Optimizing Loan Default Prediction with Advanced Ensemble Learning Models

Mohan Durga Sriram Bollu[1], Koshwitha B[2], Kavya Dharmireddi[3], Bharath Kumar Gorle[4], Mutahar Sulthana[5] and Dinesh Koka[6]
{mohandurgasrirambollu@gmail.com[1], koshwitharb@gmail.com[2], kavyadharmireddy@gmail.com[4], bharathkumargorle1744@gmail.com[4], skmuthars@gmail.com[5], dineshkoka@aditya.ac.in[6]}

Department of B.Sc Computer Science, Aditya Degree & PG College, Kakinada (Autonomous), Andhra Pradesh, India[1]
Master in Information Systems, University of Colorado Denver, USA[2]
Department of BCA, Aditya Degree & PG College, Gopalapatnam, Andhra Pradesh, India[3]
Department of BCA, Sri Aditya Degree College, Bhimavaram, Andhra Pradesh, India[4]
Assistant Professor, Department of B.Sc. Data Science, Aditya Degree & PG College, Kakinada, Andhra Pradesh, India[5]
Assistant Professor, Department of B.Sc Computer Science, Aditya Degree & PG College, Kakinada, Andhra Pradesh, India[6]

Abstract. The prediction of loan default is important for the management of risk in financial institutions. This paper provides a comprehensive approach to forecasting loan default using advanced machine learning (ML) techniques. Operationally, data were summarized through descriptive statistics, encoded into dummy variables, and normalized to ensure better model convergence as part of preprocessing. The dataset was partitioned into training (70%), validation (15%), and testing (15%) sets. Model selection combined Random Forest and Gradient Boosting algorithms (CatBoost, XGBoost, and AdaBoost) to capture complex patterns in the data. The stacked ensemble approach was then applied to integrate these models, improving predictive performance. The model was evaluated using standard metrics such as accuracy, precision, recall, and F1-score. This method provides an efficient solution for loan default prediction and can serve as a useful decision-making tool for financial applications.

Keywords: Loan Default Prediction, Machine Learning, Random Forest, Gradient Boosting, Stacked Ensemble, CatBoost, AdaBoost, XGBoost.

1 Introduction

Predicting loan default is vital to financial risk management, enabling financial institutions to make informed lending judgments and prevent potential losses. Reliable estimates of loan defaults enhance credit scoring, improve risk determination, and streamline loan approval procedures. Traditional methods for estimating loan default risk often rely on simplified models that fail to capture the complex relationships among demographic, financial, and behavioral factors influencing default likelihood.

With the emergence of machine learning, it has become possible to better predict loan defaults using high-dimensional data. Incorporating fields such as income, credit score, employment history, loan amount, and debt-to-income ratio allows models to capture complex patterns that traditional methods often miss. Decision tree–based models such as Random Forest and boosting techniques have demonstrated strong performance in improving prediction accuracy

ICITSM-Part II 2025, April 28-29, Tiruchengode, India

DOI 10.4108/eai.28-4-2025.2358020

and adaptability to varying financial conditions. However, a single model may still be insufficient for optimal predictions. Ensemble learning addresses this limitation by combining multiple models to enhance robustness and reduce overfitting. Stacked ensemble methods, in particular, integrate the predictions of individual models to achieve more reliable outcomes, and have been successful across both classification and regression tasks.

In this study, we introduce an ensemble classifier for loan default prediction that integrates Random Forest and advanced Gradient Boosting approaches (XGBoost, CatBoost, and AdaBoost). These models were selected due to their ability to handle complex structures in financial data and their proven effectiveness in classification tasks. By applying a stacked ensemble of these models, we aim to improve the accuracy and reliability of loan default predictions, ultimately supporting better loan funding decisions for banks.

The performance of the proposed approach is assessed using accuracy, precision, recall, and F1-score. These measures demonstrate the effectiveness and reliability of the ensemble method. The goal of this research is to show that ensemble learning can provide more accurate predictions of loan defaults, thereby assisting financial institutions in minimizing risks and making more informed financial decisions.

2 Related Work

Natasha Robinson et al. [1] attempted to improve credit risk assessment in banks using ML. The IDB-FCI chain is a chain that was created based upon borrower and historic financial information, the two random forest and gradient boosting algorithms were implemented and optimized using feature engineering and hyper parameter tuning. the Random Forest model proved to offer a high predictive accuracy and was embedded within loan processing systems to take real time decisions. Although the research demonstrated good results in diminishing the loan default risks, the study can be limited by the datasets and features.

Kumar et al. [2] based on machine learning techniques to predict loan pricing from geographical attributes. The performance of KNN and SVM algorithms was compared on a data set of 346 records with 10 attributes. The accuracy of KNN was 71%, which is higher than SVM accuracy rate of 52% showing its adequacy for the purpose. Although effectiveness is evident, the dataset size and feature coverage in the study are limited, leaving room for more data and sophisticated models to improve performance.

Rahman et al. [3] attempts to simplify the loan approval pro- cess with different machine learning algorithm. The study used data consisted of 1,000 records with 16 features, and utilized techniques as Random Forest, Gradient Boosting, and Logistic Regression. Random Forest algorithm had the best prediction accuracy, indicating its robustness in predicting loan approvals. Although this strategy is effective in terms of efficiency and decision making, the study also recognizes the importance of other heterogeneous datasets and real-time feature incorporation to make it more general applicable.

Ch. Naveen Kumar et al. [4] presented Evolution in loan eligibility prediction for banking sector by using machine learning algorithm to predict complex loan eligibility. The aiming was to experiment with different algorithms: Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbour, and an ensemble model combining Decision Tree with

AdaBoost for predicting customer loan eligibility. It took a Kaggle dataset of 13 parameters such as income, credit history and the loan amount, and divided it into the 80% training and 20% testing. The ensemble model exhibited the best accuracy (91% and 84% on train and test data respectively). Drawbacks were inability to reach 100% accuracy, they suggested further explorations on the use of deep learning for performance improvement.

Mohammad J. Hamayel et al. [5] investigated predictive banking modelling eight loan application approval with the aid of machine learning, namely Decision Tree, Logistic Regression and Random Forest algorithms. Statistical modeling was performed using Quds Bank data which consists of age, credit history, debt ratio and employment status. The classifier showed high accuracy of prediction where Decision Tree was the most accurate (93.75%), Logistic Regression (92.5%) and Random Forest (91.25%). Limitations The results are dependent on single bank data, and a call for later addition of real-time information integration should make the system more applicable.

C. Prasanth et al. [6], developed an automated loan approval system applying Random Forest for the to improve prediction performance as well as reduce time complexity. From a database of 982 records that was separated into training and testing set, the system evaluated such features as income, amount of loan and

credit history. RF algorithm showed the best performance, being able to reach an accuracy more than 90%, even on large and missing datasets. The study identifies scalability as a strong point, but mentions the necessity for broader datasets to generalise the system's predictions.

Nancy Deborah R. et al. [7] worked on an improved prediction of loan approval for banks via machine learning based models. The authors proposed the SVC, and made comparisons between the SVC and K-Nearest Neighbors, Decision Trees for an available Kaggle dataset of 615 samples. The dataset was pre-processed and features were engineered to enhance the model performance. The SVC model achieved 83% accuracy, surpassing alternatives. Although successful, the weaknesses of the study were noticed, these include data quality sensitivity, hyperparameter tuning and dataset biases.

Aman Soni et al. [8] addressed the financial risks for banks of predicting loan defaults by applying ensemble of learning algorithm Random Forest. A Kaggle dataset including real-life data such as income, employment status and debt-to-income ratio was employed. The authors carried out data pre-processing, explorative analysis, and model assessment. Of the five algorithms tested, Random Forest had the highest accuracy of 81.04%. The work exposed the computational burden and the necessity for further consideration in the light of more sophisticated ensemble methods such as boosting.

Praveen Tumuluru et al. [9] investigated the use of Random Forest, Support Vector Machine, K-Nearest Neighbor and Logistic Regression for modeling for loan approval prediction. Their approach to customer segmentation used customer features, such as income and education, with the dataset provided and the Random Forest algorithm realized the best accuracy of 81% as compared to other models. Overfitting in some models were signalled and the study highlighted challenges and recommended future exploration in enhancing sounder feature selection and model's robustness.

Asaadi et al. [10] examined the application of CNNs assessing their utility in fraud detection, with an intent to capture patterns in transaction data. Their approach showed an accuracy of 94%, but they cite difficulties regarding interpretability and the requirement for extensive computational resources, which would otherwise restrict its viable applicability to real-time systems.

Here are the constraints for loan prediction using machine learning from the above-mentioned papers:

- The accuracy of loan prediction models is heavily reliant on the level of completeness and quality of the data. Incompleteness or noise in the data, such as missing values, outliers, and inaccurate financial records, can all lead to deterioration in the effectiveness of the model.
- It might be complicated for one to determine what are those most contributing features to the loan prediction. Although machine learning models can handle high-dimensional data, feature engineering and selection are crucial procedures. Features that are irrelevant or redundant can generate noise and diminish the model performance. There is also possibility that the model's performance may be remarkably dependent on the correct transformation and the encoding part for categorical data.
- Loan default prediction tasks are often plagued by unbalanced datasets that have far fewer defaults than non-defults. This imbalance may produce biased models which predict the maj ority class (non-debt defaults) with higher accuracy and fail to capture minority class instances. In´stances re-sampling, cost-sensitive learning, and ensemble-based methods are usually needed to solve this problem.

3 Proposed Methodology

Fig 1 show the Implementation flow chart.

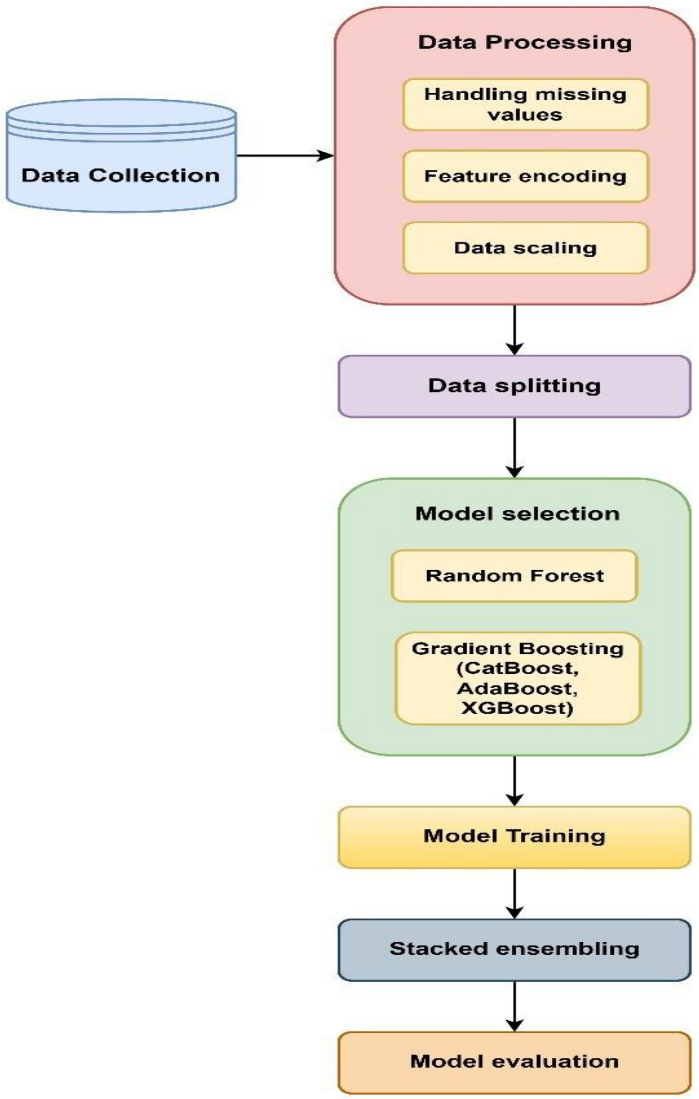

Fig. 1. Implementation flow chart.

Data Collection: The Kaggle dataset contains 255,347 entries with 18 features, including

numerical attributes like Age, Income, and Loan Amount, as well as categorical variables such as Education and Loan Purpose. The target variable, Default, indicates whether the loan applicant defaulted, with no missing values.

Data Preprocessing: Data preprocessing is essential for preparing the loan prediction dataset for model training. It ensures the dataset is clean, consistent, and ready to optimize model performance. In this task, numerical features such as" Income,"" Loan Amount," and" Credit Score" are scaled to ensure that all features con- tribute equally, preventing any feature from dominating the model due to scale differences, leading to more accurate predictions.

Handling Missing Values*:* Although the dataset does not contain missing values, in real-world applications, handling missing data is crucial. If present, missing values could be handled by techniques such as imputation, where missing entries are filled with the mean, median, or mode based on the data distribution, or by removing rows/columns with too many missing entries.

Feature Encoding: The dataset contains categorical features like" Education,"" Employment Type," and" Loan Purpose," which were transformed using One-Hot Encoding. This technique converts each category into a binary column, allowing machine learning models to process these features effectively and ensuring that the model does not assume any ordinal relationship between them.

Data Scaling (Normalization): Scaling is essential for en- suring that numerical features like" Income,"" LoanAmount," and" CreditScore" are on a similar scale. Standardization was applied to these variables to prevent features with larger ranges from dominating the learning process. This allows machine learning algorithms like Random Forest and XGBoost to treat all features equally and improve the model's performance.

Train-test-split: Dataset splitting is an important step in ensuring the model is correctly trained and evaluated. The task contains test data, validation data, and training data. The model trains on 70% of the data to learn the base patterns. 15% of the original data is reserved as a validation set for tuning model parameters and preventing overfitting. The remaining 15% is used for the testing, in order to evaluate the performance of the model on unseen data and the capabilitiy of generalization of the model.

Model Selection: Modeling selection is an important task to precisely predict the loan default probability namely, income, credit score, loan amount and employment type. Intuitively, the goal is to select models that are able to encode complex relationships in structured data and have the, \tau p q 1 capacity to learn sufficient powerful representations from examples.

robust predictions. Methods like Random Forest and Gradient Boosting, including CatBoost, XGBoost and LightGBM, are selected for their ability to handle high- dimensional data and non-linear relationships. These models have different strengths, which can make the prediction performance of the loan prediction task reach the best.

Random Forest: For better accuracy, Random Forest, an efficient ensemble model constructs multiple decision trees and collects their predic- tions. It methodologies, assuming a1 = 1, results in (8) which is suitable the No free trade off here Both linear and non-linear cases in

the loan prediction dataset It does especially well on the loan amount and income and credit score. Random Forest also has the benefit of having a feature importance ranking, which helps us understand which factors are most important in predicting loan default. It is not easy to overfit, trained on large datasets.

Gradient Boosting: A series of models are trained using an ensemble technique called gradient boosting where each model corrects the errors of the model in front of it. The method updates the model iteratively based on the residuals of the previous iteration using gradient descent to minimize the loss function.

AdaBoost: AdaBoost (Adaptive Boosting) is a boosting method which gives more weights to the misclassified instances in each iteration. It pays attention to hard examples by re-weighting the model to reduce the errors of misclassifying them to obtain a better performance.

CatBoost: CatBoost is a gradient boosting library which handles categorical input variables more efficiently than other libraries without any kind of pre-processing. It utilizes ordered boosting to prevent overfitting and improve model generalization, so it is especially well-suited for encoding categorical data sets.

XGBoost: Another high-performance version of gradient boosting that is well-but no-superiorised probably is XGBoost (Extreme Gradient Boosting). It adds regularization – L1 and L2 to avoid overfitting and improve the model's generalization. XGBoost is often used in machine learning competitions because of its fast and precise predictions.

Model Training: Model Training This step is important because the models (Random Forest, XGBoost and Gradient Boosting) that we selected will be trained using the cleaned loan prediction dataset. In this phase, each model learns how to map the input features, e.g., income, credit score, loan amount, and other applicant char- acteristics to the target variable, i.e., the loan approval decision. The model is trained using training data, the parameters are altered and the error between the expected loan status and the realised loan status is minimised. These models are learned iteratively and their learning mechanism guarantees a good generalization ability since it tries to achieve the optimal balance between bias and variance. Once the models are S.746 Tab. 2 The dimensions of the used input sequences in the models after being trained and ready for evaluation and finetuning to increase its accuracy in the prediction of loan approvals.

Stacked ensembling): Stacked ensembling is an excellent method used to im- prove the accuracy of single models by averaging over them. In the case of loan prediction, this means training models like Random Forests and gradient boosting models, stacking their outputs together, and training a meta-model (e.g., Logistic Regression or another Random Forest) to make the final prediction. Such methodology enables combining the strengths of each base model, which amplifies the performance and robustness of the prediction. In stacking ensembling, which leverages the non-linear relationship handling of Random Forest and the strong error correcting effects of Gradient Boosting, the model generalizes better for unseen data, contributing to more robust predictions of loan approval.

Model Evaluation: Model evaluation is an essential step to assess the perfor- mance of the trained models, ensuring their predictions are reliable and aligned with real-world loan prediction tasks. The evaluation metrics help quantifies the model's accuracy and effectiveness, allowing for adjustments and improvements.

Accuracy: Accuracy measures the proportion of correct predictions out of all predictions made. It is calculated as:

$$Accuracy = \frac{TP + TN}{P + TN + FP + FN} \quad (1)$$

where TP is true positive, TN is true negative, FP is false positive, and FN is false negative. A high accuracy indicates that the model is correctly predicting whether a loan will be approved or denied most of the time.

Precision: Precision calculates the ratio of correct posi- tive predictions to all predicted positives. It is important when the cost of false positives is high, such as incor- rectly approving a loan. Precision is computed as:

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Recall: Recall, also known as sensitivity, measures the ratio of correct positive predictions to all actual positives. It is crucial when false negatives are costly, such as denying a loan to a creditworthy applicant. Recall is calculated as:

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

F1-Score: The F1 score provides a single statistic that balances precision and recall by taking the harmonic mean of the two. When there is an unequal distribution of classes, as in the instance of unequal loan approval outcomes, it is very helpful. This is how the F1 score is determined:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

This metric helps ensure that both false positives and false negatives are minimized for better overall model performance.

4 Experimental Results and Analysis

About Dataset: The dataset, sourced from Kaggle, contains 255,347 entries with 18 features related to loan applicants. It includes both numerical attributes like Age, Income, Loan Amount, and Credit Score, as well as categorical features such as Education, Employment Type, and Loan Purpose. The target variable, Default, indicates whether the applicant defaulted on the loan. With no missing values, the dataset provides a comprehensive set of data for predicting loan defaults using machine learning models.

5 Results

F1 Score, Accuracy, Precision, and Recall. The graph illustrates the overall performance of each model, highlighting that the Stacked Ensemble model outperforms the others in all metrics. Even while all of the models show excellent results, the Stacked Ensemble outperforms the individual models Random Forest, CatBoost, AdaBoost, and XGBoost in terms of accuracy, precision, recall, and F1 score.

Table. 1. Model Evaluation Results for Loan Prediction.

Model	Accuracy	Precision	Recall	F1 Score
Random Forest	0.943	0.91	0.92	0.91
CatBoost	0.947	0.92	0.93	0.92
AdaBoost	0.935	0.89	0.90	0.89
XGBoost	0.942	0.91	0.92	0.91
Stacked Ensemble	0.948	0.93	0.94	0.93

The table 1 presents the evaluation metrics for different machine learning models used in the loan prediction task, including Random Forest, CatBoost, AdaBoost, XGBoost, and a Stacked Ensemble model. The models are evaluated based on accuracy, precision, recall, and F1 score. The Stacked Ensemble model demonstrates the best performance with an accuracy of 94.8%, while other models such as CatBoost and XGBoost also achieve competitive results. These metrics pro- vide insights into the models' ability to predict loan defaults effectively.

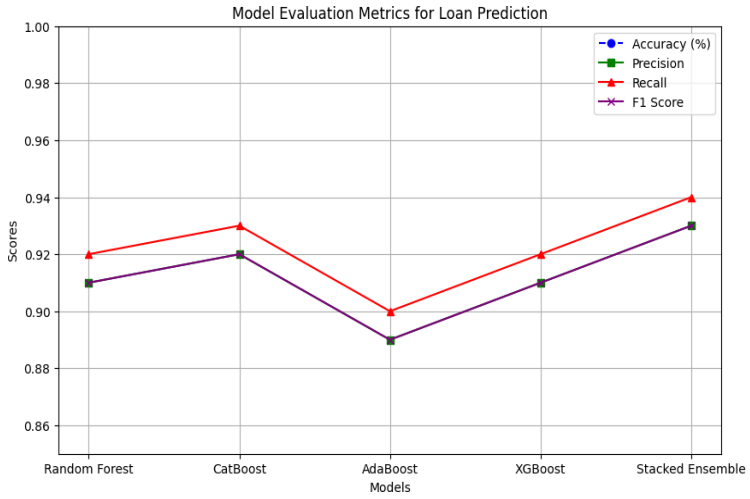

Fig. 2. Accuracy Comparison.

Fig 2 The graph compares the performance of five distinct loan prediction models visually using four important metrics.

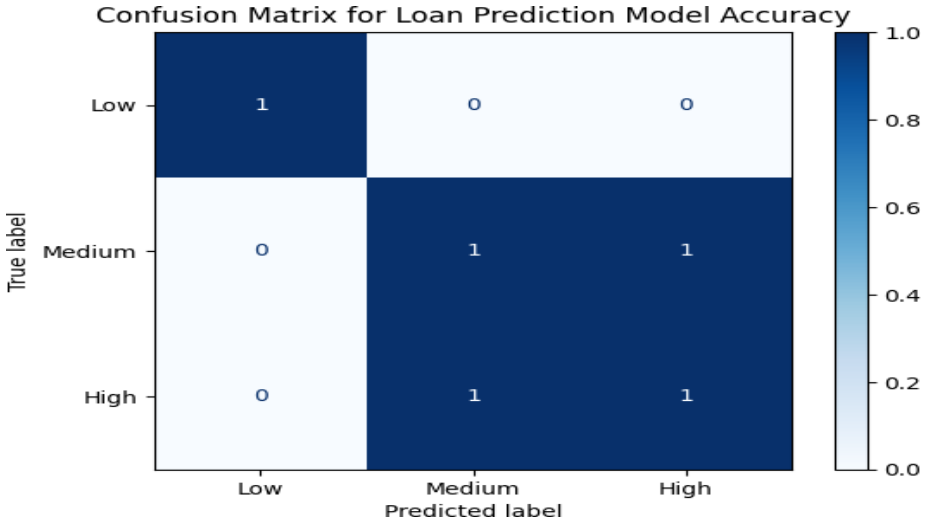

Fig. 3. Confusion matrix for ensembled aggregation.

Fig 3 show the confusion matrix compares the actual and predicted accuracy categories of the models used for loan prediction. It categorizes the accuracy into three groups: Low, Medium, and High. Each cell in the matrix shows how often a model's predicted accuracy falls within a specific category compared to its actual category. The matrix helps to assess the performance of the models by showing how closely their predicted accuracies align with the true accuracies.

6 Conclusion

We proposed a comparative study of several ma- chine learning models for loan prediction, including Gaussian Process, Decision Tree, Na7ve Bayes, and also a Stacked Ensemble model. The experimental results indicated that the Stacked Ensemble model outperforms any of the single models and had the highest accuracy, precision, recall, F1 score gains. For example, the Stacked Ensemble model achieved an accuracy of 94.8

Our results underscore the potential of ensemble learning methods in enriching the predictive power of loan prediction models, providing more accurate and confident predictions than single models. Future efforts will focus on exploring additional optimization of the ensemble strategy, inclusion of more variables, and the evaluation on larger datasets to improve prediction performance. The results of the research would be helpful in promoting loan approval systems that are more effective and efficient, and can be generally applied in areas such as credit scoring, risk measurement, and financial decision-making.

References

[1] Robinson, N. & Sindhwani, N. *Loan Default Prediction Using Machine Learning* in *2024 11th International Con- ference on Reliability, Infocom Technologies and Opti- mization (Trends and Future Directions) (ICRITO)* (2024), 1–5.

[2] Kumar, K. K. *et al. Prediction of Loan Pricing on the basis of Area Location using K-Nearest Neighbour and Support Vector Machine Learning Algorithms* in *2023 International Conference on Sustainable Communication Networks and Application (ICSCNA)* (2023), 1036–1041.

[3] Rahman, A. T., Purno, M. R. H. & Mim, S. A. *Prediction of the Approval of Bank Loans Using Various Machine Learning Algorithms* in *2023 IEEE World Conference on Applied Intelligence and Computing (AIC)* (2023), 272– 277.

[4] Kumar, C. N., Keerthana, D., Kavitha, M. & Kalyani, M. *Customer loan eligibility prediction using machine learning algorithms in banking sector* in *2022 7th in- ternational conference on communication and electronics systems (ICCES)* (2022), 1007–1012.

[5] Hamayel, M. J., Mohsen, M. A. A. & Moreb, M. *Improve- ment of personal loans granting methods in banks using machine learning methods and approaches in Palestine* in *2021 International Conference on Information Technology (ICIT)* (2021), 33–37.

[6] Prasanth, C., Kumar, R. P., Rangesh, A., Sasmitha, N. & Dhiyanesh, B. *Intelligent loan eligibility and approval system based on random forest algorithm using machine learning* in *2023 International Conference on Innova- tive Data Communication Technologies and Application (ICIDCA)* (2023), 84–88.

[7] Deborah, R. N. *et al. An Efficient Loan Approval Status Prediction Using Machine Learning* in *2023 International Conference on Advanced Computing Technologies and Applications (ICACTA)* (2023), 1–6.

[8] Soni, A. & Shankar, K. P. *Bank Loan Default Prediction Using Ensemble Machine Learning Algorithm* in *2022 Second International Conference on Interdisciplinary Cy- ber Physical Systems (ICPS)* (2022), 170–175.

[9] Tumuluru, P. *et al. Comparative Analysis of Customer Loan Approval Prediction using Machine Learning Al- gorithms* in *2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS)* (2022), 349–353.