Machine Learning for Risk Assessment: A Comparative Study of Models Predicting Loan Approval

Amara Gnana Sirishma, Patil Hemanth Kumar Reddy, Yoshitha Tulasi,
Kasala Harsha Vardhan Reddy, Prem Reddy YS, Gayatri Ramasamy
Department of Artificial Intelligence,
Amrita School of Computing, Bengaluru
Amrita Vishwa Vidyapeetham, India
bl.en.u4aie22105@bl.stundents.amrita.edu, bl.en.u4aie22134@bl.stundents.amrita.edu,
bl.en.u4aie22177@bl.stundents.amrita.edu, bl.en.u4aie22169@bl.stundents.amrita.edu
bl.en.u4aie22166@bl.stundents.amrita.edu, r_gayathri@blr.amrita.edu

***Abstract*—Implementing an automated system for loan approval using machine learning offers significant advantages by enhancing the speed and accuracy of credit decisions, reducing human bias, and increasing the efficiency of the lending process. These systems can process large volumes of data rapidly, recognizing patterns and evaluating credit risks with a precision that exceeds traditional manual methods. Implementation of different machine-learning models in this paper. This paper proposes a machine learning-based approach for predicting the approval status of loan applications using a dataset containing applicant attributes. Having various preprocessing techniques and implemented nine different types of classifiers which include Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Gradient Boosting, AdaBoost, and XGBoost., and MLP Classifier for making the prediction, the system aims to automate and optimize the loan approval process in financial institutions. The successful implementation offers a valuable tool for enhancing the efficiency of whether a loan must be approved or not in the financial sector so that decisions can be quicker. After implementing all the classifiers, the best classifier was found to be the Support Vector Machine classifier, achieving an accuracy of 90.5% and an F1 score of 0.907.**

***Index Terms*—Loan Approval, Machine Learning, Predictive Modeling, Financial Institutions, Transparency, Interpretability**

I. Introduction

In the rapidly evolving financial landscape, the efficiency and reliability of loan approval processes are crucial for both customers and financial institutions. Traditional methods of loan approval, reliant on manual assessment, are not only time-consuming but also prone to human error and bias. Automated loan approval prediction, leveraging advanced machine learning techniques, emerges as a crucial innovation to enhance the speed, accuracy, and objectivity of credit decisions. Automation in loan approval not only streamlines the process but also democratizes financial services, ensuring fair and equitable treatment of all applicants based on data-driven insights. By automating the evaluation process, banks can process applications more swiftly and with greater accuracy, reducing the incidence of non-performing assets while increasing customer satisfaction.

Machine learning (ML) has introduced radical changes in diverse sectors by introducing advanced means of analysis and interpretation of complex data. In several areas, such as cyberbullying detection, identification of fake news, and spam detection, this is one application where the value of ML is high, but traditional approaches have a poor performance. In the context of cyberbullying, different word embeddings are tested using deep learning classifiers for the task of classifying social media comments as cyberbullying comments or not. The best performance across most datasets was achieved using GloVe embeddings in a Bi-LSTM GRU classifier[1]. Detection of fake news has also been done with automatic features and the various deep learning models and word embeddings, in which the BERT classifiers perform much better than many other models. These classifiers get a test accuracy of 99.20% on benchmark datasets[2]. Indeed, when several machine learning algorithms were combined and integrated into the PySpark framework, spam detection in messages improved. Experiments show that the SVM classifier, enhanced using the SMOTE method for balancing datasets, obtains the highest accuracy and F1 score. The main focus areas of these studies would be machine learning toward enhancing the accuracy and efficiency of automated systems involved in cyberbullying, fake news, and spam detection. This research shows that the transformational potential of ML in the design of safe and reliable digital environments is high[3].

This project shows different Machine Learning models, including Logistic Regression, Decision Trees, Random Forest, and Support Vector Machines (SVM) to predict loan approval outcomes. based on comprehensive datasets encompassing applicants' demographics, financial history, and other pertinent factors. The methodology follows a structured process for data collection, feature engineering, and thorough model selection and evaluation, with performance measured using metrics like accuracy, precision, recall, and the Area Under the ROC Curve (AUC-ROC). By presenting a detailed exploration of these models, the study contributes important insights into the role of automation and machine learning in enhancing the efficiency and fairness of loan approval processes. These findings guide future implementations of machine learning in financial decision-making,opening the door to more innovative, efficient, and fair financial services.

The structure of this paper is organized as follows: Section 2 presents an overview of related work. Section 3 details the methodology applied in this study. Section 4 provides a discussion on results. Finally, Section 5 concludes the paper and suggests areas for future research to further enhance the results obtained.

II. Literature Survey

In our literature review, we have studied various papers that explore the application of machine learning in predicting loan approval.

Khan, A. et al. highlight the efficacy of Random Forest in accurately forecasting loan outcomes, potentially reducing approval time and manpower for banks. Additionally, the paper discusses the potential of Genetic Algorithms to optimize lending decisions, emphasizing the importance of model selection in minimizing errors and maximizing profitability[4]. Alagic,A and et al. investigate the incorporation of mental health data into loan approval prediction models through the use of machine learning techniques. It demonstrates the potential of leveraging diverse data sources to enhance credit risk analysis, aiding in the identification of customers at higher risk of default. The findings emphasize the importance of comprehensive data utilization for more accurate loan approval predictions, benefiting financial institutions in minimizing potential losses[5]. Uddin et al. propose an ensemble machine learning-based system for predicting bank loan approvals, addressing the challenges of manual assessment processes. By leveraging diverse ML models and a user-friendly application interface, it enhances accuracy and efficiency in identifying qualified loan applicants, contributing to improved risk management practices in the banking industry[6].

Dansana et al. investigate the influence of loan features on bank loan prediction using the Random Forest algorithm[7], with the goal of improving the loan approval process and reducing the risk of defaults. Through analysis of various parameters and classification models, it provides insights into improving the efficiency and reliability of loan approval systems, crucial for maintaining financial stability in the banking industry. In the context of optimizing resource-constrained systems, Gurupriya et al. (2023)[8] introduced the Combinatorial Bat Optimization (CBOA) algorithm for Wireless Sensor Networks (WSNs) to improve energy efficiency and fault tolerance. By dynamically selecting backup cluster heads and optimizing routing based on multi-objective criteria, the approach ensures reliability in decision-making processes. This methodology parallels optimization challenges in loan approval systems, where efficient resource management and accurate classification are critical. The principles of CBOA can be adapted to optimize machine learning models for loan approval by enhancing decision pathways and minimizing errors under constrained condition. Priscilla, R and et al.[9] focus on baseline modeling for early prediction of loan approval systems, aiming to improve the accuracy of identifying potential defaulters. By employing machine learning techniques, particularly the Random Forest algorithm, it offers a promising approach to automate loan approval processes, reduce default risks, and enhance the overall efficiency of lending operations in financial institutions. Yasaswini, P. and et al. [10]focus on analyzing and forecasting bank loan approval data using machine learning algorithms, aiming to improve the efficiency of selecting safe loan applicants. By training models on past loan records, particularly utilizing SVM and Random Forest algorithms, it provides a promising approach to predict loan safety and enhance decision-making processes in the banking sector. Deep learning (DL) models have shown significant promise in healthcare, particularly in Alzheimer's disease detection. Advanced DL techniques combined with the Synthetic Minority Oversampling Technique (SMOTE) address the issue of class imbalance, improving the accuracy of MRI-based classification models for Alzheimer's progression (Krishnaa et al.)[11]. Models like CNN, ResNet50, and EfficientNetB2 have demonstrated enhanced reliability after SMOTE-based augmentation. This strategic handling of imbalanced data through SMOTE has significantly contributed to better generalization and accuracy in predicting Alzheimer's disease stages.

Singh and et al. [12] introduce a novel approach to incorporating Responsible AI techniques, specifically focusing on explainability and fairness, into the loan approval process. By implementing a proprietary framework with functionalities such as standardized explainability and fairness tools, it enhances trust and reliance on AI systems while addressing ethical concerns in decision-making processes.Gunjan Pahuja et al.[13] conducted a survey on existing computational intelligence strategies for disease detection, focusing on classifiers like Multilayer Perceptron[13], Support Vector Machine, and K-Nearest Neighbor. They found the Levenberg Margurandt algorithm to be the most effective, achieving a maximum accuracy of 96% in detecting Parkinson's Disease, highlighting the importance of classifier selection for accurate diagnosis Varshaa Sai Sripriya et al. (2022)[14] proposed a predictive machine learning model to assist banks in determining loan eligibility based on financial records. The model targets entries with a credit score above 700 and guarantees a minimum balance of Rs. 5000 after loan repayment. Various techniques, including Artificial Neural Network, Gradient Descent, XGBoost, Random Forest, and Support Vector Machine, were assessed, with the Artificial Neural Network showing the highest accuracy, making it the most effective solution for loan approval prediction. Sarath Krishnan et al. [15](2016) introduced a method to rank features automatically for classification tasks, aiming to identify which features are significant and which are not. The proposed method helps in reducing dimensionality and improving classifier performance by ranking features based on their significance. This approach addresses challenges in pattern recognition and is particularly useful in systems where certain features might cause confusion, leading to misclassification. The study highlights the efficiency of the proposed method in various applications, including fruit identification and industrial inspection. G. Radhika et al. (2016)[16] developed a system for virtual try-on of jewellery using Bayesian skin classifiers to enhance customer satisfaction and reliability in jewellery shops. The system detects skin pixels and places the jewellery virtually, eliminating the need for physical trials that could damage the jewellery. The method uses augmented reality to provide a realistic view of jewellery on the customer's hand, making the shopping experience more efficient and reducing the crowd during peak seasons. This approach leverages advancements in technology to create smarter retail solutions.

III. METHODOLOGY

The flow of the implementation goes as the flowchart I have given below:

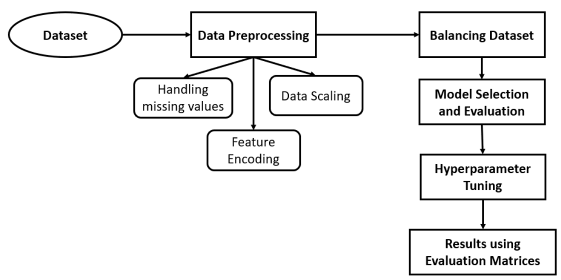

Fig. 1. Flow Chart of the methodology

A. Data Description

The dataset utilized in this study is from Kaggle and consists of records from loan applications, each containing multiple attributes that influence loan approval decisions. These attributes comprise demographic information, financial history, which are believed to affect the likelihood of default and thus the decision on loan approval, and loan details such as Gender, Married, Dependents, Self_Employed, LoanAmount, Loan_Amount_Term, and Credit_History.

B. Data Preprocessing

The number of missing entries identified and addressed in each column were as follows: Gender: 13 missing, Married: 3 missing, Dependents: 15 missing, Self_Employed: 32 missing, LoanAmount: 22 missing, Loan_Amount_Term: 14 missing, and Credit_History: 50 missing. The dataset loan approval status, with 422 instances of 'Yes' approvals compared to only 192 instances of 'No' rejections, indicating a disparity that could potentially influence the predictive performance of the models. Effective data preprocessing is crucial for optimizing the performance of machine learning models. The steps taken include:

1) Handling Missing Values: Missing data can lead to biased or incorrect model predictions and can affect the performance of almost all algorithms. Categorical Features were imputed with the mode to maintain the statistical distribution of the dataset, and Numerical Features were imputed with the median to handle outliers more robustly than the mean.

2) Feature Encoding: Most machine learning models require input to be numeric. Encoding transforms categorical data into a numerical format, ensuring that the algorithms can interpret them correctly. One-hot Encoding was applied to nominal categorical variables to convert categories into a binary matrix representation, which prevents the models from assuming a natural ordering between categories. Label Encoding was used for ordinal categorical variables or binary categories, assigning a unique integer based on alphabetical ordering.

3) Data Scaling: Many machine learning algorithms are sensitive to data scaling, and having features on a similar scale can enhance their convergence and performance. The StandardScaler was applied to standardize the features by removing the mean and scaling them to unit variance. This step is particularly important for algorithms that rely on distance computations between data points, such as K-Nearest Neighbors (KNN) and Support Vector Machines (SVM).

C. Balancing the Dataset

Addressing class imbalance is important to prevent model bias towards the majority class, especially prevalent in datasets concerning events like loan defaults or diseases. In tackling imbalanced datasets, Ramasamy et al.[17] introduced a cost-sensitive learning framework combined with multi-class classification and undersampling techniques for pest identification in the Coconut Leaf Dataset. This approach effectively addresses class imbalance by assigning different misclassification costs and utilizing undersampling to reduce bias towards overrepresented classes. Such techniques can be valuable in loan approval prediction, where class imbalance (i.e., more loan approvals than denials) is prevalent. Implementing a similar cost-sensitive approach could improve the accuracy of predicting loan denials, reducing biases and enhancing model fairness. Oversampling the minority class involves replicating the minority class instances to balance the distribution of classes, enhancing the model's capacity to learn from underrepresented classes. We

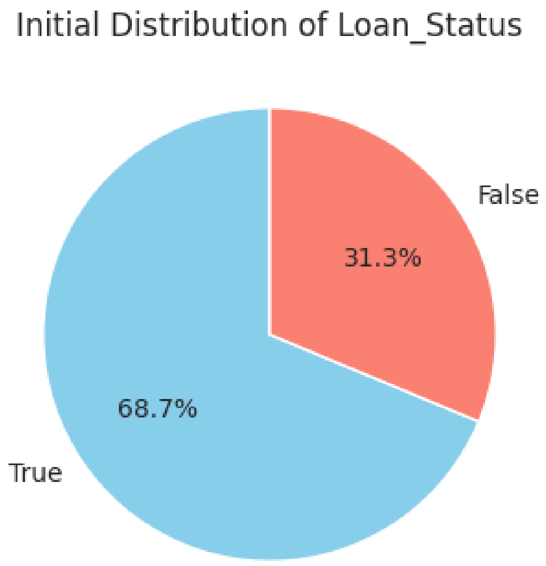

Fig. 2. Initial visualisation, showing data imbalance

have used random oversampling method to balance the dataset, which involves, replicating the minority class instances to balance the distribution of classes, enhancingthe model's capacity to learn from underrepresented classes.

D. Model Selection and Evaluation

Given the diversity of machine learning algorithms available, selecting appropriate models is vital to address the specific nature of the dataset effectively. The models selected and their rationale include:

Logistic Regression: A statistical model that estimates the probabilities using a logistic function, widely used for binary classification problems. Chosen for its efficiency and interpretability, serving as a baseline for performance comparison. In our dataset, Logistic Regression would be applied by fitting a model that estimates the probability of loan approval (Yes vs No) based on input features such as income, loan amount, credit history, etc. The output is a probability that ranges between 0 and 1, which is then thresholded (usually at 0.5) to decide the classification.

Decision Tree: A model that partitions the data into subsets using tree-like graphs of decisions and their possible consequences. Useful for handling nonlinear data patterns and easily interpretable, allowing for easy extraction of decision rules. For our dataset, a Decision Tree classifier will split the data into subsets based on the value of the most significant features which best separate the classes (approved or not approved). This results in a tree-like structure where each path from the root to the leaf represents classification rules.

Random Forest: A technique in ensemble learning for classification that constructs a multitude of decision trees at training time. It improves classification accuracy through bagging and feature randomness when building individual trees, reducing variance without increasing bias. Random Forest applied to our dataset by creating multiple trees (perhaps hundreds) and making them work on slightly different subsets of the data and/or features. Each tree casts a vote, and the most frequent classification (majority vote) is selected as the final prediction, enhancing robustness and accuracy.

Support Vector Machine (SVM): A dominant, flexible, and efficient algorithm for classification and regression, using a technique known as the kernel trick to transform the data, which helps identify an optimal boundary between the possible outputs based on these transformations. Particularly good for complex but small- or medium-sized datasets. In the context our dataset, SVM would analyze your data and find the best hyperplane that separates approved from not approved loans in the feature space. It is particularly effective if the boundary between classes is not linear, especially when using kernel SVM that can project the data to higher dimensions.

K-Nearest Neighbors (KNN): A non-parametric method used for classification and regression. A sample is classified through a plurality vote of its neighbors, with the sample being assigned to the class most common among its k nearest neighbors. KNN is very intuitive and has been selected due to its simplicity and efficacy in binary classification tasks. KNN used on our dataset by identifying the k closest samples in the data space to a given application and predicting its class based on the majority label of these neighbors. The selection of k and distance metric can significantly influence performance.

Gradient Boosting: Gradient Boosting is a machine learning technique used for both regression and classification tasks. It creates a prediction model by combining multiple weak learners, usually decision trees, into an ensemble. The model is built in a stage-wise manner, with each new model correcting the errors of the previous one, and it can be generalized by optimizing any differentiable loss function. Applied to the dataset, Gradient Boosting would sequentially add new models that correct the errors made by previous models. Each new model focuses more on the misclassified examples by the previous ones, striving to improve where the predecessors failed.

AdaBoost: A boosting algorithm that can be used along with various other types of learning algorithms to improve performance. The outputs of other learning algorithms ('weak learners') are aggregated into a weighted sum, which forms the final output of the boosted classifier. It is often used to boost the performance of decision trees on binary classification problems. In the loan approval dataset, AdaBoost begins by fitting a classifier to the original dataset. It then iteratively fits additional classifiers to the same dataset, adjusting the weights of incorrectly classified instances so that subsequent classifiers give more attention to the harder cases.

XGBoost: An implementation of gradient boosted decision trees designed for speed and performance. XGBoost provides a powerful prediction framework; it's frequently used in winning solutions to data science competitions and is widely used in industry. XGBoost would handle your dataset by building an ensemble of trees like Random Forest but incorporating techniques like gradient boosting to optimize loss functions, which can be particularly effective for large and complex datasets.

MLP Classifier: A neural network model that optimizes the log-loss function using LBFGS or stochastic gradient descent. Chosen for its capability to capture non-linear relationships through its layers and neurons. MLP would process the loan dataset through layers of neurons, each capable of learning different aspects of the data through their connections and weights. The final layer would make a decision on loan approval based on the inputs received from previous layers and the training process.

Each model was trained using the balanced dataset and evaluated across several metrics: accuracy, F1 score, recall, precision, and ROC AUC. We performed hyperparameter tuning to all the classifiers possible.

E. Hyperparameter Tuning

Hyperparameter tuning is a crucial step in optimizing the performance of the machine learning models. Each algorithm has specific parameters, known as hyperparameters, which can be adjusted to improve the model's accuracy, precision, recall, and other performance metrics. In this study, we utilized GridSearchCV, a comprehensive search method, to explore various combinations of hyperparameters for each classifier. The process involved defining a parameter grid for each model and conducting a cross-validation to determine the optimal set of hyperparameters that maximize the model's performance.

The hyperparameters tuned for Random forest include the number of estimators (trees), the maximum depth of the trees, the maximum number of features considered for splitting, and the criterion used for splitting nodes. The hyperparameters tuned for the SVM include the regularization parameter (C) and the kernel coefficient (gamma). The hyperparameters tuned for the KNN classifier include the number of neighbors (k) and the weight function used in prediction. For Gradient Boosting, the hyperparameters tuned include the number of boosting stages, learning rate, and maximum depth of the individual estimators. For AdaBoost, the hyperparameters tuned are the number of estimators and the learning rate. For XGBoost, the hyperparameters tuned consist of the number of estimators, learning rate, and maximum depth of the trees. For the MLP (Multilayer Perceptron) classifier, the hyperparameters tuned include the size of the hidden layers, activation function, and solver used for weight optimization. The parameter grid explored for the classifiers was:

TABLE I
HYPERPARAMETER TUNING FOR VARIOUS CLASSIFIERS

Model	**Parameter Grid**
Random Forest	{ 'n_estimators': [100, 200], 'max_features': ['auto', 'sqrt'], 'max_depth': [10, 20], 'criterion': ['gini', 'entropy'] }
SVM	{ 'C': [0.1, 1, 10], 'gamma': [1, 0.1, 0.01] }
KNN	{ 'n_neighbors': [3, 5, 7], 'weights': ['uniform', 'distance'] }
Gradient Boosting	{ 'n_estimators': [100, 200], 'learning_rate': [0.01, 0.1, 0.5] }
AdaBoost	{ 'n_estimators': [50, 100], 'learning_rate': [0.01, 0.1, 1] }
XGBoost	{ 'n_estimators': [100, 200], 'learning_rate': [0.01, 0.1], 'max_depth': [3, 6, 10] }
MLP Classifier	{ 'hidden_layer_sizes': [(50, 50), (100,)], 'activation': ['tanh', 'relu'], 'solver': ['adam', 'sgd'] }

Each model was trained using the balanced dataset and evaluated across several metrics: accuracy, F1 score, recall, precision, and ROC AUC. By performing hyperparameter tuning, we ensured that the classifiers were optimized to their full potential, leading to improved predictive performance.

IV. RESULTS AND DISCUSSION

A. Impact of Data Preprocessing

The importance of feature scaling and data balancing was evident in the improved performance of certain models. For instance, KNN and SVM, which are sensitive to the scale of the data, exhibited significantly better results after the application of StandardScaler. This preprocessing step ensured that all features contributed equally to the model, preventing features with larger scales from overpowering the decision-making process.

B. Evaluation Metrics Explained

1) Accuracy: Accuracy measures the proportion of correct results (both true positives and true negatives) out of the total number of cases examined. It offers a simple performance metric across all classes, making it suitable for balanced datasets.

2) F1 Score: The F1 Score is the harmonic mean of precision and recall, offering a balanced measure between the two. It is especially valuable when the costs of false positives and false negatives are significant and equally important.

3) Recall: Recall, also known as sensitivity, measures a model's ability to identify all relevant instances (true positives) within a dataset. High recall is essential in situations where missing a positive instance (e.g., a loan default) would result in higher risk or cost.

TABLE II
PERFORMANCE METRICS OF CLASSIFIERS

Classifier	Accuracy	Precision	Recall	F1 Score	ROC AUC
Logistic Regression	0.7638	0.6883	0.8983	0.7794	0.8719
Decision Tree	0.8622	0.8268	0.8898	0.8571	0.8640
Random Forest	0.8583	0.7808	0.9661	0.8636	0.9025
SVM	0.9055	0.8357	0.9915	0.9070	0.9258
KNN	0.8150	0.8198	0.7712	0.7948	0.9003
Gradient Boosting	0.8740	0.7986	0.9746	0.8779	0.9043
AdaBoost	0.8150	0.7415	0.9237	0.8226	0.8852
XGBoost	0.8465	0.8319	0.8390	0.8354	0.8971
MLP Classifier	0.8858	0.8346	0.9407	0.8845	0.9221

4) Precision: Precision quantifies the number of true positive predictions made out of all positive predictions (true and false positives). This metric is significant when the consequence of a false positive is critical.

5) Area Under the ROC Curve (AUC-ROC): The AUC-ROC curve is a performance metric for classification problems across different threshold settings. ROC is a probability curve, and AUC quantifies the model's ability to distinguish between classes. A higher AUC indicates a better model's capacity to separate the classes.

Table II summarizes the performance metrics for the classifiers evaluated in this study.

C. Graphical Representation of Results

We have used the ROC curves to visually represent our outputs, But in the figure below, we have shown only the four classifiers, that have given us the best results, first being the SVM, then MLP classifier, followed by gradient boosting and then the Random forest classifier is depicted in the fig 2:

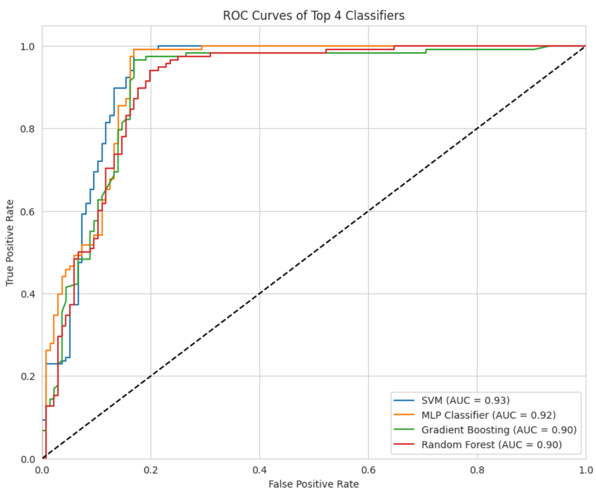

Fig. 3. Best 4 classifiers ROC curve outputs

These figures show ROC curve for each classifier, illustrating their performance across different thresholds and providing a visual comparison of their capability to manage trade-offs between recall and precision.

D. Detailed Performance Analysis

Each model's performance can be discussed in detail, noting specific strengths and weaknesses based on the results obtained. Random Forest, after hyperparameter tuning, demonstrated superior performance in terms of accuracy and AUC, highlighting its robustness across various features and its ability to handle high-dimensional data effectively. It consistently provided high accuracy and balanced precision and recall, making it a reliable choice for predicting loan approval. SVM and MLP also performed well, especially in high-dimensional spaces. SVM excelled in distinguishing complex patterns due to its ability to find the optimal hyperplane for classification. However, it required careful parameter tuning to avoid overfitting, which could otherwise compromise its generalizability. MLP, with its deep learning capabilities, captured nonlinear relationships effectively but demanded extensive computational resources and meticulous tuning to prevent overfitting, which posed challenges in achieving optimal performance without significant resource allocation.

Other models, such as Logistic Regression and Decision Tree, worked as important benchmarks in the evaluation process. Logistic Regression, while simple and interpretable, showed lower accuracy compared to more complex models but provided valuable insights due to its transparency in predicting probabilities. Decision Tree offered easy interpretability and handled nonlinear data patterns but was prone to overfitting, resulting in decreased performance on unseen data. K-Nearest Neighbors (KNN) displayed intuitive and effective classification for binary tasks but was sensitive to the choice of k and distance metric, impacting its accuracy. Gradient Boosting and AdaBoost, as ensemble methods, improved classification accuracy by focusing on difficult cases and correcting errors iteratively. However, they required careful tuning to balance bias and variance effectively. XGBoost, known for its speed and performance, handled large and complex datasets efficiently but required meticulous parameter tuning for optimal results. In summary, while Random Forest proved to be the most effective model in this study, other classifiers provided valuable insights and demonstrated specific strengths and weaknesses, contributing to a a thorough understanding of the dataset and model performance.

V. Conclusion and Future Development

This study explored the efficacy of various machine learning algorithms for predicting loan approval. Our analysis included Logistic Regression, Decision Tree, Random Forest, SVM, KNN, Gradient Boosting, AdaBoost, XGBoost, and MLP Classifier. The SVM classifier demonstrated the the best performance, achieving an accuracy of 90.55%, a recall of 99.15%, and an ROC AUC of 0.93.indicating its robustness in correctly identifying loan approvals. The MLP Classifier also showed high efficacy with an accuracy of 88.58% and an ROC AUC of 0.92, highlighting its ability to model complex relationships. Gradient Boosting and Random Forest also performed well, with ROC AUC values of 0.90 and 0.90, respectively. These results highlight the potential of advanced machine learning techniques in automating and improving the loan approval process, thereby reducing human error and improving decision-making efficiency in financial institutions.

For future work, we recommend exploring more sophisticated ensemble methods and deep learning architectures to further improve prediction accuracy and robustness. Incorporating additional features such as applicant behavioral data and economic indicators could provide deeper insights and enhance model performance. Moreover, deploying the best-performing models in a real-time environment would be valuable to assess their practicality and effectiveness in a live setting. Further research could also focus on interpretability techniques to ensure that the models provide transparent and explainable decisions, which is crucial for gaining trust from stakeholders and complying with regulatory standards. Lastly, investigating the impact of different data balancing techniques and feature engineering approaches could yield significant improvements in model generalization and performance.

References

1 Kumar, A. V., Gupta, D., and Venugopalan, M., "Cyberbullying text classification for social media data using embedding and deep learning approaches," in *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. IEEE, 2023.

2 Subhash, P. M. *et al.*, "Fake news detection using deep learning and transformer-based model," in *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. IEEE, 2023.

3 Sulochana, B. C. *et al.*, "Pyspark-powered ml models for accurate spam detection in messages," in *2023 2nd International Conference on Futuristic Technologies (INCOFT)*. IEEE, 2023.

4 Khan, A. and et al., "Loan approval prediction model: a comparative analysis," *Advances and Applications in Mathematical Sciences*, vol. 20, no. 3, 2021.

5 Alagic, A. and et al., "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*, vol. 6, no. 1, pp. 53–77, 2024.

6 Uddin, N. and et al., "An ensemble machine learning-based bank loan approval predictions system with a smart application," *International Journal of Cognitive Computing in Engineering*, vol. 4, pp. 327–339, 2023.

7 Dansana, D. and et al., "Analyzing the impact of loan features on bank loan prediction using random forest algorithm," *Engineering Reports*, 2023, e12707.

8 Gurupriya, M., Ramasamy, G., and Karthikeyan, B., "Combinatorial bat optimization for wsn with hoover index clustering," in *2023 International Conference on Next Generation Electronics (NEleX)*. IEEE, December 2023, pp. 1–5.

9 Priscilla, R. and et al., "Baseline modeling for early prediction of loan approval system," in *2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF)*. IEEE, 2023.

10 Yasaswini, P. and et al., "Analysis and forecasting of bank loan approval data using machine learning algorithms."

11 Krishna, K. S., Kumar, B. D., Reddy, M. D., Saketh, C. H., and Ramasamy, G., "A multi-class classification framework with smote based data augmentation technique for alzheimer's disease progression," in *2024 International Conference on Advances in Modern Age Technologies for Health and Engineering Science (AMATHE)*. IEEE, May 2024, pp. 1–6.

12 Purificato, E. and et al., "The use of responsible artificial intelligence techniques in the context of loan approval processes," *International Journal of Human–Computer Interaction*, vol. 39, no. 7, pp. 1543–1562, 2023.

13 Tallapureddy, G. and Radha, D., "Analysis of ensemble of machine learning algorithms for detection of parkinson's disease," in *2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, 2022, pp. 354–361.

14 Sripriya, S. V. S., Varrey, S. D. S., and Venkateshkumar, M., "Predictive model to compute eligibility test for loans," in *2022 IEEE Industrial Electronics and Applications Conference (IEACon)*, 2022, pp. 185–190.

15 S., P. and Krishnan, S., "Feature ranking procedure for automatic feature extraction," in *International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)-2016*, 2016.

16 G., R., T Kumar, S., Mangayarkarasi, N., and Nirmala, R., "Virtual try of a jewellery using skin classifiers," *International Journal of Control Theory and Applications*, vol. 9, no. 5, pp. 191–194, 2016.

17 Ramasamy, G., Gurupriya, M., Vasavi, C. S., and Karthikeyan, B., "A cost-sensitive learning approach with multi-class classification and undersampling techniques for pest identification in the coconut leaf dataset," in *2024 3rd International Conference on Artificial Intelligence For Internet of Things (AIIoT)*. IEEE, May 2024, pp. 1–5.