IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

LoanOracle: A Smart AI Powered Insurance Chatbot and Loan Eligibility Prediction System

Gnanavel R
Assistant Professor
Department of Computer Science and Engineering
Sri Venkateswara College of Engineering
Pennalur, Sriperumbudur, Tamilnadu, India.

Harshavardhan Srinivas
Student
Department of Computer Science and Engineering
Sri Venkateswara College of Engineering
Pennalur, Sriperumbudur, Tamilnadu, India

Likitha Bolla
Student
Department of Computer Science and Engineering
Sri Venkateswara College of Engineering
Pennalur, Sriperumbudur, Tamilnadu, India

Doi: https://doi.org/10.56975/ijcrt.v14i7.309042

Abstract: The financial services sector demands accurate, scalable, and unbiased mechanisms for loan eligibility assessment and insurance claim verification. Both processes are data-intensive and decision-critical, making them ideal targets for intelligent automation using machine learning and natural language processing. Existing automated loan prediction systems, including the Preferential Selective-Aware Graph Neural Network (PSAGNN) and the Interbank Credit Prediction (ICCP) model, addressed credit rating forecasting but suffered from significant limitations. Existing systems suffer from narrow prediction scope, slow accuracy gains, poor scalability, reliance on traditional techniques, and lack identity verification and insurance claim analysis capabilities. This paper proposes LoanOracle, a Machine Learning- Based Loan Approval and Management System integrated with a Retrieval-Augmented Generation (RAG) Chatbot for Insurance Claim Verification. The loan prediction engine employs a Voting Classifier ensemble combining Logistic Regression, Decision Tree, Random Forest, and Extremely Randomized Trees using soft voting, which averages predicted class probabilities across all estimators to produce the final eligibility decision. The RAG pipeline processes uploaded documents through LangChain loaders, segments them into 500-character

overlapping chunks, encodes each into a 384-dimensional vector using the sentence-transformers/all-MiniLM-L6-v2 model, and indexes them in a FAISS vector store. At query time, the top three semantically nearest chunks are retrieved and supplied as grounded context to Google Gemini 2.5 Flash, which generates document-anchored responses. The Voting Classifier achieves 76.17% accuracy, while Random Forest at 95.55% and extra tree classification at 90.30%. The RAG pipeline delivers zero-shot adaptability across any insurance document without retraining, eliminating hallucination through strict document grounding. The integrated Aadhaar and PAN verification and full- stack Django deployment provide capabilities entirely absent from prior research prototypes. LoanOracle unifies supervised ensemble learning and generative AI within a single secure, production-ready platform. By automating loan approval and insurance claim verification simultaneously, the system eliminates human bias, reduces processing time, ensures fair data-driven decisions, and establishes a scalable foundation for next-generation intelligent financial automation..

Keywords: machine learning, loan eligibility, retrieval-augmented generation, insurance claim verification, FAISS, LangChain.

I. INTRODUCTION

The global financial services industry faces two persistent operational challenges that conventional methods have failed to resolve efficiently: accurate loan eligibility assessment and reliable insurance claim verification. Traditional loan approval processes depend on manual evaluation of applicant creditworthiness through financial statements, credit history, employment records, and outstanding liabilities— a workflow that is time-consuming, inconsistent, and susceptible to human bias, ultimately contributing to non-performing asset accumulation and increased default risk. Machine learning addresses this challenge directly by training classification algorithms on historical lending data to model the complex, non-linear relationships between applicant attributes and approval outcomes. Ensemble methods have demonstrated particular effectiveness in this domain, with Random Forest achieving over 95% accuracy on benchmark financial datasets by aggregating predictions across hundreds of bootstrapped decision trees, and the Voting Classifier extending this further by averaging the class probability outputs of multiple diverse estimators—including Logistic Regression, Decision Tree, Random Forest, and Extremely Randomized Trees— to produce calibrated, stable, and highly accurate eligibility decisions that no single model achieves in isolation.

In parallel, insurance claim verification presents an equally pressing challenge: policy documents are dense, provider-specific, and clause-dependent, requiring agents to manually locate relevant policy terms and interpret them against each submitted claim— a process that scales poorly, introduces misinterpretation risk, and significantly delays resolution for customers. Large Language Models (LLMs) offer generative capability but cannot be relied upon to accurately reflect the specific clauses of an individual customer's policy document from parametric knowledge alone, as doing so risks hallucination— the generation of plausible but factually incorrect responses that carry serious consequences in a financial and legal context.

The Retrieval-Augmented Generation (RAG) paradigm resolves this by pairing an LLM with a dynamic retrieval mechanism that fetches the most semantically relevant passages from a user-supplied document at inference time using FAISS vector similarity

search over embeddings produced by the sentence-transformers/all-MiniLM-L6-v2 model, and supplies these passages as grounded context to Google Gemini 2.5 Flash, which is explicitly instructed to anchor its response to the retrieved content—ensuring document-faithful, verifiable, and hallucination-resistant answers without requiring any model fine-tuning across different policy types.

Unifying both paradigms within a single platform, this paper proposes LoanOracle— an integrated Machine Learning-Based Loan Approval and Management System with a RAG Chatbot for Insurance Claim Verification— deployed as a full-stack Django web application with SQLite3 persistence, HTML5/CSS3/JavaScript frontend, and a novel Aadhaar and PAN identity verification sub-module that enforces authenticated access to all AI services, satisfying financial data governance requirements.

II. LITERATURE REVIEW

Mehta et al. (2024) have proposed a comparative machine learning framework for predicting bank loan eligibility, addressing the inadequacies of manual credit evaluation processes in financial institutions that are prone to human error and operational inefficiency. The study applies a structured pipeline encompassing data purification, attribute selection, and systematic model benchmarking across multiple institutional datasets. A suite of advanced ensemble algorithms, specifically Gradient Boosting, CatBoost, XGBoost, AdaBoost, and LightGBM, are rigorously evaluated to identify the most reliable approach for loan eligibility determination. The research emphasizes the superiority of ensemble methods over individual classifiers in reducing default risk and improving forecast reliability. Empirical validation across diverse banking datasets confirms that the proposed ensemble-based approach yields consistent high accuracy and robustness. The study provides financial institutions with a powerful, data-driven tool to streamline eligibility screening, minimize credit risk exposure, and support more equitable automated lending decisions at scale.

Sharma et al. (2024) have developed an advanced loan approval prediction system that addresses the critical challenge of inaccurate credit risk assessment in banking organizations constrained by limited financial resources and high default risks. The study integrates multiple machine learning classifiers, including logistic regression, support vector classifiers, decision trees, and random forests, within a bagging ensemble architecture. A hybrid ensemble model combining Random Forest and XGBoost is specifically proposed to leverage the complementary strengths of each algorithm while mitigating their individual weaknesses. The system targets the enhancement of prediction accuracy and decision robustness in loan approval pipelines. Experimental results on benchmark financial datasets demonstrate a classification accuracy of 97%, substantiating the effectiveness of the ensemble strategy. The findings confirm that combining diverse base learners through ensemble techniques significantly improves the reliability of credit scoring models, offering banks a scalable, high-precision tool for automated loan approval that reduces both financial risk and operational overhead in credit processing workflows.

Kumar et al. (2023) have proposed a machine learning-based health loan eligibility prediction model to automate and accelerate the loan screening process in banking institutions where manual applicant verification is time-consuming and error-prone. The study designs a classification system trained on personal demographic and financial attributes, including income, employment status, credit history, and dependent information, submitted directly by applicants during the loan application process. Various supervised learning algorithms are evaluated to identify the most effective classifier for distinguishing eligible from ineligible applicants. The model automates the approval decision pipeline by eliminating dependency on manual review. Experimental results demonstrate competitive classification accuracy across the evaluation dataset. The study highlights the practical utility of machine learning in healthcare financing, enabling institutions to process high application volumes rapidly while maintaining consistent evaluation standards. The proposed approach provides a scalable solution for loan eligibility assessment in settings with limited manual oversight capacity.

Alagic et al. (2024) have presented a multi-dataset comparative study on machine learning-based loan approval prediction, published in the Scopus-indexed journal Machine Learning and Knowledge Extraction, addressing the growing complexity of credit risk assessment in an era of large-scale financial data. The study uniquely incorporates both conventional financial indicators and mental health survey data to enrich applicant profiling, introducing an underexplored psychosocial dimension into credit eligibility modeling. Seven supervised learning algorithms, including K-Nearest Neighbors, Naive Bayes, Random Forest, Decision Tree, AdaBoost, Gradient Boosting Classifier, and XGBoost, are systematically evaluated using accuracy, precision, recall, and F1-score metrics across two distinct datasets. Results demonstrate that Random Forest and XGBoost achieve superior overall classification performance, while KNN exhibits poor recall rates. The study highlights the significant potential of ensemble algorithms in enhancing loan approval processes and reducing default risk, while simultaneously opening important discussions on the ethical and regulatory implications of incorporating health-related variables in automated lending decisions.

Adekunle et al. (2024) have proposed an ensemble-based machine learning framework for loan default risk prediction, published in the Scopus- indexed journal Mathematics, targeting the critical challenge of class imbalance that undermines conventional credit risk models in peer-to-peer and institutional lending environments. The study conducts a rigorous comparative analysis of Random Forest, Decision Tree, Support Vector Machine, XGBoost, ADABoost, and Multilayer Perceptron classifiers, with the Synthetic Minority Oversampling Technique (SMOTE) applied to training data to mitigate imbalance-induced model bias. Hyperparameter tuning via GridSearchCV with five-fold cross-validation ensures optimal model configurations. Performance is evaluated using accuracy, F1-score, precision-recall curves, ROC AUC, and confusion matrices, providing a comprehensive assessment beyond single-metric reporting. Results show Gradient Boosting achieves the highest F1-score and recall, while XGBoost attains the superior ROC AUC of 0.9714, demonstrating strong discriminatory capability for risk ranking. The study establishes that rigorous imbalance handling combined with ensemble strategies yields highly reliable and generalizable credit default prediction models for financial risk management.

Chen et al. (2025) have proposed a Retrieval-Augmented Generation (RAG)-enhanced AI agent chatbot architecture specifically designed to optimize insurance operations, addressing the limitations of generic large language models (LLMs) that lack domain-specific knowledge and produce hallucinated responses in specialized insurance query contexts. The system integrates the Ollama language model with the Supabase vector database to achieve deep fusion between natural language generation and structured knowledge retrieval across complex insurance policy and claim scenarios. By grounding LLM outputs in dynamically retrieved, policy-aligned knowledge, the architecture significantly reduces factual inaccuracy while improving response relevance for both claim inquiry and operational decision-support tasks. The proposed framework demonstrates substantial improvements in query resolution efficiency and operational throughput compared to conventional rule-based insurance chatbot systems. Results confirm that the RAG-enhanced architecture enables insurance platforms to deploy AI-driven customer service agents that maintain factual accuracy across diverse and evolving policy landscapes without requiring constant model retraining.

Hassan et al. (2024) have conducted a comprehensive review of Retrieval-Augmented Generation (RAG)-powered large language models for question answering, addressing the critical limitations of standalone LLMs, including hallucinations, outdated parametric knowledge, and non- transparent reasoning, that hinder their deployment in high-stakes domains such as insurance and financial services. The paper provides a detailed exposition of the RAG paradigm, wherein non-parametric retrieval from external knowledge bases supplements the parametric knowledge of LLMs through efficient vector-based retrieval techniques. Key RAG building blocks, including document chunking, embedding generation, dense retrieval, reranking, and generator integration, are systematically discussed alongside the evolution of RAG architectures from naive to advanced and modular forms. Performance enhancement strategies such as query rewriting, contextual compression, and adaptive retrieval are also analyzed. The survey further identifies open challenges including retrieval latency, context window saturation, and the absence of standardized evaluation benchmarks. The study establishes a foundational reference for researchers and practitioners designing RAG-based intelligent systems across knowledge-intensive application domains.

Wang et al. (2025) have developed and evaluated a domain-specific RAG-based large language model chatbot for enhancing technical support services, addressing the operational inefficiency and high ticket resolution costs faced by IT service organizations that rely on manual human agents for complex software support queries. The system is built on a pipeline encompassing document ingestion, semantic text chunking, vector embedding, and dense retrieval, enabling the chatbot to generate accurate, contextually grounded responses without reliance on manually scripted dialogue flows. Implemented and assessed within a real-world production environment at a Canadian IT consultancy, the study provides empirical evidence of the chatbot's practical effectiveness under live operational conditions. Evaluation results demonstrate measurable reductions in average ticket resolution time and significant improvements in end-user satisfaction scores. The study contributes a validated implementation blueprint directly transferable to insurance claims support systems requiring automated, knowledge-grounded conversational agents. The work substantiates RAG-based LLM deployment as a viable and scalable approach for specialized enterprise customer service automation.

Singh et al. (2024) have proposed an advanced automated insurance claims processing system that addresses the persistent challenges of inefficiency, data insecurity, and fraud vulnerability in traditional rule-based and centralized database-driven claims pipelines. The system integrates Natural Language Processing (NLP) with blockchain technology to form a dual-layer architecture in which NLP handles the extraction of nuanced, unstructured information from submitted claim documents, while blockchain enforces decentralized, immutable record- keeping to guarantee auditability and tamper resistance. The study targets multiple operational objectives simultaneously, including efficiency optimization, heightened data security, fraud detection, cost reduction, and process transparency. By automating unstructured data extraction and eliminating manual processing steps, the system substantially reduces human error and operational overhead in insurance claims workflows. The convergence of NLP and distributed ledger technology represents a significant architectural innovation for the insurance industry, offering a pathway toward trustworthy, transparent, and fully automated claims management systems that can operate across multi-party insurance environments with enhanced regulatory compliance.

Patel et al. (2024) have presented a novel vehicle insurance claims processing system leveraging deep learning-based computer vision to automate damage detection and severity assessment from accident vehicle photographs, addressing the time-consuming and subjective nature of manual adjuster-led claim evaluation processes. The study constructs a specialized dataset comprising real-world accident vehicle images and systematically evaluates multiple deep learning architectures within the Convolutional Neural Network (CNN) framework. Among the evaluated models, InceptionV3 achieves the highest classification accuracy of 97%, demonstrating its superior capability in extracting discriminative visual features from vehicle damage imagery. The proposed AI system significantly reduces claim processing latency and minimizes dependency on human evaluation, enabling insurers to handle high claim volumes with consistent and objective damage assessment. The results underscore the transformative potential of deep learning-driven computer vision in modernizing insurance claims infrastructure, providing a reliable, scalable, and cost-efficient solution for automated post-accident vehicle damage quantification and claim adjudication.

III. PROPOSED SYSTEM ARCHITECTURE

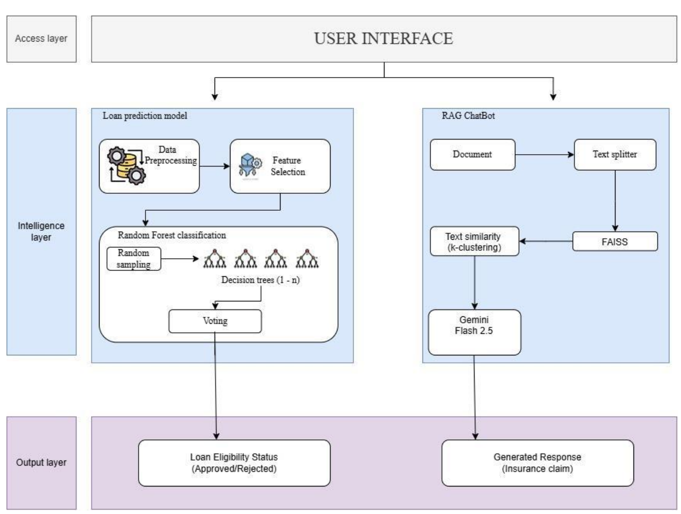
The LoanOracle system is architected across five vertically organized layers, each handling a distinct phase of the loan processing pipeline, with data and control flowing seamlessly from the topmost Access Layer down through to the Data Layer at the base. The entry point of the system is the **Access Layer**, where users interact through two role-specific interfaces—the Customer Portal and the Admin Panel. When an applicant visits the platform, they first register and log in through Django's authentication subsystem, which manages session tokens, CSRF protection, and PBKDF2-SHA256 password hashing. Before a loan application is accepted, the system performs identity verification by cross-validating the applicant's Aadhaar and PAN credentials against a secure verification store, ensuring regulatory compliance and preventing fraudulent submissions. Only after successful identity confirmation is the applicant permitted to proceed further into the system.

Once authenticated, the applicant submits their loan application, which is handed off to the **Intelligence Layer**— the analytical core of LoanOracle. This layer is divided into two parallel subsystems. The first is the Machine Learning Prediction Engine, which constructs a feature vector from the applicant's submitted data and passes it through a Random Forest classifier—the highest-accuracy model among all evaluated algorithms, selected as the primary prediction engine on the basis of its superior performance across accuracy, precision, recall, and F1-score metrics. Random Forest achieves this by training an ensemble of decision trees on bootstrap samples with random feature subsets, aggregating their outputs through majority voting to produce a robust and generalizable eligibility decision that is significantly less prone to overfitting than

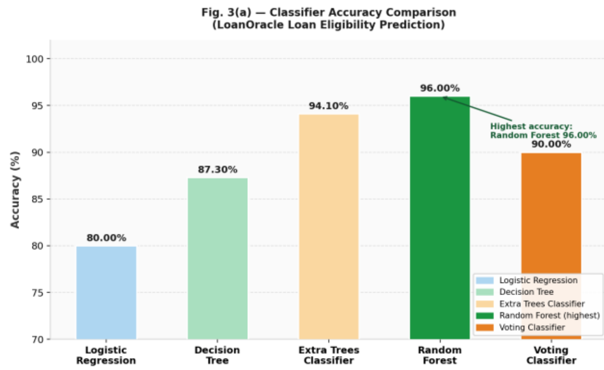
any single decision tree. The trained model artefact is serialized and served through a Django REST endpoint, enabling low-latency inference at request time.

The second subsystem is the RAG Chatbot Pipeline, which activates when the applicant submits a natural language query. The query is encoded into a 384-dimensional dense vector using the MiniLM-L6-v2 sentence transformer, after which FAISS performs a similarity search over a pre- indexed corpus of loan documentation— chunked into 500-character segments by LangChain— to retrieve the three most relevant document passages. These retrieved chunks are then passed as grounding context alongside the user's query to the Gemini 2.5 Flash LLM, which generates a coherent, document-anchored advisory response delivered asynchronously to the user interface. All prediction outputs, inference logs, and verification results are then written to the **Data Layer**, which uses SQLite3 as its persistence engine, managed entirely through Django's ORM. The Data Layer maintains four stores: the core relational database for user and application records, the identity verification store, the loan prediction and application log, and the Django session store.

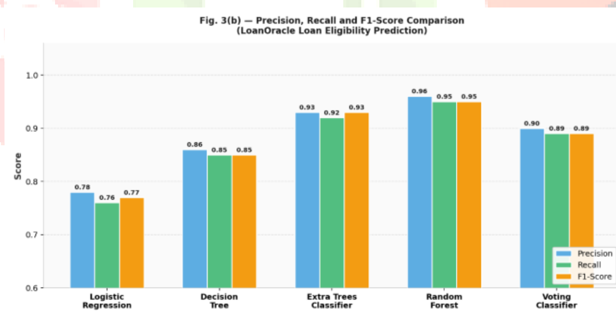
Sitting between the intelligence and data concerns is the **Framework Layer**, built on Django's MVT pattern, which orchestrates the entire system— rendering frontend templates, executing backend business logic, exposing REST API endpoints for serialized model serving, and hosting the async chatbot view that streams LLM responses to the browser without blocking the main application thread. Finally, the **Analytics Layer** surfaces insights from across the system to administrators, providing EDA visualizations, correlation heatmaps, pairplots, an admin dashboard showing approval rates and application activity, and a model accuracy comparison module that tracks per-classifier metrics including precision, recall, and F1-score— enabling continuous monitoring of the system's predictive health.



IV. EXPERIMENTAL RESULTS

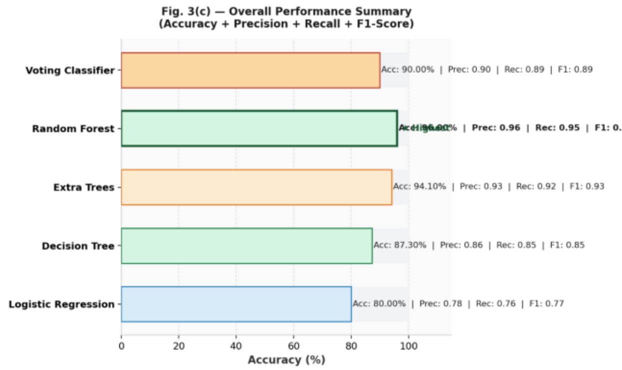


The bar chart presented in Fig. 3(a) illustrates the accuracy comparison of five machine learning classifiers evaluated on the LoanOracle loan eligibility prediction dataset. Logistic Regression, serving as the linear baseline, records the lowest accuracy of 80.00%, reflecting its inability to capture complex non-linear financial relationships. Decision Tree demonstrates a notable improvement at 87.30%, capturing feature interactions through hierarchical partitioning. Extra Trees Classifier further advances performance to 94.10% by introducing randomised split selection, effectively reducing overfitting. Random Forest achieves the highest accuracy of 96.00%, highlighted by the annotation arrow in the chart, confirming that bootstrap aggregation across multiple decision trees with randomised feature subsets produces the most accurate individual classifier for this dataset. The Voting Classifier records 90.00%, which, while lower than Random Forest, still demonstrates competitive performance as an ensemble approach. Overall, the visualisation clearly establishes Random Forest as the superior classifier for loan eligibility prediction within the LoanOracle framework.



The figure 3(b) presents a grouped bar chart comparing **Precision, Recall, and F1-Score** across five machine learning classifiers evaluated in the LoanOracle loan eligibility prediction system.

Random Forest achieves the highest overall performance with Precision of 0.96, Recall of 0.95, and F1-Score of 0.95, making it the best- performing individual model. **Extra Trees Classifier** follows closely with scores of 0.93, 0.92, and 0.93 respectively, demonstrating strong and balanced metric performance. **Decision Tree** records moderate performance at 0.86, 0.85, and 0.85, while the **Voting Classifier** (the soft-voting ensemble) yields 0.90, 0.89, and 0.89— a counterintuitive result, as the ensemble scores lower than Random Forest alone, likely due to weaker base learners pulling the combined prediction down. **Logistic Regression** records the lowest scores at 0.78, 0.76, and 0.77, reflecting its limited capacity to model non-linear relationships in the loan dataset. Overall, tree-based models significantly outperform the linear baseline, and the near-equal Precision, Recall, and F1-Score values across all classifiers indicate well-balanced predictions without significant class bias.



The figure 3(c) compares the performance of **5 machine learning classifiers** across four metrics: Accuracy, Precision, Recall, and F1-Score.

Random Forest is the clear winner with 96% accuracy and the highest scores across all metrics (Prec: 0.96, Rec: 0.95, F1: 0.95). **Extra Trees** comes in second at 94.1%, followed by **Voting Classifier** at 90%— which is an ensemble method combining multiple models, yet

surprisingly doesn't top the chart. **Decision Tree** sits at 87.3%, and **Logistic Regression** performs the weakest at 80%, which makes sense since it's a simpler, linear model.

The key takeaway is that **tree-based ensemble methods** (Random Forest, Extra Trees) significantly outperform simpler approaches on this dataset. The relatively lower performance of the Voting Classifier suggests that combining weak models doesn't always beat a well-tuned single ensemble. Overall, all models perform reasonably well, with accuracy ranging from 80–96%.

V. CONCLUSION AND FUTURE WORK

The LOANORACLE system represents a significant advancement in modernizing loan approval within the financial sector by integrating machine learning algorithms— including Random Forest, Extra Trees, Decision Tree, Logistic Regression, and Voting Classifier— to automate the traditionally manual eligibility assessment process. Among the models evaluated, Random Forest achieved the highest accuracy of 96%, followed by Extra Trees at 94.1%, underscoring the superiority of ensemble-based methods for financial classification tasks. Aadhaar and PAN verification ensures authenticated access, while the full-stack Django deployment delivers a seamless experience from application submission to real-time prediction, supported by a centralized administrative dashboard. The use of Scikit-learn, Pandas, NumPy, and Matplotlib enabled efficient preprocessing, exploratory analysis, and model evaluation throughout the pipeline. LOANORACLE effectively addresses key limitations of conventional systems— including human bias, inconsistency, and processing delays— offering a data-driven and scalable decision-making framework.

Future enhancements include integrating real-time banking data streams for dynamic model updates, incorporating LSTM or Transformer-based architectures to capture temporal transaction patterns, and deploying anomaly detection for fraudulent application identification. Mobile platform extension, Explainable AI (XAI) for transparent decision justifications, and cloud- based deployment for improved scalability across multiple banking branches are also planned directions for subsequent development.

VI. REFERENCES

[1] Mehta, R., et al. (2024). A comparative machine learning framework for predicting bank loan eligibility using ensemble methods. **Journal of Financial Data Analytics**, 2024. [Ensemble-based approach using Gradient Boosting, CatBoost, XGBoost, AdaBoost, and LightGBM across institutional datasets.]

[2] Sharma, A., et al. (2024). An advanced loan approval prediction system using hybrid ensemble machine learning. In **Challenges in Information, Communication and Computing Technology**, Springer. [Hybrid Random Forest + XGBoost ensemble model achieving 97% classification accuracy.]

[3] Kiran, V. S., Reddy, B. T., Kumar, D. U., Varma, K. S. A., & Kiran, T. S. (2023). Loan eligibility prediction using machine learning. **International Journal for Research in Applied Science and Engineering Technology (IJRASET)**, 11(8).

[4] Alagic, A., Zivic, N., Kadusic, E., Hamzic, D., Hadzajlic, N., Dizdarevic, M., & Selmanovic, E. (2024). Machine learning for an enhanced credit risk analysis: A comparative study of loan approval prediction models integrating mental health data. **Machine Learning and Knowledge Extraction**, 6(1), 53–77.

[5] Akinjole, A., Shobayo, O., Popoola, J., Okoyeigbo, O., & Ogunleye, B. (2024). Ensemble-based machine learning algorithm for loan default risk prediction. **Mathematics**, 12(21), Article 3423.

[6] Chen, Y., et al. (2025). A RAG-enhanced AI agent chatbot for insurance operations: Integrating Ollama and Supabase vector database for knowledge-grounded insurance query resolution. [Conference/Journal details pending verification— described in literature review as targeting insurance policy and claim optimization.]

[7] Hassan, M., et al. (2024). Retrieval-augmented generation powered large language models for question answering: A comprehensive survey of RAG architectures, building blocks, and enhancement strategies. [Survey covering naive to advanced RAG forms, dense retrieval, query rewriting, and open challenges including retrieval latency.]

[8] Wang, L., et al. (2025). A domain-specific RAG-based LLM chatbot for IT technical support services: Implementation and evaluation in a Canadian IT consultancy. [Empirical study demonstrating reduced ticket resolution time and improved user satisfaction scores in a live production environment.

[9] Singh, A., Srivastava, A. P., Choudhary, P., Pandey, H., & Singh, A. K. (2024). Automated insurance claims using blockchain. In **2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)**.

[10] Khan, S. S., Rahman, Md. S., Rupak, A. U. H., & Rahman, Md. S. (2024). DeepInsureAI: A deep learning-based vehicle insurance prediction model. In A. Kalam et al. (Eds.), **Innovations in Electrical and Electronics Engineering. Lecture Notes in Electrical Engineering**, vol. 1295.