International Journal of Advanced Multidisciplinary Research and Educational Development
Volume 2, Issue 1 | January – February 2026 | www.ijamred.com
ISSN: 3107-6513

Ensemble based Loan Approval Prediction System

Rohan Kumar Jha[#1], Rohit Kumar[*2], S Hiten[#3]

[#1]*Student, Department of Information Science & Engineering, CMR Institute of Technology, Bengaluru-560037, Karnataka, India*

[*2]*Student, Department of Information Science & Engineering, CMR Institute of Technology, Bengaluru-560037, Karnataka, India*

[#3]*Student, Department of Information Science & Engineering, CMR Institute of Technology, Bengaluru-560037, Karnataka, India*

[1]rojh22ise@cmrit.ac.in, [2]roku22ise@cmrit.ac.in, [3]shi22ise@cmrit.ac.in

***Abstract*—In the rapidly evolving financial sector, automating credit risk assessment is essential to improve efficiency and reduce default rates. Traditional rule-based systems often fail to handle complex financial data, resulting in delays and inaccurate decisions. This paper presents an Ensemble-Based Loan Approval Prediction System that leverages machine learning to deliver accurate, real-time, and transparent decision support. The study utilizes a dataset of 4,269 records with 12 financial attributes, including income, CIBIL score, and loan terms. A robust preprocessing pipeline incorporating Z-score–based outlier removal, categorical encoding, and a composite Assets feature was implemented. Seven supervised learning algorithms were evaluated, along with three ensemble techniques—Bagging, Random Forest, and AdaBoost—to enhance predictive stability. The Bagging Classifier achieved the best performance, with a testing accuracy and F1-score of 98.36%. To address the interpretability challenge of ensemble models, SHAP (SHapley Additive exPlanations) was integrated to quantify feature contributions, identifying the CIBIL score as the dominant factor in loan approval. The final model is deployed using a Streamlit web application, providing instant predictions and visual explanations for improved decision transparency.**

***Keywords*—Ensemble Model, Bagging Classifier, SHAP, Model Deployment, Feature Importance, Feature Engineering, Exploratory Data Analysis, Comparitive Analysis**

I. INTRODUCTION

The financial services sector is undergoing rapid digital transformation, fundamentally altering how institutions manage risk and interact with customers. Within this evolving environment, credit risk assessment remains a critical function, directly impacting both profitability and financial stability. The widespread adoption of fintech platforms and digital lending services has led to a sharp increase in online loan applications, significantly intensifying the demand for fast, accurate, and scalable decision-making systems. As a result, traditional manual and rule-based credit evaluation processes are increasingly inadequate.

Conventional loan approval systems rely heavily on manual document verification and static eligibility rules. While these methods have historically supported risk management, they suffer from limited scalability, slow processing times, and susceptibility to human error and subjective bias. Manual evaluations can take several days per application and often fail to capture the complex financial behaviors of modern borrowers, particularly individuals with non-traditional income sources or limited credit histories. Consequently, such rigid frameworks may reject creditworthy applicants or approve high-risk individuals, leading to inefficient lending outcomes. Machine Learning (ML) has emerged as a powerful alternative to traditional credit assessment techniques. By learning from historical data, ML models can identify complex, non-linear relationships among financial variables, enabling more accurate predictions of creditworthiness. Predictive analytics improves decision speed and reduces the likelihood of non-performing assets (NPAs). However, individual ML models such as Logistic Regression or standalone Decision Trees often struggle with high-dimensional, noisy financial data. These models may suffer from overfitting or underfitting, limiting their reliability when deployed in real-world environments.

To address these limitations, Ensemble Learning techniques have gained prominence. By combining multiple base classifiers, ensemble methods such as Bagging, Random Forest, and Boosting reduce variance and bias, resulting in more robust and stable predictions. This study investigates the effectiveness of ensemble-based approaches for loan approval prediction, with particular emphasis on Bagging-based classifiers to enhance accuracy and generalization. Despite improvements in predictive performance, accuracy alone is insufficient in the highly regulated financial domain. Advanced ML models are often criticized for their "black-box" nature, which hinders trust and regulatory compliance. Financial institutions, regulators, and applicants require transparent and interpretable decision-making processes. Therefore, modern credit risk systems must balance predictive power with explainability, ensuring that decisions can be clearly justified and understood.

The central objective of this research is to address this dual requirement of accuracy and transparency. The proposed system integrates ensemble learning techniques with Explainable Artificial Intelligence (XAI) mechanisms to provide both reliable predictions and meaningful explanations. By employing the SHAP (SHapley Additive exPlanations) framework, the model offers detailed insights into the influence of key financial features—such as credit scores, loan terms, and asset values—on approval decisions. Furthermore, the system is deployed through an interactive Streamlit-based web interface, enabling real-time, transparent, and user-friendly credit risk assessment.

II. LITERATURE SURVEY

The domain of automated credit risk assessment has witnessed a significant paradigm shift from traditional statistical methods to advanced machine learning (ML) architectures. Recent research has predominantly focused on enhancing predictive accuracy through algorithmic optimization, ensemble learning, and rigorous data preprocessing.

Optimization and Meta-Heuristics Chittimalla et al. (2024) explored the integration of meta-heuristic optimization techniques to refine loan approval mechanisms. Their research introduced a Tabu Search Optimization layer atop standard classifiers like Logistic Regression and Support Vector Machines (SVM). The study demonstrated that while traditional models like Random Forest achieved a respectable accuracy of 97%, the application of Tabu Search optimization pushed the performance to 98%. Their work highlights the necessity of minimizing risk analysis through sophisticated search algorithms, though it primarily focused on accuracy maximization rather than model interpretability.

Ensemble Learning and Inclusivity Building on the need for robust classification, Saha et al. (2025) presented a framework specifically designed for diverse applicant demographics using the same benchmark Kaggle dataset utilized in this study. Their comparative analysis of multiple algorithms revealed the superiority of ensemble methods over standalone classifiers. Specifically, their implementation of XGBoost and Ensembled Bagged Trees (EBT) achieved accuracies of 98.72% and 98.59% respectively, outperforming traditional Logistic Regression (91.3%). This study is particularly relevant as it validates the efficacy of bagging techniques in handling financial data, noting that ensemble architectures significantly reduce operational variance and enhance fairness in lending decisions for economically unprivileged groups.

Data Preprocessing and Feature Selection While algorithm selection is critical, Ahmadani et al. (2023) emphasized the foundational importance of data quality. Their research investigated the impact of preprocessing stages—specifically Z-score standardization—and feature selection methods like Information Gain and the Gini Index. The study concluded that applying these preprocessing steps significantly improved the performance of the Random Forest algorithm, elevating its accuracy to 97.1%. This underscores that effective creditworthiness prediction is dependent not just on the model architecture, but on the rigorous removal of noise and the selection of high-value financial indicators. Research Gap Despite the high predictive accuracies reported in these studies—ranging from 97% to 98.7%—a critical limitation persists: the "black-box" nature of these advanced models. Chittimalla et al. focused on optimization, and Saha et al. prioritized accuracy and inclusivity, yet neither fully addressed the regulatory requirement for local interpretability. Modern banking regulations increasingly demand to know why a specific applicant was rejected, not just the probability of rejection. This research aims to bridge this gap by integrating the Bagging Classifier (which matches the high accuracy of the cited studies) with the SHAP (SHapley Additive exPlanations) framework, thereby providing the granular, instance-level transparency that previous optimization-focused studies have overlooked.

III. PROPOSED METHODOLOGY

The comprehensive procedural workflow adopted for this study, visually delineated in Figure 1, or chest rates a sequential pipeline designed to transform raw financial data into actionable, explainable insights. The process commences with the acquisition of Data Customer & Loan Records, which serves as the foundational input. This raw data is subjected to rigorous Data Preprocessing to address inconsistencies and formatting requirements, setting the stage for Exploratory Data Analysis (EDA), where statistical trends and feature distributions are examined to inform model selection. The core computational phase involves the training of the Ensemble-Based Loan Classifier, specifically leveraging the Bagging Ensemble technique to maximize predictive stability and accuracy. To address the opacity often associated with such advanced models, the pipeline integrates Explanatory 3 AI–SHAP, a framework that computes feature importance to validate decision logic. The workflow culminates in the Streamlit based Web Application Interface Deployment, a user-facing layer that synthesizes these components to deliver the final Result—comprising the loan status, a confidence probability score, and a visual explanation—directly to the end-use.

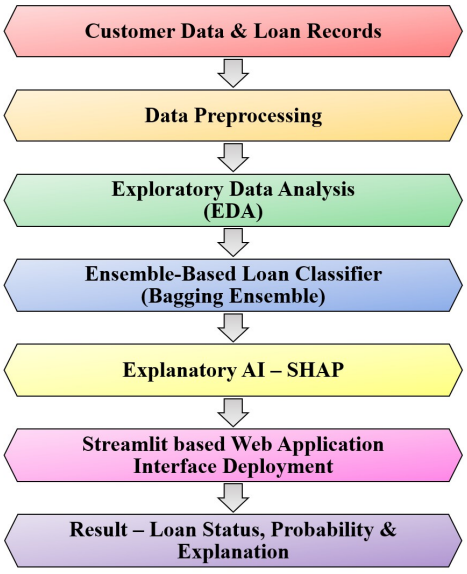

Fig 1. Flow Diagram

1. Data Collection and Understanding

The experimental framework utilizes a structured financial dataset obtained from the Kaggle repository, providing a standardized benchmark for predictive modelling. The dataset consists of 4,269 loan applicant records with 13 attributes representing key dimensions of credit risk, including demographic factors (education level, self-employment status, dependents), financial indicators (annual income, loan amount, loan term, CIBIL score), and collateral-related assets (residential, commercial, and luxury holdings). A

comprehensive data understanding phase was conducted to evaluate data quality and distribution. The target variable, loan status, is binary, where '1' indicates approval and '0' denotes rejection. Statistical analysis confirmed the absence of missing values, eliminating the need for data imputation. Class distribution analysis revealed approximately 62% approved and 38% rejected applications, indicating a reasonably balanced dataset suitable for supervised learning without requiring oversampling techniques.

2. Data preprocessing

Data preprocessing ensured dataset integrity through binary encoding of categorical variables such as education and self-employment status. Numerical features were standardized using a StandardScaler to normalize feature scales. Feature engineering was applied by aggregating asset-related attributes into a composite *Assets* variable to enhance predictive performance. The dataset was subsequently divided into an 80:20 training–testing split for robust model evaluation.

3. Exploratory data Analysis

Exploratory Data Analysis (EDA) was conducted to validate data integrity and examine the distribution and relationships of key features prior to model training. The analysis focused on understanding feature behaviour, identifying anomalies, and evaluating associations with loan approval outcomes. The dataset consists of 4,269 records and 13 attributes. Descriptive statistics revealed substantial variation in feature scales, such as annual income versus loan term, justifying feature standardization to avoid bias in distance-based algorithms. The class distribution (62% approved, 38% rejected) indicated a reasonably balanced dataset, supporting the use of accuracy-based evaluation metrics. Correlation analysis identified the CIBIL score as the most influential predictor, particularly for values above 748, reinforcing its significance in credit evaluation. The engineered *Assets* feature further showed that higher asset-to-loan ratios substantially increase approval likelihood. To enhance model robustness, outlier detection was performed using Z-score analysis, and observations with $|Z| > 3$ were removed to mitigate overfitting.

4. Model Development

To develop a robust predictive framework, seven supervised learning algorithms were evaluated, each representing a distinct modelling strategy for credit risk assessment.

- **Logistic Regression:** Served as a linear baseline model, using balanced class weights to evaluate linear relationships between financial features and loan approval outcomes.
- **K-Nearest Neighbours (KNN):** Assessed distance-based classification performance, demonstrating strong sensitivity to feature scaling and validating the need for standardized inputs.
- **Support Vector Machine (SVM):** Applied to maximize class separation in high-dimensional space, with probability estimates enabling ROC-AUC–based performance evaluation.
- **Gaussian Naïve Bayes:** Used as a probabilistic benchmark to test whether simple independence assumptions among features could yield competitive predictions.
- **Ridge Classifier:** Introduced to address multicollinearity among correlated financial attributes through L2 regularization, reducing overfitting in linear models.
- **Linear Discriminant Analysis (LDA):** Provided a statistical classification perspective by projecting data to maximize inter-class separation under normality assumptions.
- **Decision Tree Classifier:** Offered high interpretability through rule-based splits but exhibited high variance, motivating the use of ensemble techniques. Its limitations led to the adoption of Bagging, which achieved the highest test accuracy of **98.36%**.

5. Ensemble Learning Model

To address the high variance and overfitting of individual decision trees, three ensemble learning techniques were evaluated. These methods combine multiple base estimators to improve stability and generalization, which is particularly effective for credit risk modelling.

- **Bagging Classifier:** Implemented using 100 decision trees trained on 60% bootstrapped samples, Bagging effectively reduced variance and achieved the best performance, with a testing accuracy and F1-score of **98.36%**.
- **Random Forest Classifier:** By introducing feature-level randomness during tree construction, Random Forest further decorrelated estimators, resulting in a competitive testing accuracy of **97.54%**.
- **AdaBoost Classifier:** This sequential boosting approach emphasized misclassified instances using shallow decision trees (max depth = 2). With a learning rate of 0.5, it achieved an accuracy of **96.84%**, though it was outperformed by variance-reduction methods.

6. Model Deployment

To translate the predictive framework into a practical application, the optimized Bagging Classifier was deployed as an interactive Streamlit web application. The system follows a standardized *train–save–serve* pipeline, using serialized .pkl files for both the trained model and the StandardScaler to ensure consistency across data ingestion, processing, and inference stages. The user interface employs constrained sliders and dropdowns to minimize input errors, enabling real-time feature scaling and prediction. To satisfy regulatory transparency requirements, the application integrates a custom SHAP-based explanation module, presenting the loan decision alongside a dynamic SHAP waterfall chart that quantifies the influence of key features such as asset value and loan term.

 ISSN: **3107-6513**

7. System Architecture

The system architecture, illustrated in Figure 2, presents an end-to-end workflow for automated credit assessment. The process begins with the Loan Approval Dataset, where key financial attributes such as credit score, assets, and education are ingested into the computational pipeline. The framework is organized into three primary modules: data loading, model training, and prediction. Input data undergo preprocessing and exploratory data analysis (EDA) to ensure quality and consistency before being passed to the ensemble-based loan classifier, which employs a Bagging strategy for optimized predictive performance. An integrated SHAP-based explainability module then generates interpretable insights into the model's decisions. The workflow concludes with deployment through a Streamlit web interface, where users receive an immediate, color-coded loan decision—approved or rejected—effectively translating complex model inference into an accessible real-world application.

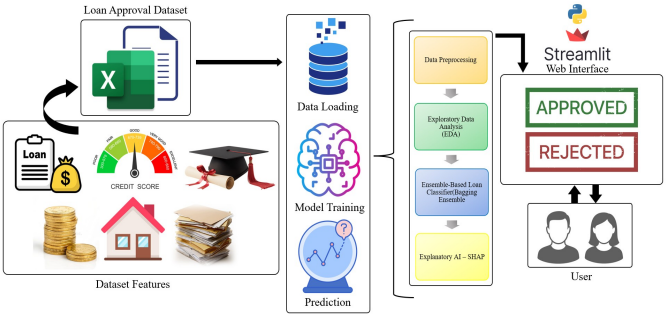

Fig 2. System Architecture

IV. RESULTS

1. Performance Comparison of Model Used

Model	Accuracy	Execution Time
Logistic Regression	0.91	0.14 s
KNN	0.90	0.02 s
SVM	0.93	1.13 s
Naive Bayes	0.94	0.00 s
Ridge Classifier	0.92	0.06 s
LDA	0.92	0.04 s
Decision Tree	0.96	0.03 s
Random Forest	0.97	1.04 s
Bagging	0.98	0.06 s
AdaBoost	0.97	0.37 s

Table 1 – Model Performance Comparison

2. Diagnostic Performance Analysis

Among all evaluated models, the **Bagging classifier** emerges as the best-performing approach, achieving the **highest accuracy of 0.98** with a **low execution time of 0.06 seconds**. This balance of superior predictive performance and computational efficiency makes Bagging the most suitable model for real-time loan approval applications.

The table compares multiple classification models based on **prediction accuracy** and **execution time**, highlighting the trade-off between performance and computational cost. Traditional models such as Logistic Regression and KNN show faster execution but comparatively lower accuracy, while more complex models like SVM and Random Forest achieve improved accuracy at the expense of higher computation time.

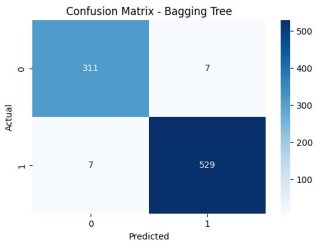

Fig 3. Confusion Matrix

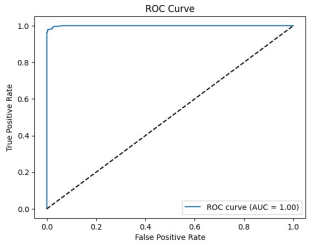

Fig 4. ROC Curve

To validate the performance of the optimized Bagging Classifier, evaluation was performed using a Confusion Matrix and ROC curve (Figure 3). The Confusion Matrix indicates strong predictive accuracy, with 529 true positives and 311 true negatives. Misclassifications were minimal and balanced, with only 7 false positives and 7 false negatives, demonstrating reliable and unbiased decision-making.

The ROC curve further confirms the model's effectiveness, with the curve closely following the top-left boundary. An AUC score of **1.00** indicates perfect discriminatory capability, highlighting the model's ability to accurately distinguish between approved and rejected loan applications across all classification thresholds.

3. Feature Impact Analysis

Feature	Importance Score
CIBIL Score	0.8254
Loan Term	0.0807
Loan Amount	0.0415
Income Annum	0.0315

Table 2 – Feature Importance

Beyond predictive performance, understanding feature influence is essential for model transparency. Table 3 presents global feature importance, identifying the CIBIL score as the dominant factor (importance ≈ 0.83), consistent with standard credit evaluation practices. The influence of the top features is further examined using distribution plots (Figures 4a–5b). The CIBIL score exhibits a clear threshold beyond which approval probability increases sharply. Analysis of loan term and loan amount indicates a preference for shorter durations and moderate principal values to reduce default risk. Additionally, annual income shows a positive correlation with approval likelihood, reflecting improved repayment capacity at higher income levels.

ISSN: **3107-6513**

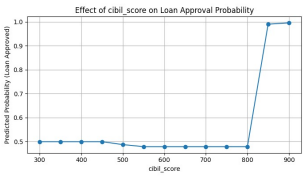

Fig 5. CIBIL Score

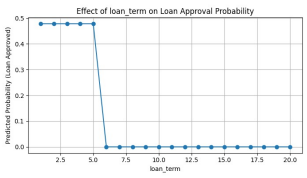

Fig 6. Loan Term

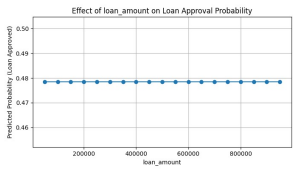

Fig 7. Loan Amount

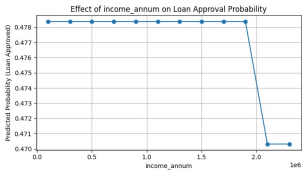

Fig 8. Income Annum

4. Final Deployment and Output

The deployed system integrates the Bagging classifier with SHAP to deliver instance-level interpretability. A dynamic SHAP waterfall plot decomposes each prediction, clearly showing how individual features influence the final outcome. In approval cases, positive contributors such as high asset value and strong CIBIL score incrementally raise the base probability to a final approval score of 99%. In rejection cases, negative factors—including low annual income and sub-optimal CIBIL score—are visualized as reducing the approval probability to 47%. This bidirectional explanation capability ensures that the ensemble model meets the interpretability requirements of modern banking systems.

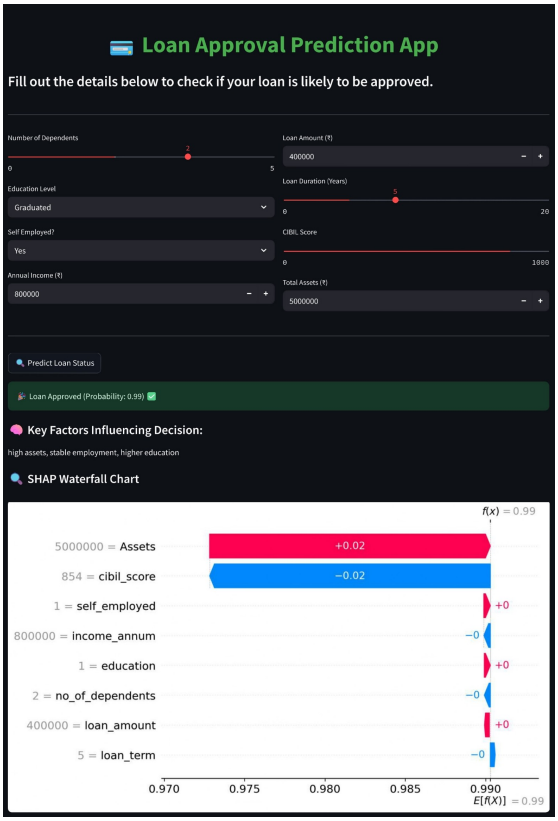

Fig 9. Loan Approved

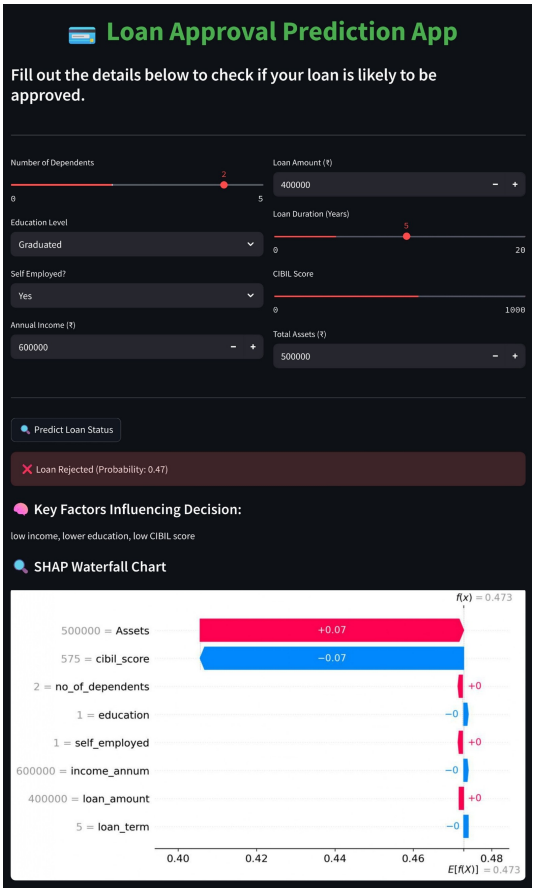

Fig 10. Loan Rejected

V. Conclusion & Future Scope

This study presents a robust, automated loan approval prediction system that addresses the limitations of traditional manual credit assessment. Through a comparative evaluation of seven supervised learning models and three ensemble techniques, the Bagging Classifier emerged as the most effective approach, achieving a testing accuracy of **98.36%** and an **AUC of 1.00**, significantly outperforming baseline models such as Logistic Regression and KNN.

The integration of the SHAP explainability framework enabled transparent decision-making, mitigating the "black-box" limitation of high-performing models. Feature attribution analysis identified the CIBIL score as the most influential factor in loan approval decisions, followed by loan term and loan amount.

The system was successfully deployed via a Streamlit-based web application, delivering real-time, interpretable predictions and reducing loan processing time from days to seconds while lowering the risk of non-performing assets. Future work will explore deep learning models to capture more complex non-linear patterns in large-scale and unstructured financial data.

Future Scope

- **Integration of Deep Learning Models:** Future enhancements may incorporate deep learning

techniques such as Artificial Neural Networks (ANNs) or hybrid ensemble–deep learning architectures to capture complex non-linear relationships in large-scale financial datasets and improve predictive accuracy.

- **Scalable Cloud-Based Deployment:** The system can be deployed on cloud-based infrastructure using microservices architecture to support high-volume, real-time loan processing while ensuring scalability, reliability, and fault tolerance.

REFERENCES

[1] S. Sharmila, P. Sandhya, P. S. Kousar, P. Anuradha, and S. Deekshitha, "Bank loan approval using machine learning," 2024 International Conference on Integrated Circuits and Communication Systems (ICICACS), 2024.
[2] C.Prasanth, A. Rangesh, R. P. Kumar, N.Sasmitha, andB.Dhiyanesh, "Intelligent loan eligibility and approval system based on random forest algorithm using machine learning," 2023 Interna tional Conference on Innovative Data Communication Technologies and Application (ICIDCA), 2023.
[3] E. Kadam, A. Gupta, S. Jagtap, I. Dubey, and G. Tawde, "Loan approval prediction system using logistic regression and cibil score," 2023 IEEE Fourth International Conference on Electronics and Sustainable Communication Systems (ICESC), 2023.
[4] X. Wang, Z. Kr¨aussl, M. Zurad, and M. Brorsson, "Effective automatic feature engineering on f inancial statements for bankruptcy prediction," IEEE International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME), 2023.
[5] A. Shepard and N. Naheed, "Application of data transformation techniques and train-test split," 2021 International Conference on Computational Science and Computational Intelligence (CSCI), 2021.
[6] R. Karthiga, G. Usha, N. Raju, and K. Narasimhan, "Transfer learning based breast cancer clas sification using one-hot encoding technique," 2021 IEEE International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021.
[7] D. Jurafsky and J. H. Martin, "Logistic regression (chapter 5 draft)," Speech and Language Pro cessing, 2024.
[8] P. Zang, "Application of id3 decision tree classification algorithm in mathematical data analysis," 2023 IEEE ICIICS Conference, 2023.
[9] P. Kumar, U. L. Maneesh, and G. M. Sanjay, "Optimizing loan approval decisions: Harnessing ensemble learning for credit scoring," 2024 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), 2024.
[10] A. Badhan, P. Kaur, A. Rana, B. Saha, and S. S. Malhi, "A comparative analysis for loan approval prediction using machine learning," 2024 International Conference on Electrical Electronics and Computing Technologies (ICEECT), 2024.
[11] N. R. Deborah, A. S. Rajiv, A. Vinora, M. C. Devi, M. S. Arif, and M. G. S. Arif, "An efficient loan approval status prediction using machine learning," Proc. InCACCT 2025, 2025.
[12] V. Singh, A. Yadav, R. Awasthi, and N. Partheeban, "Prediction of modernized loan approval system based on machine learning approach," 2021 International Conference on Intelligent Tech nologies (CONIT), 2021.
[13] J. Tsiligaridis, "Tree-based ensemble models and algorithms for classification," 2023 Interna tional Conference on Artificial Intelligence in Information and Communication (ICAIC), 2023.
[14] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001.
[15] U. E. Orji, C. H. Ugwuishiwu, J. C. N. Nguemaleu, and P. N. Ugwuanyi, "Machine learning models for predicting bank loan eligibility," 2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development (NIGERCON), 2022.
[16] M. Karntardzic, "Data mining: Concepts, models, methods, and algorithms," IEEE Press, 2003.
[17] V. Bahel, S. Pillai, and M. Malhotra, "A comparative study on various binary classification algo rithms and their improved variant for optimal performance," 2020 IEEE Region 10 Symposium (TENSYMP), 2020.
[18] R. R. Aritro Saha and B. C. Sahana, "Advancement of loan approval system for diverse applicants with machine learning framework," 2025 IEEE 14th International Conference on Communication Systems and Network Technologies (CSNT), IEEE, 2025.
[19] R. Chen and D. Li, "Mutual information reduction techniques and its applications in feature engineering," 2025 IEEE International Conference on Consumer Electronics (ICCE), 2025.
[20] J. Maul and K. Moon, "Neural network ensembling with random features," 2024 International Conference on Machine Learning and Applications (ICMLA), 2024