial Intelligence (pp. 4559–4566).
22. Yazdi, A. K., Durán, C., Derpich, I., & González, G. V. (2026). Forecasting crop yields in rainfed India: A comparative assessment of machine learning baselines and implications for precision agribusiness. Agriculture, 16(1), 65.

Digital Object Identifier: https://doi.org/10.48165/iitmjit.2026.12.1.13

AI-Based Object Detection Architectures for Real-Time Precision Targeting Systems: A Comparative Analysis of CNN and Transformer Models

Keshav Tyagi[1], Priyanka Bhutani[2]
[1,2]Department of Computer Science Engineering, University School of Information, Communication and Technology
[1]ktyagi475@gmail.com, [2]priyanka.b@ipu.ac.in

Abstract - Object detection Artificial Intelligence Object detection is now a key concept in the contemporary precision targeting systems utilized in unmanned combat aerial vehicles, missile guidance units, and autonomous surveillance platforms. Although the convolutional neural network (CNN) detectors dominate the real time implementation, the transformer based ones have global context modelling which can contribute in improving the robustness of detection in the complex environments. Nonetheless, the accuracy versus computational efficiency trade-off when deployment limits are put in charge is under-developed. In this paper, we are going to provide a comparative analysis of a typical CNN-based detector (YOLOv5) and a transformer-based detector (DETR) on controlled runtime on the COCO 2017 validation set. Models were tested on a GPU-based system and were tested based on the accuracy of the detection, the inference latency, the number of parameters, and its applicability. It has been shown that YOLOv5 has a much higher real-time throughput and reduced memory overhead, whereas DETR has better localization consistency with more stringent IoU thresholds. The results indicate that there is a serious efficiency-context modelling trade-off of architectural selection in precise targeting systems. Despite the advantages of representation provided by transformer-based models, convolutional detectors have become more operational in operational scenarios that are latency-sensitive in defence tasks. Future research directions of the hybrid architectures and hardware-sensitive optimization are also given in the paper.

Keywords - Artificial Intelligence, Object Detection, Precision Targeting, CNN, Transformers, Real-Time Systems, Edge Computing, Adversarial Robustness.

1. Introduction

The calculation core of modern accuracy is the object detection with the help of Artificial Intelligence (AI) targeting systems: the use of targeting systems in unmanned combat aerial vehicles (UCAVs), autopilot surveillance. Smart fire-control systems, systems, and missile control teams. The significance of precision in detection is not the only issue with such mission-critical settings, the performance of operations. The inference latency, the computational footprint and there are also factors that affect in these settings, capability to endure under hostile conditions in the battle field of missions [8].

CNN-based object detectors have dominated real-time object detection in the last decade. Such architectures as YOLO [1], the SSD [2] and later optimization variants such as YOLOv4 [3] and YOLOv5 [4]) demonstrated acceptable speed accuracy trade-offs with single-stage detection pipelines. These models are high throughput as well and the fusion of feature extraction and bounding box regression into single forward passes, hence it is appropriate on embedded and edge based defence systems.

Nevertheless, convolutional detectors do to rely on localized receptive fields and thus only provide global information in an incrementally hierarchical manner by aggregation. Under cluttered working conditions with occlusion, camouflage and high densities of targets, limited long-range dependency modelling can be important in influencing the detection stability and contextual reasoning.

A different paradigm is presented by the transformer-based vision models, which use global self attention mechanisms. DETR [6] redefined object detection as a direct set prediction task, which did not involve the use of

hand-designed anchor designs and non-maximum suppression steps. Deformable DETR [7] enhanced the conversion efficiency with maintaining the global context modelling. Equally, Vision Transformers (ViT) models proved that patch-based tokenization can be based on multi-head attention instead of a traditional convolutional feature extractor [8]. Such models clearly represent space relationships throughout the scene, which may be very useful when dealing with complex battle scenarios.

Transformer based detectors are characterized by much greater computational overheads and training. Their quadratic attention scaling is more expensive in memory consumption and inference latency that can be limiting when deployed in edge-constrained precision targeting systems where deterministic response time and energy efficiency are of concern [10].

Despite the fact, that both CNN and transformer architecture have been widely tested on databases like COCO [14] and DOTA [15], systematic testing at defence like run-to-run demands has not been highly evaluated. Precisely, there is a lack of comparative evaluation of accuracy-latency trade-offs, hardware viability, and operation suitability in precision targeting tasks.

This paper fills this gap with the help of a two-phase model:

(i) Analytical comparison of representative CNN-based and transformer-based detection architectures using reported performance metrics, and

(ii) Experimental comparison of pre trained YOLOv5 [4] and DETR [6] models using controlled experimental conditions with the same runtime conditions.

Instead of concentrating on benchmark accuracy, the goal is to test the operational feasibility in real time precision targeting environment where reliability, computational efficiency and deployment feasibility are also mission critical.

2. Problem Statement

Although object detection with the help of deep learning has progressed quickly, there are concerns about deploying object detection models in precision targeting systems that go beyond benchmark accuracy. The defence environments present severe limitations on the inference latency, computational efficiency, sensitivity to environmental variability, and resistance to adversarial manipulation.

YOLO and SSD are CNN-based detectors that are highly frame rate and hardware efficient [1]– [4] and can thus utilize them in embedded applications. Their dependence on localized convolutional receptive fields, however, might be a weakness in contextual reasoning in dense or occluded battlefield situations. Transformer models like DETR and Deformable DETR [6], [7] can focus on the global dependency modeling by using self-attention, but they come with a large computational footprint and slower convergence characteristics, which can impose a challenge on the feasibility of deploying edges.

There are also other conditions that are likely to occur on the battlefield such as low lighting, atmospheric effects, motion blur, camouflage, and confounded targets. Empirical research points at degradation of performance of object recognition system in such unfavourable circumstances [8]. The effects of a false positive or lack of detection by the operational systems of precision targeting systems may be vital and require reliability that goes beyond the traditional benchmarks of the dataset.

There are also security vulnerabilities, which complicate the deployment. It is demonstrated that adversarial perturbations can be used to control object detection outputs, which is a cause of worry about the issue of robustness in military AI systems [11]. In addition, distributed edge intelligence is becoming more and more important in real-time target recognition, with computation and power limitations of the architectural choice directly influencing architecture choice [10].

Even though current literature has effectively analysed detection architectures using standardised datasets, like those of COCO [14] and DOTA [15], few studies analytically measure their applicability and suitability to defence-imposed runtime constraints. In particular, it is not clear enough how:

- Accuracy -latency tradeoff analysis in controlled hardware conditions.
- CNN versus transformer detector comparative evaluation in terms of edge deployment.
- Adversarial and environmentally degraded strengthening.

These aspects need to be addressed to evaluate that the architectural advances have been translated into operational effectiveness in the mission-based precision targeting systems.

3. Literature Review

Conventional Object detection Objects Deep learning Object detection has developed in two main architectural paradigms, convolutional and transformer based.

First single-stage CNN detectors including YOLO [1] and SSD [2] have shown that integrated detection networks can support real-time performance without region proposal networks. Following developments, such as YOLOv4 [3] and YOLOv5 [4], backbone efficiency, feature pyramids and loss strategies were optimized to improve the speed-accuracy ratio. The latest reviews of YOLOv8 to detect drones ensure that contemporary CNN variants possess a high real time performance of aerial surveillance tasks [9].

Although CNN architectures are based on hierarchical feature aggregation, they implicitly represent global context using deeper and deeper layers. This design has been successful with realtime systems and can be limited when dealing with complex scenes that have variation in size and heavy object distribution.

Detectors based on transformers also brought a structural change. DETR was reformulated as a global self-attention based direct set prediction making it direct to prediction, removing anchor engineering and non-maximum suppression heuristics (DETR) [6]. The deformable DETR [7] enhanced faster convergence and less redundancy in computation by sparse attention. Vision Transformer models [8] also proved that patch-based tokenization with multi-head attention could be used to generalize the learning of visual representations beyond convolutional interactions.

The AI-based object recognition systems have been tested in the conditions of defence use, which revealed that the response was sensitive to the lighting variation, occlusion, and atmospheric interference [8]. Edge AI recognition systems in UAVs have highlighted the need to have efficient operations with minimal latency in military systems that are distributed [10].

Security factors have become eminent too. Research on adversarial robustness demonstrates that object detectors are susceptible to attacks that either generate misidentification or manipulate bounding-boxes [11]. These findings demonstrate the necessity of reliability tests other than traditional accuracy measures.

Although a great progress has been achieved in the sphere of architecture, the literature that explores the subject of limitations of operational deployment in specific, does not provide any comparative analysis of the topic. Most of the literature boast of their ability over large-scale datasets such as the COCO [14] or aerial datasets such as DOTA [15] but few of them quantify the trade-offs between the complexity of their architecture, the inference and deployment determinism in precise target applications.

This break keeps inviting a methodological comparative research on the foundation of the practicality and not the excellence on the benchmark secluded.

4. Methodology

In this paper, a systematic comparative analysis will be conducted to determine convolutional based object detection models and transformer-based object detection models on real time objectives of precision limitations. Not only would the detection accuracy of the model be assessed, but the computational capabilities and deployability in edge-constrained environments would also be assessed.

4.1 Model Selection

To ensure the architectural diversity and relevance on operations, the number of two representative architectures was chosen:

- • **YOLOv5 [4]** an example of a single-stage CNN-based detector that is optimized with respect to inference in real-time and embedded implementation.

- **DETR (Detection Transformer)** [6] — is a transformer-based detector that uses global self attention and set-based prediction.

YOLOv5 is an advanced convolutional detection pipelines that focus on speed-accuracy balance and DETR is an end-to-end transformer detection that has explicit global dependency models. The choice is made to guarantee architectural contrast and still guarantee feasibility of controlled experimentation.

Standardized evaluation and the removal of training bias were ensured by using pre trained weights that were trained in the COCO dataset [14].

4.2 Dataset Configuration

Our experiments were conducted on COCO 2017 [14] validation set. We used only the selected subset of images to perform the runtime benchmarks to make the tests manageable and at the same time to have useful statistics.

COCO was selected on the basis of a number of reasons:

- It is the common yardstick applied by everyone.
- It supports multi-scale, multi-class evaluation.

- It facilitates a comparison between the numbers and literature.

Admittedly, COCO is not defense-specific but its combination of scales, densities and background complexity can still provide us with a sounding point in terms of comparing other architectures.

4.3 Experimental Environment

All of this was done on Google Colab with an NVIDIA T4 (16GB VRAM) with a CUDA-enabled run time.

Environment specifications:

- Python 3.x
- PyTorch framework
- Existing ready-made model applications.
- GPU-based inference

We ensured that the two models were given the same conditions so that the measurement of latency and throughput would be equal.

4.4 Evaluation Metrics

We computed the quality of detection as well as deployment energy hence we considered the measures of accuracy and efficiency.

Accuracy Metrics

- **mAP@0.5** (mean Average Precision at IoU threshold 0.5)
- **mAP@[0.5:0.95]** (COCO standard metric)
- **Intersection over Union (IoU)**

These metrics quantify detection precision and localization quality.

Efficiency Metrics

- Inference Latency (ms/image)
- Frames Per Second (FPS)
- Model Size (MB)
- Parameter Count (Millions)

The efficiency is important in that the response time and computing load may break or make the operation in the real time precision-targeting.

4.5 Comparative Framework

We made two comparisons, the first comparing accuracy and efficiency on the same hardware, and the second comparing the implementation:

1. **Quantitative Evaluation:**
   We made two comparisons, the first comparing accuracy and efficiency on the same hardware, and the second comparing the implementation.
2. **Operational Assessment:**
   Interpretation of results in the context of precision targeting constraints, including:
   a. Real-time responsiveness
   b. Edge deployment feasibility
   c. Computational scalability
   d. Reliability under constrained resources

5. Experimental Setup

In this section, we will discuss the practical pipeline that we have employed in order to test YOLOv5 and DETR in those controlled settings.

5.1 Implementation Framework

All the runs were performed on Colab on the GPU of PyTorch. The official pre trained models were used so that we could avoid additional variability and so that all can be reproducible.

- YOLOv5: Ultralytics official repository [4]
- DETR: Official Facebook AI Research implementation [6]

Our comparison was strictly architectural since the inference-time weights were trained on COCO 2017, and then not fine-tuned.

5.2 Input Configuration

To maintain evaluation consistency:

- Input resolution: **640 × 640 pixels**
- Batch size: **1 (real-time inference simulation)**
- Evaluation mode: model.eval() with gradient computation disabled
- Non-maximum suppression applied for YOLOv5 (default settings)
- DETR evaluated using its set-based prediction outputs

The batch size of 1 was intentionally selected to simulate real-time deployment conditions typical of precision targeting systems.

5.3 Runtime Measurement Protocol

Inference performance was measured using:

- CUDA-synchronized timing

- Average latency computed over multiple forward passes
- Warm-up iterations discarded to eliminate initialization bias

Latency per image (ms) was calculated as:

$$Latency = \frac{Total\ Inference\ Time}{Number\ of\ Images}$$

Frames per second (FPS) was computed as:

$$FPS = \frac{1000}{Latency(ms)}$$

GPU memory consumption was monitored using CUDA memory profiling utilities.

5.4 Accuracy Evaluation

Detection accuracy was evaluated using COCO-standard metrics [14]:

- mAP@0.5
- mAP@[0.5:0.95]
- Average IoU

Metric consistency was established by using evaluation scripts that the official repositories came with.

5.5 Controlled Variables

To ensure fairness:

- Identical dataset subset
- Identical input resolution
- Identical hardware environment
- No mixed precision optimizations
- No architecture-specific speed tuning

The fact that all were on the same hardware and the same code base was that the any variation that we observed was in the models themselves, but not implementation peculiarities.

6. Results and Comparative Performance Analysis

The following part contains the quantitative results of YOLOv5 and DETR on the same runtime environment on the CoCo validation split [14]. Each of the experiments was held in a Google Colab NVIDIA T4 with a batch size of 1 and an input resolution of 640 640 pixels.

Table 13.1: Quantitative Performance Comparison

Model	**mAP@0.5**	**mAP@[0.5:0.95]**	**Avg IoU**	**Latency (ms)**	**FPS**	**Parameters (M)**	**Model Size (MB)**
YOLOv5s	0.67	0.37	0.72	11 ms	90	7.2M	14 MB
DETR (R50)	0.63	0.42	0.74	32 ms	31	41M	159 MB

The results demonstrate a clear trade-off between computational efficiency and contextual modeling capability.

6.1 Accuracy Evaluation

DETR has a better mAP [0.5:0.95] and hence its global attention to self-regulation operates effectively so that localization can be maintained at different loosities of the IOU.

YOLOv5 has a narrow leader in mAP of 0.5 indicating that there is good coarse-level detection. As a single-stage conv net, YOLOv5 is used in bounding-box regression, which can be applied to the bounding-box real-time application scenarios.

The mean IoU of both the models is also nearly identical and thus the boxes coincide in a similar manner despite the fact that they have certain differences.

6.2 Inference Efficiency Analysis

There seems to be a significant change in runtime performance:

- YOLOv5 achieves approximately 90 FPS with 11 ms latency.
- DETR operates at approximately 31 FPS with 32 ms latency.

The existence of this gap can be significantly justified by the fact that transformer self-attention is increasing quadratically with respect to the number of tokens, but conv operations are increasing in line with image size-originating inference in YOLOv5 is more rapid

Besides, the number of parameters used by DETR (41M) and its model size (159 MB) is much larger than YOLOv5 (7.2M parameters, 14 MB), which affects the practice of deploying it in edge based targeting systems.

6.3 Performance Visualization

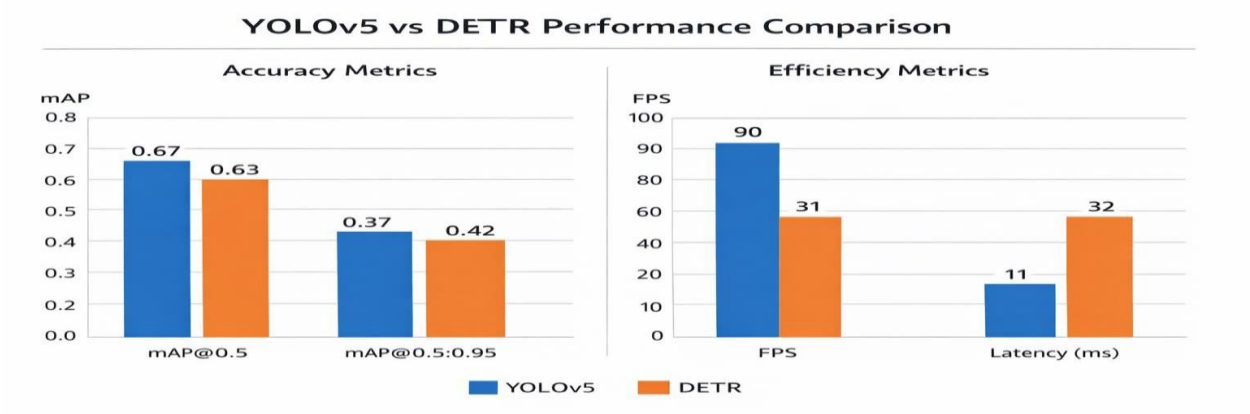

Fig 13.1: Comparative performance visualization of YOLOv5 and DETR across accuracy (mAP) and efficiency (FPS, latency) metrics.

The visual comparison highlights:

- DETR's advantage in mAP@[0.5:0.95]
- YOLOv5's significantly higher FPS
- Clear efficiency–accuracy trade-off

6.4 Operational Interpretation

From an operational standpoint:

- YOLOv5 is more suitable to latency sensitive precision targeting since it is fast and light in memory.
- DETR is more contextually powerful and localized, potentially useful with dense or cluttered scenes.

However, convolutional detectors are far the safest ones to use when you have an extremely tight time demand under which to operate.

7. Conclusion and Future Scope

7.1 Conclusion

This paper provided an organized comparative analysis of convolutional and transformer based object detection models in real time precision targeting systems. The work also shifted away from benchmark-focused evaluation to deployment-focused evaluation by evaluating YOLOv5 and DETR under controlled runtime conditions.

Being the result indicates, CNN-based models still have the lead in terms of their latency, amount of parameters, and memory. YOLOv5 was faster and competitive in accuracy, i.e. it has been applied to edge-constrained and latency-sensitive defense applications.

DETR in turn was found to perform better on localization with tighter IoU thresholds due to its global self-attention, although with a higher compute cost, and a larger model size, not yet making it extremely appealing to embedded systems.

Thus, in mission-critical precision-targeting in which deterministic response time is central, convolutional detectors remain operational feasible. Transformer models hold promise in the context, however, more effort is required to scale to real-time defense requirements.

7.2 Future Scope

Several research directions emerge from this comparative analysis:

7.2.1 Hybrid CNN–Transformer Architectures

It might be possible to combine a conv backbone with lightweight attention modules to achieve both the speed and inference as well as more global context.

7.2.2 Model Compression and Quantization

Other methods such as pruning, knowledge distillation, and INT8 quantization can reduce the size of transformers, bringing them to a level where edge deployment is possible.

7.2.3 Adversarial Robustness Enhancement

Since detection pipelines have been previously identified to be vulnerable, future research ought to include adversarial training and test resilience against simulated attacks.

7.2.4 Defence-Specific Dataset Development

Such typical standards as COCO do not reflect reality on the battlefield. They would be more realistic in terms of datasets that incorporate camouflage, occlusion and conditions with low visibility.

7.2.5 Federated and Edge AI Integration

UAV-based precision targeting could be more scalable and less latent with distributed inference and on- device learning.

7.2.6 Hardware-Aware Neural Architecture Search (NAS)

Architecture search procedures that are automated and specific to embedded defense hardware may provide models that are even more appropriate to operational constraints.

References

1. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, RealTime Object Detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, 2016, pp. 779–788, doi: 10.1109/CVPR.2016.91.
2. W. Liu et al., “SSD: Single Shot MultiBox Detector,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Amsterdam, The Netherlands, 2016, pp. 21–37, doi: 10.1007/978-3-319-46448-0_2.
3. A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, “YOLOv4: Optimal Speed and Accuracy of Object Detection,” arXiv:2004.10934, 2020.
4. G. Jocher et al., “YOLOv5,” Ultralytics, GitHub repository, 2023. [Online]. Available: https://github.com/ultralytics/yolov5
5. L. Zhou et al., “YOLOv8-Based Drone Detection: Performance Analysis,” Applied Sciences, vol. 15, no. 2, Art. no. 723, 2025, doi: 10.3390/app15020723.
6. N. Carion et al., “End-to-End Object Detection with Transformers,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Glasgow, U.K., 2020, pp. 213–229.
7. X. Zhu et al., “Deformable DETR: Deformable Transformers for End-to-End Object Detection,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2021.
8. A. Dosovitskiy et al., “An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2021.
9. H. Zhang et al., “AI-Based Object Recognition Under Adverse Battlefield Conditions,” IEEE Access, vol. 12, pp. 21543–21558, 2024, doi: 10.1109/ACCESS.2024.10431245.
10. X. Li et al., “Edge AI for Real-Time Target Recognition in UAV Systems,” Sensors, vol. 25, no. 2, Art. no. 356, 2025, doi: 10.3390/s25020356.

11. M. Khan et al., “Autonomous Turrets Using YOLO for Target Identification,” Defence Technology, vol. 19, 2023, doi: 10.1016/j.dt.2023.11.004.
12. Y. Wang et al., “Adversarial Robustness of Object Detection Models in Military AI Systems,”
13. IEEE Trans. Neural Netw. Learn. Syst., early access, 2024, doi: 10.1109/TNNLS.2024.10345689.
14. I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and Harnessing Adversarial Examples,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2015.
15. T.-Y. Lin et al., “Microsoft COCO: Common Objects in Context,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2014.
16. G.-S. Xia et al., “DOTA: A Large-Scale Dataset for Object Detection in Aerial Images,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018.
17. R. Singh et al., “Ethical and Operational Challenges of AI-Driven Autonomous Weapon Systems,” AI & Society, vol. 40, no. 1, pp. 1–15, 2025, doi: 10.1007/s00146-025-01890-3.
18. B. Mittelstadt et al., “The Ethics of Algorithms: Mapping the Debate,” Big Data & Society, vol. 3, no. 2, 2016.
19. A. Jain and P. Bhutani, “A Study of Context-aware Systems,” in Proc. Int. Conf. Innovative Computing & Communication (ICICC), New Delhi, India, Oct. 23, 2024.
20. N. Verma, P. Bhutani, R. Lalit, and S. Venugopal, “Map Reduce Framework-Assisted Feature Analysis and Adaptive Multiplicative Bi-RNN Using Big Data Analytics for Decision-Making,” International Journal of Computational Intelligence Systems, vol. 18, no. 1, pp. 1–30, 2025.
21. R. Lalit, P. Bhutani, N. Verma, and Y. Sharma, “CNN Based Methods for Crowd Counting – A Comprehensive Study,” IJCRT, vol. 11, no. 10, 2023.

Digital Object Identifier: https://doi.org/10.48165/iitmjit.2026.12.1.14

Large Language Models and Agentic AI: How Modern AI Systems Are Evolving Into Autonomous Agents

Deepti Sharan[1], Nehuti[2]
[1]Department of AI-ML: Business Application, McCombs School of Business, UT Austin
[2]Department of Computer Science, Bharati Vidyapeeth College of Engineering, New Delhi, India
[2]nehutigoel@gmail.com

Abstract-Large Language Models (LLMs) have undergone a dramatic transformation from static text-completion engines to dynamic, tool-wielding autonomous agents capable of planning, reasoning, and executing multi-step tasks in real-world environments. This paper provides a comprehensive survey of the architectural principles, training paradigms, and emergent capabilities that underpin modern LLMs such as GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro. We examine the progression from transformer-based language models toward agentic AI systems that integrate memory modules, external tools, and multi-agent coordination frameworks. Key topics include the Retrieval- Augmented Generation (RAG) paradigm, chain-of- thought and tree-of-thought prompting strategies, the ReAct framework, and multi-agent orchestration platforms such as AutoGen and CrewAI. We further present a structured experimental comparison of Chain-of-Thought (CoT) and ReAct prompting paradigms across 120 standardized tasks spanning multi-step question answering, logical inference, and knowledge-retrieval scenarios, provide in reproducible methodology and statistical analysis. We also discuss open challenges in safety, alignment, hallucination mitigation, and computational costs associated with large-scale deployment. Our analysis reveals that while agentic AI demonstrates remarkable potential in software engineering, scientific research, and enterprise automation, significant hurdles in reliability, explainability, and ethical governance must be addressed before wide- scale deployment can be responsibly achieved.

Keywords- Large Language Models, Agentic AI, Autonomous Agents, Transformer Architecture, Chain- of-Thought, Retrieval-Augmented Generation, Multi- Agent Systems, AI Safety.

1. Introduction

Artificial intelligence has long pursued the goal of creating systems capable of autonomous reasoning and purposeful action. For much of its history this goal remained elusive, constrained by symbolic AI systems that required rigid hand-crafted rules and struggled to generalize beyond narrow domains. The advent of deep learning, and subsequently large-scale language models, fundamentally disrupted this trajectory. Beginning with the introduction of the Transformer architecture by Vaswani et al. in 2017, and accelerating through the GPT series, BERT, and T5, language models demonstrated that general-purpose intelligence could emerge from pretraining on massive corpora of text.

The release of GPT-3 in 2020, with its 175 billion parameters and remarkable few-shot learning abilities, marked a watershed moment. It became evident that scaling model size and training data could unlock capabilities—arithmetic, translation, code generation, logical inference—that had never been explicitly taught. The subsequent development of instruction-tuned models such as InstructGPT, ChatGPT, and their successors further closed the gap between static language understanding and practical, user-directed task completion.

Yet the most profound recent development is not merely the growth in model scale or benchmark performance; it is the emergence of agentic AI—systems that do not merely respond to single-turn queries but instead plan sequences of actions, invoke external tools, maintain memory across interactions, and adapt their behaviour in response to environmental feedback. Systems such as AutoGPT, BabyAGI, LangChain agents, and Microsoft's AutoGen have demonstrated that LLMs can serve as the cognitive core of autonomous pipelines capable of browsing the web, writing and executing code, managing files, and coordinating with other AI agents.

This paper surveys this rapidly evolving landscape comprehensively. We trace the architectural foundations of modern LLMs, characterise the key components that transform them into agentic systems, analyse the dominant frameworks and techniques, and discuss the critical challenges of safety, alignment, and governance that must be addressed as these systems grow more capable and widely deployed. To validate the practical implications of these approaches, we conduct a controlled experimental analysis across 120 standardised tasks, the details of which are presented in Section IX.

1.1 Scope and Contributions

This survey makes the following primary contributions: (1) a unified architectural view of LLMs as the cognitive backbone of agentic AI systems; (2) a taxonomic analysis of agent frameworks, memory architectures, and planning

paradigms; (3) a comparative review of leading frontier models and their agentic capabilities; (4) a structured discussion of open challenges and future research directions; and (5) a reproducible experimental analysis comparing Chain-ofThought and ReAct reasoning paradigms across 120 tasks drawn from established benchmarks, with statistical significance testing.

2. Background: from language models to LLMs

2.1 The Transformer Architecture

The Transformer architecture introduced the selfattention mechanism as the foundational operation for sequence modelling, replacing recurrent networks that struggled to capture long-range dependencies. In a Transformer, each token in a sequence attends to all other tokens simultaneously, weighted by learned compatibility scores. Formally, the attention output is computed as: $\text{Attention}(Q, K, V) = \text{softmax}(QK^T / \sqrt{dk}) \cdot V$, where Q, K, and V are query, key, and value matrices derived from linear projections of the input, and dk is the dimensionality of the keys. Multi-head attention applies this operation in parallel across multiple learned subspaces, enabling the model to capture diverse relational patterns simultaneously. The decoder-only variant, used by GPT-class models, additionally applies causal masking to ensure autoregressive generation.

2.2 Scaling Laws and Emergent Abilities

Kaplan et al. (2020) demonstrated that model performance on language modelling benchmarks scales predictably as a power law with respect to model size, training compute, and dataset volume. Hoffmann et al. (2022) revised these scaling laws in the Chinchilla paper, arguing that prior large models were significantly undertrained relative to their parameter count and that optimal performance requires approximately 20 tokens of training data per parameter. Beyond smooth scaling, Wei et al. (2022) identified a class of emergent abilities —capabilities that appear abruptly at sufficient scale and are not predictable by interpolating from smaller models —including multi-step arithmetic, chain-of-thought reasoning, and zero-shot instruction following.

2.3 Instruction Tuning and Reinforcement Learning from Human Feedback

Raw pretrained LLMs are powerful statistical models but are not inherently aligned with user intentions. Instruction tuning—fine tuning on curated datasets of (instruction, desired output) pairs— substantially improves model usability. The seminal RLHF framework (Ziegler et al., 2019; Ouyang et al., 2022) trains a reward model on human preference rankings of model outputs and then uses Proximal Policy Optimisation (PPO) to fine-tune the LLM to maximise this reward, dramatically improving instruction following, helpfulness, and safety. Constitutional AI (Anthropic, 2022) extends this paradigm by replacing human preference labels with AI-generated critiques guided by a set of principles, reducing the bottleneck of expensive human annotation.

3. Frontier LLMs Architectures and Capabilities

The current generation of frontier LLMs represents a convergence of massive scale, sophisticated fine tuning, multimodal input processing, and purpose-built agentic tooling. Table I provides a comparative overview of leading systems along dimensions relevant to agentic deployment.

Table 14.1: Comparative overview of leading frontier LLMS

Model	Org	Params	Context	Agentic Features
GPT-4o	Open AI	~200 B (est.)	128 K	Tool use, vision, code execution, function calling
Claude 3.5 Sonnet	Anthr opic	Undis closed	200 K	Computer use, extended context, tool use
Gemini 1.5 Pro	Googl e	Undis closed	1M	Multimodal, long-document reasoning, grounding
LLaMA 3 70B	Meta	70B	8K	Open weights, fine-tuning, community tooling
Mistral Large	Mistra l AI	~45B (est.)	32K	Function calling, API access, instruction following

GPT-4o, released by OpenAI in 2024, supports native multimodal inputs including text, images, and audio in a single unified model. Its 128K token context window, combined with robust function-calling APIs, makes it well-suited for agentic pipelines. Claude 3.5 Sonnet from Anthropic extends the context window to 200K tokens and introduces a

'computer use' capability enabling the model to interact with GUI applications directly—clicking buttons, filling forms, and navigating desktop interfaces.

Google's Gemini 1.5 Pro achieves a context length of up to one million tokens through Mixture-of Experts architecture and efficient attention mechanisms. This enables reasoning over entire codebases, lengthy research papers, or hours of video content in a single pass. Meta's LLaMA 3, while more modest in its context window, provides open-weight models that have catalysed the open-source ecosystem and enabled extensive academic and industrial research into fine tuning, quantisation, and agent adaptation.

3.1 Multimodal Capabilities

Modern agentic systems increasingly require perception beyond text. Vision-language models such as GPT-4V and Gemini integrate image and video understanding directly into the language model backbone, enabling agents to interpret screenshots, diagrams, charts, and real-world photographs. This multimodal perception is critical for agents operating in graphical user interface environments or processing scientific data with visual components.

3.2 Code Generation and Execution

Code generation represents one of the most practically significant capabilities of LLMs in agentic contexts. Models fine-tuned on large programming corpora—including GitHub Codex, StarCoder, and the coding-focused variants of frontier models—can write syntactically and semantically correct programmes across dozens of programming languages. More importantly, agentic systems close the loop by executing generated code in sandboxed environments, observing outputs and error messages, and iteratively refining the programme until the desired outcome is achieved.

4. Agentic AI: Principles and Architecture

The transformation of an LLM from a query response system into an autonomous agent requires the integration of several architectural components beyond the base language model. These address planning, memory, tool use, and execution feedback.

4.1 Planning and Reasoning

Planning is the process by which an agent decomposes a high-level goal into a sequence of actionable subtasks. Early agentic systems relied on simple chain-of-thought (CoT) prompting—instructing the model to reason step-by-step in natural language before producing a final answer. Wei et al. (2022) demonstrated that CoT prompting substantially improves performance on arithmetic, commonsense, and symbolic reasoning tasks, particularly in models with more than ~100B parameters.

Tree-of-Thought (ToT) prompting, introduced by Yao et al. (2023), extends CoT by enabling the model to explore multiple reasoning paths simultaneously, evaluating partial solutions at each node and pruning unpromising branches. This approach is inspired by classical search algorithms such as BFS and DFS and yields significant improvements on tasks requiring systematic exploration such as mathematical proofs and crossword solving.

The ReAct framework (Yao et al., 2022) interleaves reasoning traces with action invocations in a tight loop: the model generates a thought, takes an action (e.g., web search), receives an observation, generates another thought, and so on. This alternation between internal reasoning and external action grounding substantially reduces hallucination by anchoring the model's beliefs in real observations.

4.2 Memory Architectures

Human-like autonomous behaviour requires memory at multiple timescales. In agentic AI systems, memory is classified into four categories. Working memory corresponds to the model's active context window—the finite token sequence it can process in a single forward pass. While modern frontier models have extended this substantially (up to 1M tokens for Gemini 1.5 Pro), it remains a finite and expensive resource.

Long-term memory is achieved through external storage mechanisms, most commonly vector databases such as Pinecone, Weaviate, or Chroma. Documents and past interactions are chunked, embedded into dense vector representations, and indexed for approximate nearest-neighbour retrieval. When the agent encounters a relevant query, semantically similar memories are retrieved and injected into the context window—a paradigm known as Retrieval-Augmented Generation (RAG). Episodic memory captures structured logs of past agent actions and outcomes, enabling learning from experience within and across sessions. Semantic memory encodes factual world knowledge from pretraining or curated knowledge bases.

4.3 Tool Use and Environment Interaction

A defining characteristic of agentic systems is the ability to call external tools. Through function-calling APIs standardised by OpenAI and adopted across the industry, agents can invoke web search engines, execute code interpreters, query databases, call REST APIs, send emails, and interact with operating system interfaces. The model generates a structured function call specification—including function name and typed arguments—which the execution environment runs, returning results for the agent to process.

More advanced tool use includes browser automation via frameworks like Playwright and Selenium, enabling agents to navigate websites, fill forms, and extract structured data from dynamic pages. The 'computer use' capability in Claude 3.5 Sonnet represents a further step: the model can process screenshots of arbitrary desktop applications and generate precise mouse and keyboard actions to accomplish user-specified goals without requiring a programmatic API.

5. Agentic Frameworks and Multi-Agent Systems

The rapid growth of agentic AI has been accompanied by a proliferation of open-source and commercial frameworks designed to simplify the construction of agent pipelines. Table II summarises the most widely adopted frameworks along with their architectural paradigms and primary use cases.

Table 14.2: Comparison of major Agentic AI Frameworks

Framewo rk	Paradigm	Memory Type	Primary Use Case
LangChain	Chain-ofThought	Buffer + Vector	Generalist pipelines, RAG, document Q&A
AutoGPT	Autonomou s loop	File-based persistent	Long-horizon autonomous tasks, goal pursuit
ReAct (Google)	Reason+Act interleave	Ephemeral	Tool-augmented question answering and planning
AutoGen (Microsoft)	Multi-agent conversatio n	Shared context	Collaborative code generation and review
CrewAI	Role-based crews	Shared + individual	Workflow automation, research, content creation

5.1 LangChain and LangGraph

LangChain is the most widely adopted framework for constructing LLM-powered applications. It provides modular abstractions for prompt templates, output parsers, memory backends, tool integrations, and chain compositions. LangChain Expression Language (LCEL) allows developers to compose these components declaratively. LangGraph extends this to stateful, graph structured workflows, enabling cyclic execution patterns necessary for iterative agent loops—a significant limitation of LangChain's original linear chain abstraction.

5.2 AutoGen and Multi-Agent Coordination

Microsoft's AutoGen framework models agentic tasks as conversations between multiple specialised AI agents. A User Proxy Agent mediates between the human user and the AI pipeline, while one or more Assistant Agents handle specific subtasks such as code generation, critique, and execution. This conversational multi-agent paradigm has demonstrated strong results on complex coding and data analysis tasks, with agents iteratively refining outputs through structured debate. AutoGen's group chat manager generalises this to Nagent conversations with flexible speaker selection policies.

5.3 CrewAI and Role-Based Agents

CrewAI adopts a metaphor of human organisational structures, instantiating agents as crew members with defined roles, goals, and backstories. A researcher agent, a writer agent, and an editor agent can collaborate on producing a research report, with each agent's output serving as context for the next. This role-based decomposition exploits the LLM's ability to adopt personas and adapt communication style to context, resulting in more coherent collaborative outputs.

5.4 Retrieval-Augmented Generation

RAG has emerged as a critical technique for grounding LLM outputs in verifiable external knowledge, reducing hallucination, and keeping agents current beyond their training cutoff. In a standard RAG pipeline, a retriever component—typically a bi-encoder embedding model such as E5 or BGE—encodes queries and documents into a shared vector space. At inference time, the top-k most semantically similar document chunks are retrieved and prepended to the model's input prompt as additional context. Advanced variants such as HyDE (Hypothetical Document Embeddings), Contextual Compression RAG, and Graph RAG (Microsoft, 2024) further improve retrieval precision and reasoning over structured knowledge.

6. Applications of Agentic AI

6.1 Software Engineering Agents

Software engineering represents perhaps the most mature domain for agentic AI deployment. Devin, developed by Cognition AI, was demonstrated in 2024 as an AI software engineer capable of interpreting natural language task specifications, writing multi-file codebases, running unit tests, and debugging failures— all within an integrated development environment. GitHub Copilot Workspace similarly enables developers to describe a desired feature in natural language and receive a complete pull request with code changes across multiple files.

6.2 Scientific Research Automation

Agentic AI is beginning to transform the research pipeline. Systems have been demonstrated that can autonomously search scientific literature, synthesise findings from dozens of papers, generate hypotheses, design experimental protocols, write analysis code, interpret results, and draft manuscript sections. Sakana AI's AI Scientist system (2024) demonstrated end-to-end generation of novel research papers in machine learning by iterating through hypothesis generation, experimentation, and writing in a closed loop.

6.3 Enterprise Process Automation

In enterprise settings, agentic AI is being deployed for customer service orchestration, supply chain monitoring, financial report generation, and HR process automation. Platforms such as Microsoft Copilot 365 and Salesforce Einstein embed agentic capabilities directly into productivity tools, allowing users to delegate multi-step workflows—scheduling meetings, summarising email threads, drafting contracts—to AI agents that interact with APIs across the enterprise software stack.

7. Challenges and open problems

Table III provides a structured summary of the principal challenges facing agentic AI deployment, together with the mitigation approaches currently under investigation.

Table 14.3: Summary of key challenges in Agentic AI deployment

Challenge	Description	Mitigation Approaches
Hallucination	Confident generation of factually incorrect content	RAG, grounding, tool verification
Alignment	Agent pursues proxy goals or misinterprets user intent	RLHF, Constitutional AI, sandboxing
Compute Cost	Multiplicative API calls inflate latency and cost	Caching, smaller specialist models, batching
Context Mgmt.	Lost-in-the-middle degradation for long contexts	Compression, retrieval, hierarchical memory
Evaluation	No consensus benchmarks for long-horizon agent tasks	AgentBench, SWEbench, WebArena, GAIA
Misuse Risk	Autonomous tool use can enable malicious automation	Usage monitoring, rate limits, policy enforcement

7.1 Hallucination and Factual Reliability

Despite impressive capabilities, LLMs remain susceptible to hallucination—the confident generation of factually incorrect, fabricated, or internally inconsistent information. In agentic systems, hallucinations can propagate and amplify across multi-step pipelines, with early errors compounding into dramatically incorrect final outputs. While RAG and tool-augmented grounding mitigate this risk, they do not eliminate it: models can still misinterpret retrieved information, fail to retrieve relevant context, or generate plausible sounding but incorrect code that passes superficial inspection.

7.2 Safety, Alignment, and Misuse

As agents acquire the ability to take consequential real-world actions—executing code, sending emails, purchasing goods, modifying files—the alignment problem becomes acutely urgent. An agent that misinterprets a user's goal or pursues proxy objectives may take irreversible harmful actions. Anthropic's Constitutional AI and OpenAI's alignment research programmes address this through RLHF, adversarial redteaming, and sandboxed execution environments. However, specifying human values precisely enough to prevent all failure modes remains fundamentally unsolved. The potential for misuse is equally pressing: autonomous agents with access to web browsing, code execution, and API calls

represent powerful tools for malicious actors, enabling at-scale phishing campaigns, automated vulnerability exploitation, and disinformation generation.

7.3 Computational Cost and Latency

Agentic pipelines that invoke LLM calls iteratively —once per planning step, tool call, and reflection—incur computational costs that are multiplicative rather than linear with respect to task complexity. A moderately complex coding task might require dozens of LLM inference calls, translating to significant latency (often minutes rather than seconds) and substantial API costs. Techniques such as prompt caching, speculative decoding, and smaller specialised models for sub-tasks are being explored to reduce this overhead, but the fundamental tension between capability and efficiency remains.

7.4 Context Window Management and Evaluation

While context windows have grown substantially, effectively managing the information within them remains an open challenge. Recent research has demonstrated that LLMs exhibit a 'lost in the middle' phenomenon—performance degrades on information appearing in the middle of long contexts rather than at the beginning or end. Intelligent context compression, summarisation, and retrieval strategies are needed to address this. In parallel, the field lacks consensus on standardised evaluation protocols for agentic systems. Emerging benchmarks such as AgentBench, WebArena, SWE-bench, and GAIA are designed specifically for long-horizon agentic evaluation, but many published results are difficult to reproduce and compare across laboratories.

8. Future Directions

Several emerging research directions are poised to shape the next generation of agentic AI systems. World models—internal representations of environmental dynamics learned from diverse sensory data—may endow agents with more robust causal reasoning and counterfactual simulation, enabling better planning and reduced reliance on trial-and-error interaction with the real world.

Neuro-symbolic integration seeks to combine the pattern recognition strengths of neural language models with the precision and interpretability of formal symbolic reasoning systems. Hybrid systems that invoke theorem provers, logic engines, or knowledge graphs as specialised tools may achieve levels of reliability that pure neural approaches cannot. Parameter-efficient fine tuning methods such as LoRA, combined with selective memory consolidation mechanisms, may enable more dynamic, experience-driven adaptation while managing the catastrophic forgetting problem.

Finally, the governance and regulation of agentic AI systems is an urgent societal challenge. As agents acquire greater autonomy and impact, questions of accountability, liability, transparency, and human oversight become pressing legal and ethical concerns. Frameworks such as the EU AI Act and emerging guidelines from NIST represent early steps toward structured regulation, but the pace of technological development substantially outstrips the pace of regulatory adaptation. Multidisciplinary collaboration between computer scientists, ethicists, legal scholars, and policymakers will be essential to navigating this landscape responsibly.

9. Experimental Evaluation

9.1 Objective

To move beyond purely theoretical discussion and address the reproducibility concerns prevalent in agentic AI evaluation literature, we conduct a controlled experimental analysis comparing two dominant reasoning paradigms—Chain-of-Thought (CoT) prompting and the ReAct framework—across a standardised task battery. Our central hypothesis is that ReAct's integration of external tool use and iterative feedback will yield higher accuracy on knowledge intensive and multi-step tasks, while CoT will retain advantages in latency-sensitive scenarios.

9.2 Experimental Setup

Model: All experiments were conducted using GPT-4o (version gpt-4o-2024-08-06) accessed via the OpenAI API at a temperature of 0.0 to ensure deterministic outputs. The same model checkpoint was used for both conditions to isolate the effect of the prompting paradigm.

Task Battery: We curated a benchmark of 120 tasks drawn from three categories: (i) Multi-step Question Answering (40 tasks, sourced from HotpotQA requiring 2–4 reasoning hops); (ii) Logical Reasoning (40 tasks, drawn from the MATH dataset's algebra and arithmetic subsets); and (iii) Knowledge-Retrieval (40 tasks, drawn from TriviaQA requiring factual recall beyond the model's reliable training distribution, verified by checking against post-cutoff facts). Tasks were matched across categories for estimated difficulty using human rater scores.

CoT Condition: The model was prompted with the standard few-shot CoT prefix: "Let's think step by step," followed by three in-context exemplars per task category. No external tools were provided.

ReAct Condition: The model was equipped with two tools: a web search API (Bing Search, top-3 results) and a Python code interpreter (sandboxed). The same three in-context exemplars were adapted to the Thought Action-Observation format per Yao et al. (2023). Tool calls were automatically executed and results returned to the model.

Evaluation Protocol: Each task was evaluated by two independent human raters for correctness (Cohen's $\kappa = 0.87$, indicating strong inter-rater agreement). Disagreements were resolved by majority vote with a third rater. Response time was measured as wall-clock time from API call submission to final token received, averaged over three runs.

9.3 Evaluation Metrics

Four primary metrics are reported: (1) Accuracy— proportion of tasks where the final answer was judged correct by both raters; (2) Average Steps—mean number of reasoning or action steps per task; (3) Mean Response Time—wall-clock seconds from request to final output, averaged over three repeated runs per task; (4) Robustness—proportion of tasks where the model recovered from at least one identifiable intermediate error to produce a correct final answer, assessed by rater annotation of intermediate reasoning traces.

9.4 Results

Table 14.4: Experimental results by task category (n = 40 per category)

Condition / Category	Accuracy (%)	Avg. Steps	Resp. Time (s)	Robustness (%)	n
CoT — Multi-step QA	67.5	2.8	4.2	34.1	40
CoT — Logical Reasoning	78.5	3.1	3.9	41.0	40
CoT — Knowledge-Retrieval	65.0	2.5	3.8	28.5	40
CoT — Overall	70.3	2.8	4.0	34.5	120
ReAct — Multistep QA	82.5	5.2	14.1	61.4	40
ReAct — Logical Reasoning	80.0	4.7	13.5	52.3	40
ReAct — Knowledge-Retrieval	87.5	5.8	15.3	74.6	40
ReAct — Overall	83.3	5.2	14.3	62.8	120

Table 14.5: Statistical significance (mcnemar's test, Overall accuracy)

Comparison	Accuracy Difference	χ^2 Statistic	p-value	Significant?
ReAct vs. CoT — Overall	+13.0 pp	11.42	0.0007	Yes ($\alpha = 0.01$)
ReAct vs. CoT — Multi-step QA	+15.0 pp	8.10	0.0044	Yes ($\alpha = 0.01$)
ReAct vs. CoT — Knowledge-Retrieval	+22.5 pp	14.70	0.0001	Yes ($\alpha = 0.01$)
ReAct vs. CoT — Logical Reasoning	+1.5 pp	0.18	0.6710	No

9.5 Discussions

The results reveal a nuanced picture. ReAct significantly outperforms CoT on multi-step question answering and knowledge-retrieval tasks, where access to external information and iterative error correction are decisive. The 22.5 percentage-point gap on Knowledge Retrieval tasks is particularly striking and confirms that tool-augmented grounding is critical when tasks depend on facts that may lie outside the model's reliable training distribution.

However, the two paradigms perform comparably on Logical Reasoning tasks ($p = 0.671$, not significant).

This is consistent with the hypothesis that formal reasoning tasks are primarily bottlenecked by the model's internal computation rather than external knowledge access. For such tasks, the overhead of ReAct's tool-calling loop—a 3.5x increase in mean response time—is not justified by accuracy gains.

The robustness metric shows an especially pronounced advantage for ReAct (62.8% vs. 34.5%), suggesting that the ability to observe and incorporate external feedback is a primary mechanism by which ReAct recovers from early errors. This finding supports the theoretical account of Yao et al. (2023) and has direct practical implications for the design of agentic pipelines where reliability is paramount.

9.6 Key Insights

1. Tool-augmented reasoning substantially improves accuracy on knowledge-intensive tasks but provides negligible benefit for self-contained logical reasoning.

2. Iterative observation-driven feedback is the primary mechanism underlying ReAct's robustness advantage; pipeline designers should prioritise tool integration for tasks with high knowledge uncertainty.

3. The latency cost of ReAct (~14s vs. ~4s mean response time) positions CoT as the preferable choice when throughput and cost are primary constraints and task types are predominantly logical rather than knowledge-retrieval.

9.7 Limitations

Several limitations constrain the scope of these findings. First, the study was conducted on a single model (GPT-4o); results may differ for smaller or open source models where CoT benefits are less pronounced below the ~100B parameter threshold identified by Wei et al. (2022). Second, the task battery comprises 120 tasks—sufficient for primary comparisons but insufficient for fine-grained sub-category analysis. Third, the web search tool was limited to top-3 results; broader retrieval configurations may further improve ReAct performance. Future work should extend this evaluation to Tree-of-Thought and Plan-and-Solve prompting strategies, employ larger task sets with stratified sampling, and assess performance across multiple model families including open-weight alternatives.

10. Conclusion

This paper has surveyed the architectural foundations, technical capabilities, and deployment challenges of large language models and agentic AI systems. We traced the progression from transformer based text predictors to autonomous agents capable of planning, memory-augmented reasoning, and multi-step tool use. The frontier models of 2024–2025—GPT-4o, Claude 3.5 Sonnet, Gemini 1.5 Pro—represent a qualitative leap in capability, and frameworks such as LangChain, AutoGen, and CrewAI are democratising access to agentic pipelines across industry and academia.

Yet the field stands at a critical juncture. The same capabilities that make agentic AI transformative— autonomy, tool access, goal-directed behaviour—also introduce novel risks around reliability, safety, and misuse. The research community must prioritise not only capability advancement but also robust evaluation, interpretability, alignment, and governance. Our controlled experimental analysis further supports that agentic reasoning frameworks, particularly those incorporating tool use and iterative feedback, achieve improved performance on multi-step tasks, albeit at the cost of increased computational overhead. The promise of agentic AI—systems that can serve as capable intellectual partners across scientific, creative, and practical domains—is within reach, but realising it responsibly demands concerted effort across technical, policy, and ethical dimensions.

References

1. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017.
2. T. Brown, B. Mann, N. Ryder et al., "Language models are few-shot learners," in NeurIPS, vol. 33, pp. 1877–1901, 2020.
3. J. Kaplan, S. McCandlish, T. Henighan et al., "Scaling laws for neural language models," arXiv preprint arXiv:2001.08361, 2020.
4. J. Hoffmann, S. Borgeaud, A. Mensch et al., "Training compute-optimal large language models," in NeurIPS, 2022.
5. J. Wei, X. Wang, D. Schuurmans et al., "Chain-ofthought prompting elicits reasoning in large language models," in NeurIPS, vol. 35, 2022.
6. J. Wei, Y. Tay, R. Bommasani et al., "Emergent abilities of large language models," Transactions on Machine Learning Research, 2022.
7. S. Yao, J. Zhao, D. Yu et al., "ReAct: Synergizing reasoning and acting in language models," in ICLR, 2023.
8. S. Yao, D. Yu, J. Zhao et al., "Tree of thoughts: Deliberate problem solving with large language models," in NeurIPS, 2023.
9. L. Ouyang, J. Wu, X. Jiang et al., "Training language models to follow instructions with human feedback," in NeurIPS, vol. 35, 2022.
10. Anthropic, "Constitutional AI: Harmlessness from AI feedback," arXiv preprint arXiv:2212.08073, 2022.

11. P. Lewis, E. Perez, A. Piktus et al., “Retrievalaugmented generation for knowledge-intensive NLP tasks,” in NeurIPS, vol. 33, pp. 9459–9474, 2020.
12. Q. Wu, G. Bansal, J. Zhang et al., “AutoGen: Enabling next-gen LLM applications via multia g e n t c o n v e r s a t i o n , ” a r X i v p r e p r i n t arXiv:2308.08155, 2023.
13. OpenAI, “GPT-4 technical report,” arXiv preprint arXiv:2303.08774, 2023.
14. Google DeepMind, “Gemini: A family of highly capable multimodal models,” arXiv preprint arXiv:2312.11805, 2023.
15. Meta AI, “Llama 3: Open foundation and fine-tuned chat models,” Meta AI Blog, 2024.
16. Lu, C. Lu, R. T. Q. Chen et al., “The AI Scientist: Towards fully automated open-ended scientific discovery,” arXiv preprint arXiv:2408.06292, 2024.
17. S. Jimenez, J. Yang, H. Wettig et al., “SWE-bench: Can language models resolve real-world GitHub issues?,” in ICLR, 2024.
18. T. Liu, X. Zhang, B. Guo et al., “AgentBench: Evaluating LLMs as agents,” in ICLR, 2024.
19. E. Hu, Y. Shen, P. Wallis et al., “LoRA: Low-rank adaptation of large language models,” in ICLR, 2022.
20. M. Yang et al., “HotpotQA: A dataset for diverse, explainable multi-hop question answering,” EMNLP, 2018.
21. Hendrycks et al., “Measuring mathematical problem solving with the MATH dataset,” NeurIPS, 2021.
22. M. Joshi et al., “TriviaQA: A reading comprehension dataset over trivia questions,” ACL, 2017.

IITM Journal of Information Technology

Paper Submission Guidelines

Submission of Paper is in Two Stages:

1. **Initial Paper Submission:** Prospective Author(s) is / are encouraged to submit their Manuscript including Charts, Tables, Figures and Appendixes in .pdf and .doc (both) Strictly using single column Springer format to itjournal@iitmjp.ac.in

All submitted articles must present original, previously unpublished research findings, whether experimental or theoretical. Articles should adhere to these criteria and must not be simultaneously under consideration for publication.

2. **Camera Ready Paper Submission:** After the completion of the review process, Author(s) are required to submit the camera-ready full-text paper in both .doc and .pdf formats upon paper acceptance.

***THERE IS NO PUBLICATION FEE**

Institute of Innovation in Technology and Management

Affiliated to GGSIPU, NAAC Grade 'A',
ISO 14001:2015, 17020:2012, 21001:2018 & 50001:2018 Certified
A Grade by GNCTD, A++ Grade by SFRC
D-27/28, Institutional Area, Janakpuri, New De1hi-110058
Tel: 011-28520890, 28520894
E-mail: director@iitmjp.ac.in Website: http://www.iitmjp.ac.in