IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, January 2026
Impact Factor: 7.67

Intelligent Loan Prediction of Defaults with Machine Learning for Financial Risk Analysis in Digital Lending

Bhalchandra Bapat
Independent researcher
bhalchandra@ssbm.ch

Abstract: *Financial institutions are constantly at risk of default by borrowers, which can result in significant financial losses. To minimize these risks and financial losses, it is necessary to create a proper predictive model of loan default. In order to mitigate these limitations, this paper concentrates on loan default prediction to conduct effective financial risk analysis through advanced machine learning (ML) techniques. Support Vector Machine (SVM), Decision Tree (DT), Logistic Regression (LR), Random Forest (RF) and Extreme Gradient Boosting (XGBoost) models are implemented and tested to measure their performance. Through experimental findings, it has been determined that the traditional models, such as SVM, DT and LR, are moderate in accuracy, whereas the ensemble methods are much more effective than the traditional models. Precisely, Random Forest model achieves highest accuracy of 99.96, and the XGBoost model achieves 99.99, both of which demonstrate high quality in addressing complex financial data. The results of this research highlight benefits of ensemble learning methods in improving prediction accuracy and facilitating more reliable financial risk decision-making.*

Keywords: Loan Default Prediction, Customer Behavioral Data, Credit Risk, ML, DL, Credit Scoring, Predictive Modeling

I. INTRODUCTION

Individuals rely on financial institutions for loans to address financial challenges, achieve personal goals or manage unforeseen expenses. Loans provide financial support that enables people to achieve their objective for purchasing a house, funding education, starting a business or paying off deft [1][2]. Loan borrowing has become an essential economic activity in a dynamic and highly competitive financial landscape. Lending serves as a key revenue that generates business for financial institutions [3]. Financial institutions have used credit scoring models that assess borrower creditworthiness through three specific factors that have been used since ancient times, such as debt-to-income ratio, income level, and credit history [4]. The risk assessments become incomplete and less accurate when critical factors such as economic conditions, geographic location and broader socio-economic aspects remain unaddressed [5].

Financial institutions continue to face a serious risk from loan default as it immediately affects their capacity to make money and preserve their financial stability. When borrowers fail to fulfill their obligations to repay their loans, lenders may suffer financial losses [6]. This means that proper forecasting of loan defaults is necessary for efficient allocation of credit and risk management [7][8]. Traditional statistical techniques, including logistic regression, can do this but are limited: they cannot process large, complex, non-linear data, so they cannot make accurate predictions. This brings out the necessity of more advanced and scaled analytical methods [9][10]. This fast influx of digital transactions in the form of online banking, e-commerce, and mobile payment systems has led to the situation where analysts are able to access huge amounts of both structured and unstructured data [11]. Previously, loan appraisal was mostly manual and thus time-consuming and subject to human bias [12]. Adoption of data-driven decision-making practices has achieved significant improvements in the decision-making processes by providing improved efficiency and uniform outcomes [13]. Machine

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, January 2026
Impact Factor: 7.67

learning (ML) techniques enable the processing of large datasets, discovering hidden patterns and developing models that describe complex relationships among variables, resulting in better prediction accuracy for loan default risk than conventional techniques.

The study develops a sophisticated intelligent loan default predictor, a system that combines both ML and DL models to conduct financial risk evaluations in online lending [14][15]. The suggest solution is based on an analysis of various borrower characteristics to sort applicants according to their defaulting probability [16][17]. The model uses advanced algorithms that can extract complex non-linear relationships to enhance its predictive accuracy, reduce default risk and assist financial institutions in making better decisions. The system proposed is more effective because it incorporates both the techniques of ML and DL, which enhances its credit risk assessment abilities.

A. Motivation

The current research paper aims to create more sophisticated credit scoring strategies based on the use of advanced ML and DL algorithms that generate more precise loan default forecasts. The research aims to use extensive digital financial data to identify complex relationships between borrower characteristics and their resulting behavior. Also, the research aims to improve risk management and minimize the risk of financial losses and facilitate efficient and data-driven sound decision-making in present-day lending systems.

B. Problem Statement

The growing number and complexity of financial data render the conventional credit scoring techniques insufficient for accurately forecasting loan defaults. Thus, it is necessary to have a data-driven, smart method of credit risk assessment and decision-making based on sophisticated ML methods.

C. Contribution

This research offers several key contributions as listed below:

- As a result, developed a complete data preprocessing pipeline of the Lending Club data, where outliers were removed and data normalized to have similar feature scales.
- One-hot encoded categorical variables using the used technique to successfully convert nominal data into a format appropriate for ML to enhance model performance.
- Remedied issue of imbalance in classes through SMOTE that rendered a balanced dataset, which improves model's dependability and predictability.
- Proposed an ML system that was made up of RF and XGBoost to precisely predict loan defaults and improve financial risk analysis in online lending platforms.
- Provided a detailed evaluation scheme based on the ROC curve, accuracy (acc), precision (prec), recall (rec), and F1-score (F1) to offer a trustworthy assessment of model's performance.

D. Justification and Novelty

The study is motivated by fact that reducing financial losses in banking and lending businesses greatly depends on accurate loan default prediction. It is innovative as it integrates more advanced ensemble learning algorithms, including XGBoost and RF, with balanced data treatment through SMOTE, thus, providing a higher prediction accuracy. Additionally, the new methodology captures intricate associations among financial data, thus enhancing the quality of risk measurement.

E. Organization of the Paper

The structure of paper is as follows: Section II includes a review of related works, and Section III outlines model implementation, the dataset, and the preparation techniques. Section IV describes the outcome of the experiment in a more or less comparative way and Section V conclude the study with the key findings and suggestions to do further

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, January 2026
Impact Factor: 7.67

research.

II. LITERATURE REVIEW

The study is informed by a detailed review and analysis of main research studies on Loan Default Prediction as a Financial Risk Analysis tool, which were done to improve the development of this study.

H. R et al. (2025) presented prior studies that rely on simpler models, ANN leverages Batch Normalization and Dropout layers to handle data imbalances and prevent overfitting, achieving an AUC-ROC score of 0.904 significantly higher than XGBoost (0.734) and RF (0.724) [18]. G. Kaur et al. (2025) suggested a good performance in a federated environment. Experimental results show that the ROC AUC is 0.897, which is above classical ML baselines. The approach maintains the data privacy by federated learning and hence it can be applicable in sensitive financial settings [19]. B. Gao (2024) The model's performance is assessed utilized a step-by-step addition technique, and a 2-hidden-layer BP neural network is ultimately determined. Accuracy, recall and precision are evaluated on the two models BP neural network and the KNN algorithm. The BP neural network model attains an accuracy of 70 percent or higher, a recall rate of over 55 percent and a precision level of over 60 percent [20].

Equally, S. K. C et al. (2024) seek to fill this gap by applying ML techniques to forecast loan eligibility more efficiently. They use the algorithms, including DT, Extra Trees, XGBoost, and LightGBM, where the last one has the best accuracy of 98.91, to process data and make more accurate predictions [21]. R. Nancy Deborah et al. (2023) introduced a bank loan prediction system that automatically selects qualified applicants to be given a loan. The study examined the ability to predict the status of loans based on ML. One of the algorithms tested was SVM whose accuracy score was 83 percent [22].

P. Pathak et al. (2023) proposed ML models, such as LR, DT, and RF, are trained on the preprocessed data for loan default prediction. The acc of these models is about 73%. An ensemble learning method is adopted in order to enhance the accuracy of prediction. The ensemble model enhances acc by improving predictions of numerous independent models by 76.8 percent [23]. U. E. Orji et al. (2022) recommended and investigated in Kaggle's Jupyter Notebook cloud environment utilizing Python programming tools. The investigation findings showed that there was high accuracy in performance with the RF approach recording the highest accuracy at 95.55% and the Lr rating the lowest at 80% [24].

Table I provides a brief summary of new studies on the topic of Loan Default Prediction with ML, outlining the suggested models, the datasets that they use, the key results, and the challenges that are faced

TABLE 1: RECENT STUDIES ON LOAN DEFAULT PREDICTION WITH ML

Author	Proposed Work	Results	Key Findings	Limitations & Future Work
R, Patel and P G (2025)	ANN model with Batch Normalization and Dropout to handle imbalance and overfitting	AUC-ROC: 0.904 (higher than XGBoost 0.734)	ANN significantly improves prediction performance and reduces overfitting	Validation on bigger and more varied datasets is required
Kaur et al. (2025)	Federated learning-based loan prediction model ensuring data privacy	ROC-AUC: 0.897	Achieves strong performance while preserving data privacy	Future work can improve scalability and communication efficiency
Gao (2024)	BP Neural Network and KNN for loan prediction	Accuracy >70%, Recall >55%, Precision >60%	BP Neural Network performs better than KNN	Performance can be enhanced with advanced deep learning models
C et al. (2024)	ML models including Decision Tree, Extra Trees, XGBoost, LightGBM	LightGBM Accuracy: 98.91%	LightGBM provides superior prediction accuracy	Requires testing on real-world large-scale financial data

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, January 2026
Impact Factor: 7.67

Nancy Deborah et al. (2023)	Loan prediction using ML algorithms including SVM	Accuracy: 83%	SVM shows moderate performance in loan prediction	Could explore ensemble or deep learning methods for improvement
Pathak et al. (2023)	Basic ML models and ensemble learning for loan default prediction	Ensemble Accuracy: 76.8%	Ensemble improves accuracy over individual models	Accuracy still moderate; needs feature optimization
Orji et al. (2022)	ML models implemented using Python on Kaggle environment	Random Forest: 95.55%, Logistic Regression: 80%	Tree-based models outperform linear models	Limited generalization across datasets
Tahmid et al. (2021)	Feature-based ML model comparison for credit recovery prediction	Accuracy: 89%	Certain ML models show strong predictive capability	Further work needed on feature engineering and model tuning

Research gaps: The existing loan default prediction method uses structured preprocessing, class balancing and ensemble learning models, but its primary prediction engine relies on traditional algorithms, which are unable to identify intricate nonlinear trends in borrower behavior. The model is based on single-source datasets of Lending Club, which restricts its usefulness in other financial contexts. Financial risk assessment process needs more advanced ML and DL systems with a higher predictive accuracy and system stability.

III. RESEARCH METHODOLOGY

The methodology involves classification of loan defaults systematically using Lending Club data on Kaggle with preprocessing of the data, one-hot encoding to handle categorical data and a normalizer to balance machine features. To reduce the imbalance between classes, SMOTE and a train-test split are used. Subsequently, ensemble learning algorithms such as RF and XGBoost are utilized and evaluated by performance indicators such as acc, prec, rec, F1, and ROC curve to effectively predict financial hazards. The flowchart proposed is in Fig. 1.
Each stage in suggested technique is thoroughly described in the section that follows:

A. Data Gathering and Analysis

The dataset used in this research is a historical collection of peer-to-peer loan data from Lending Club on Kaggle, borrower data, loan features, and repayment data. The dataset consists of about 887379 records having 75 features, which is very effective in credit risk and default prediction activities. Distribution analysis and correlation of features through data visualizations (bar plots and heatmaps) are used:

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, January 2026
Impact Factor: 7.67

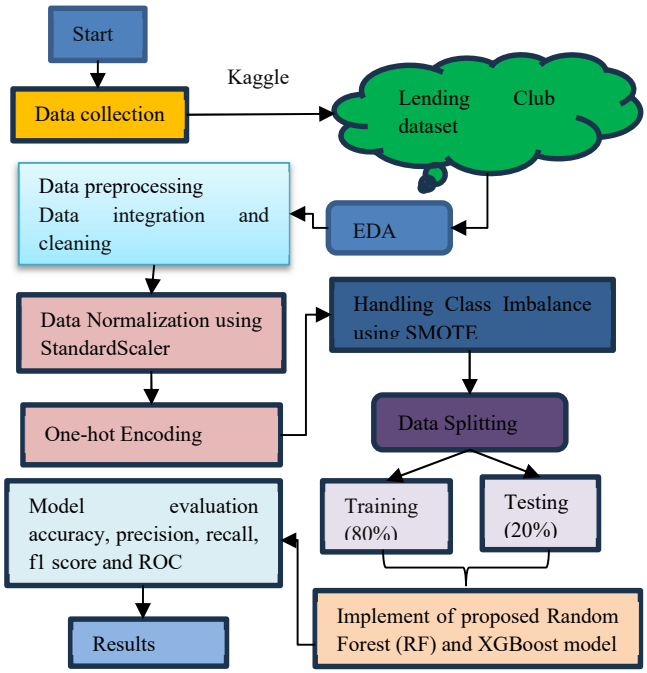

Fig. 1. Proposed flowchart for Loan Default Prediction for Financial Risk Analysis using machine learning.

The correlation heatmap in Fig. 2 demonstrates primary relationships between loan features, which show a strong positive relationship between loan amount and installment, with a correlation of (0.93), and between open accounts and total accounts, with a correlation of (0.76). Moderate correlations exist for interest rate and loan amount (0.57). Notable negative correlations include loan amount with revolving utilization (-0.71) and total accounts (-0.72). Most other features have weak correlations, indicating limited dependency.

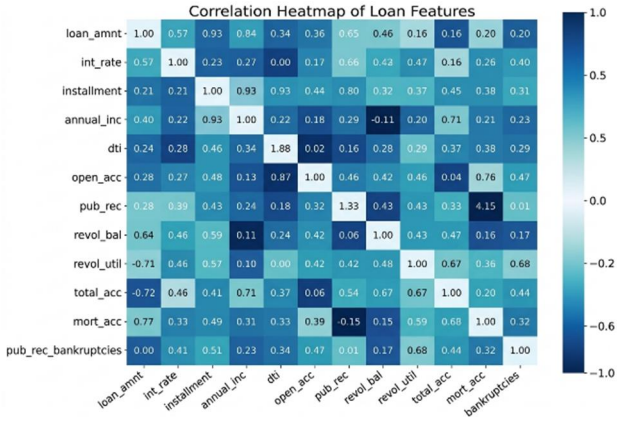

Fig. 2. Correlation Heatmap of Lending Club's dataset.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, January 2026
Impact Factor: 7.67

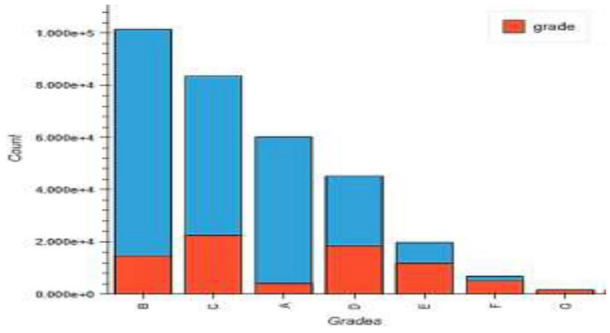

Fig. 3. Loan status by Grade.

Fig. 3 indicates that the loans are distributed according to the number of grades, whereby grades B and C represent largest volumes of loans, meaning that they are the most prevalent. Grades A to G have an evidently downward trend in the overall numbers, and this indicates that the higher-risk groups take a smaller share of the data.

B. Data Pre-Processing

Data preparation based on Lending Club dataset is done with data integration and cleaning. The preprocessing involved data normalization using data balancing, categorical variable encoding, and a StandardScaler. The main steps of preprocessing can be summarized as the following:

- **Data Normalization using Standard Scaler:** Some models may be impacted by characteristics in a dataset with a varied range. This problem is eliminated by using the normal scaler to normalize the features. The new number (n) is adjusted by scaling to fit a normal distribution, although the outliers could affect it. It is calculated using Equation (1).

$$n = \frac{n_i - n_{mean}}{\sigma} \qquad (1)$$

- **One-hot Encoding:** One-hot encoding is a procedure of converting a nominal variable into a dichotomous counterpart. It creates extra columns on each category, whereby a 1 states the category's existence, whereas 0 denotes its absence. The central concept of One Hot Encoding is to offer the possibility of successful utilization of categorical data in machine learning models.

C. Handling Class Imbalance using Synthetic Minority Over-sampling Technique (SMOTE)

The SMOTE method for dealing with class imbalance involves creating synthetic samples of minority class to achieve equal numbers of samples in both classes. This method improves model's capacity to correctly anticipate minority classes and lessens bias towards the majority class.

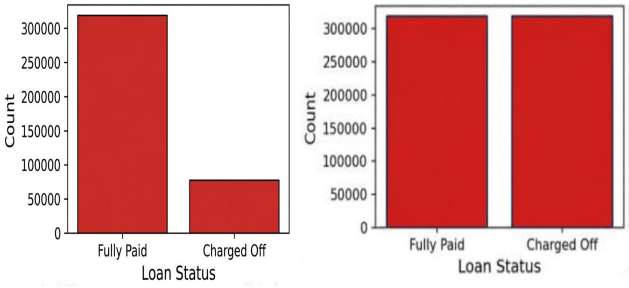

Fig. 4. Imbalanced and Balanced Class Distribution.

Fig. 4 indicates that class distribution is balanced as both the fully paid and charged off loan statuses have the same number of around 320,000, which removes the class imbalance. This even distribution leads to bias-free model training and improved performance metrics reliability in both classes.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, January 2026
Impact Factor: 7.67

D. Data Splitting

An 80:20 ratio was utilized to divide the data into a training and test set; 80% of the data was used to train the model, while 20% was kept to assess the model's performance.

E. Implement of proposed Models

The Extreme Gradient Boosting (XGBoost) and Random Forest (RF) models are designed to improve financial risk analysis in digital lending systems and forecast loan defaults with high accuracy.

1) Random Forest (RF) Model

To enhance the classification performance and minimize overfitting, using ensemble learning, the researchers' RF approach generates many DTs that are trained on various subsets of training data and feature sets. Since the system is a model of nonlinear relationships between multiple factors and can analyze extensive and complicated financial data, it is able to predict loan defaults with high accuracy.

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} T_i(x) \qquad (2)$$

In the Equation (2), $\hat{y}$ is final predicted output, N is total number of DT and $T_i(x)$is prediction of i-DT. The output of all trees (majority voting in favor of classification) is added to obtain the final decision, thus improving robustness and a high predictive performance in financial risk analysis.

The predictive model of loan default risk (financial risk assessment) is an RF model, whose hyperparameters are optimized to enhance forecasting accuracy of model without overfitting. The most important environment is 200 estimators (nestimators=200), a maximum split depth of 8 (max depth=8), a minimum split sample of 5 (min_samples_split=5), a minimum leaf sample of 2 (min_samples_leaf=2), a maximum number of features to consider in a split of 0.8 (max_features=0.8), and bootstrap on (bootstrap=True). Such settings enable the model to effectively classify borrowers and manage intricate data associations and possible imbalances in classes.

2) Extreme Gradient Boosting (XGBoost) Model

An enhanced ensemble learning technique based on GB is XGBoost model, whereby a large number of weak learners, DT, are built sequentially as the weaknesses of previous models are overcome. It is very effective in forecasting loan defaults because it optimizes the performance of the model by regularizing it, and because it can predict the performance of the model using large-scale financial data with high precision.

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)\,,\ f_k \epsilon F \qquad (3)$$

In this Equation (3), the symbol $\hat{y}_i$ denotes predicted output of i instance, K denotes number of trees, and $f_k(x_i)$ denotes each single decision tree function in the the space of functions $f_k \epsilon F$. The model combines the outcomes of the individual trees to come up with the final forecast.

$$L(\emptyset) = \sum_{i}^{n} l(y_i, \hat{y}_i) \sum_{k=1}^{K} \Omega f_k \qquad (4)$$

The loss term $(y_i, \hat{y}_i)$, which quantifies prediction error, and the regularization term Ωf_k, which regulates model complexity to prevent overfitting, make up this Equation (4) objective function. To predict loan default using the XGBoost model within the financial risk analysis, the XGBoost model is tuned with optimal hyperparameters, which are: learning rate of 0.05, maximum depth of 6, and 500 estimators, subsample and colsample by tree of 0.8, and regularization (reg_alpha=0.01, reg_lambda=1) to avoid overfitting.

F. Evaluation Metrics

This study focuses on five key performance evaluation metrics for comparing classifiers includes acc, prec, rec, F1, and ROC curve analysis. The confusion matrix evaluates efficacy of a classification model in predicting true classes by summarizing prediction results.

- **True Positive (TP):** This occurs when a positive outcome is predicted by the predictive model and it also turns out to be favorable.
- **False Positive (FP):** is when a desirable result is predicted by model, but an adverse occurrence actually occurs.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 4, January 2026 Impact Factor: 7.67

- **True Negative (TN):** is a negative prediction made by the model, and outcome is negative.
- **False Negative (FN):** Is model predicting negative, but true value of the result is positive.

1) Accuracy

The number of samples properly identified relative to accuracy is defined as the test dataset's total number of samples. Equation (5) was derived as follows-

$$Accuracy = \frac{TP+TN}{TP+Fp+TN+FN} \quad (5)$$

2) Precision

It evaluates how many accurate positive predictions a model has produced by contrasting them with the actual positive guesses, and accuracy is calculated using following Equation (6):

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

3) Recall

Recall is metric that quantifies proportion of TP instances that classifier accurately labels as positive. Recall is calculated as below in Equation (7):

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

4) F1 score

A balanced average of accuracy and recall is known as the F1-score, or balanced F1. Equation (8) illustrates it numerically:

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (8)$$

5) Receiver Operating Characteristic Curve (ROC)

ROC is a two-dimensional graph that shows how well a binary classification system works. Plotting TPR versus FPR at various thresholds shows the curve.

IV. RESULTS AND DISCUSSION

The environment for this study has been obtained, and a Core (TM) i7-1065G7 CPU running at 1.30GHz and 1.50GHz is used for the experiments. Furthermore, Python 3.7.1 is used since it provides a number of models and modules for classification.

A. Result Demonstrations

Table II presents a performance summary of results of classification of RF and XGBoost models for forecasting loan defaults from Lending Club data. The two models have very high performance, with XGBoost slightly outperforming RF with near-perfect scores of 99.99% in the acc, prec, rec, and F1. Such findings prove the suggested models to be very dependable when analyzing financial risks with XGBoost having a better predictive performance.

TABLE 2: CLASSIFICATION RESULTS OF PROPOSED MODELS LOAN DEFAULT PREDICTION FOR FINANCIAL RISK ANALYSIS USING LENDING CLUB'S DATASET

Matrix	RF	XGBoost
Accuracy	99.96	99.99
Precision	99.98	99.99
Recall	99.97	99.99
F1-score	99.98	99.99

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, January 2026
Impact Factor: 7.67

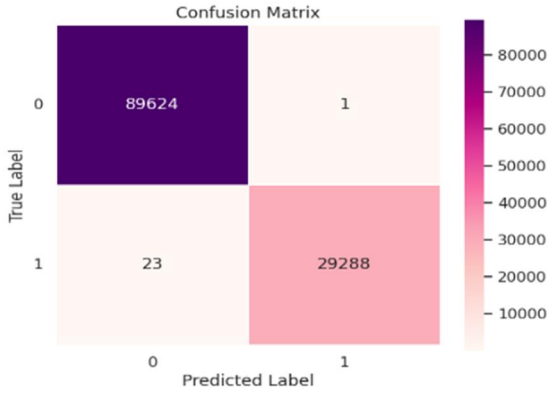

Fig. 5. Confusion Matrix for the Random Forest Model.

Fig. 5 illustrates a confusion matrix for the RF model, indicating excellent classification performance with 89,624 TN and 29,288 TP, and minimal misclassifications (1 FP and 23 FN). The model's high prec and rec, as evidenced by strong diagonal dominance and a pink-purple gradient, demonstrate that it provides highly accurate predictions for both classes. The confusion matrix seen in Fig. 6 illustrates the XGBoost model, which achieves near-perfect acc with 89,623 TN and 29,310 TP, and 2 FP and 1 FN. The high diagonal concentration and pink-to-purple gradient indicate high precision and recall in the model, confirming highly accurate predictions with very low error rates across both classes.

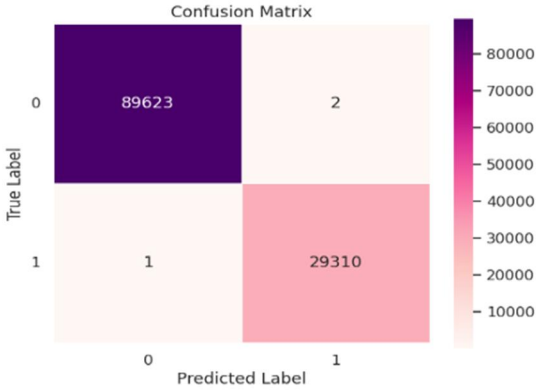

Fig. 6. Confusion Matrix for the XGBoost Model.

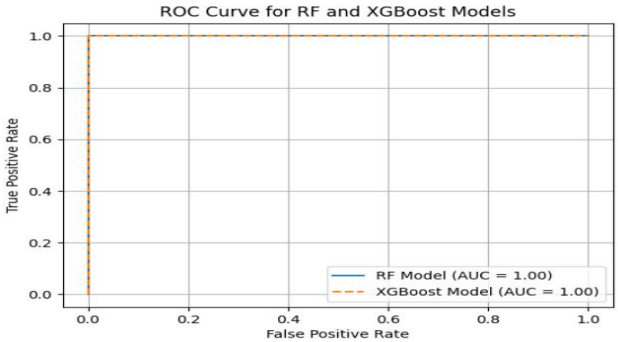

Fig. 7. ROC Curve for Proposed Models.

Fig. 7 demonstrates the ROC curve of proposed RF and XGBoost, where both models have a perfect AUC of 1.00, with the curve in both being well aligned along the upper-left-hand corner, which represents an ideal performance in classification. Such overlap establishes the fact that both models have perfect discriminative capacity and thus attain the TPR of 100% and a FPR of zero at all the thresholds.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, January 2026
Impact Factor: 7.67

B. Comparative Analysis

To evaluate efficiency of suggested RF and XGBoost models using dataset, a comparative assessment is presented in Table III based on accuracy estimates. The RF model is 99.96% accurate, while the XGBoost model is 99.99% accurate, which is much better than the SVM, DT, and LR models. This illustrates that ensemble models have a better performance and dependability in forecasting loan failure and financial risks.

TABLE 3: COMPARISON OF DIFFERENT MACHINE LEARNING MODELS FOR LOAN DEFAULT PREDICTION FOR FINANCIAL RISK ANALYSIS

Model	Accuracy	Precision	Recall	F1-score
SVM[25]	73.1	94.7	66	88.2
DT[26]	86.1	63.7	85.9	74.7
LR[27]	87.06	66.11	85.91	74.72
RF	99.96	99.98	99.97	99.98
XGBoost	99.99	99.99	99.99	99.99

Table IV is the comparative performance of financial risk assessment models on various datasets, as LightGBM, Neural Network (NN), Deep Neural Network (DNN), Random Forest (RF), and XGBoost. It shows that the models that were trained on Lending Club dataset, in particular, RF (99.96% accuracy) and XGBoost (99.99% accuracy), are significantly more effective than those used on other datasets, such as Credit Risk and Credit Card Customer datasets. The research proves that ensemble models can deliver effective results based on high-quality datasets that can provide the best financial risk assessment performance.

TABLE 4: COMPARISON PERFORMANCE OF FINANCIAL RISK ASSESSMENT MODELS USING DIFFERENT DATASETS

Model	Dataset	Acc.	Pre.	Rec.	F1
LightGBM[28]	Credit Risk dataset	79.6	78.9	76.2	77.6
NN[29]	Credit Card Customer Dataset	87.2	87.2	87.2	87.2
DNN[30]	Loan Risk Dataset	94	94	94	94
RF	Lending Club's dataset	99.96	99.98	99.97	99.98
XGBoost		99.99	99.99	99.99	99.99

C. Discussion

The findings indicate that ensemble models, including RF and XGBoost, perform better at predicting loan defaults than traditional models and DL systems. The evaluation metrics provide precise, stable results during testing due to their strong financial risk analysis capabilities. Research shows that proper model selection, combined with the quality of dataset use, is a very important criterion for increasing predictive performance.

V. CONCLUSION AND FUTURE STUDY

Credit risk, which is also called loan default, is a persistent problem for financial organizations and banks due to the resultant uncertainty about whether borrowers will be able to service their debts. Financial institutions need to reduce the possible losses, so they use statistical techniques and ML algorithms to better predict loan defaults, and their predictive algorithms are better when based on more detailed semantic information in their data. The results of the experiment indicate that ensemble learning models are more effective than traditional ones: SVM, DT, and LR achieved accuracies between 73.1% and 87.06%, whereas RF and XGBoost achieved an acc of 99.96% and 99.99%, respectively. The findings indicate that ensemble methods are useful to process complicated financial data and improve the accuracy of risk prediction. The future study should integrate the DL models with real-time financial data and behavioral data to enhance accuracy of prediction and strength of the model.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, January 2026
Impact Factor: 7.67

REFERENCES

[1] A. Ghosh, "Banking sector globalization and bank performance: A comparative analysis of low income countries with emerging markets and advanced economies," *Rev. Dev. Financ.*, vol. 6, no. 1, pp. 58–70, Jun. 2016, doi: 10.1016/j.rdf.2016.05.003.

[2] M. G. Kavussanos and D. A. Tsouknidis, "Default risk drivers in shipping bank loans," *Transp. Res. Part E Logist. Transp. Rev.*, vol. 94, pp. 71–94, Oct. 2016, doi: 10.1016/j.tre.2016.07.008.

[3] N. Radhasharan, "Real-Time AI and Data Transparency in Financial Services: A New Era of Trust and Liquidity Optimization," *J. Comput. Anal. Appl.*, vol. 35, no. 1, Jan. 2026, doi: 10.48047/jocaaa.2026.35.01.18.

[4] K. Ramkumar, R. R. Sagunthala, A. Professor, A. Nerella, S. Kilaru, and G. Gladson Battu, "A Temporal Graph Neural Network Approach for Deep Fraud Detection in Real-Time Financial Transactions," in *16th International IEEE Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2025, pp. 1–5.

[5] C. Curi and A. Lozano-Vivas, "Probability of Default and Banking Efficiency: How Does the Market Respond?," 2020, pp. 209–220. doi: 10.1007/978-3-030-41618-8_13.

[6] H. P. Cyril, "Serialization of Telecom Provisioning Transactions in Distributed Systems," *Int. J. Eng. Adv. Technol. Stud.*, vol. 15, no. 6, pp. 526–533, 2025, doi: 10.14741/ijcet/v.15.6.6.

[7] S. B. Shah, "Evaluating the Effectiveness of Machine Learning in Forecasting Financial Market Trends: A Fintech Perspective," in *2025 3rd International Conference on Integrated Circuits and Communication Systems (ICICACS)*, 2025, pp. 1–6. doi: 10.1109/ICICACS65178.2025.10968297.

[8] M. Madaan, A. Kumar, C. Keshri, R. Jain, and P. Nagrath, "Loan default prediction using decision trees and random forest: A comparative study," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1022, no. 1, 2021, doi: 10.1088/1757-899X/1022/1/012042.

[9] A. I. Marqués, V. García, and J. S. Sánchez, "Exploring the behaviour of base classifiers in credit scoring ensembles," *Expert Syst. Appl.*, vol. 39, no. 11, pp. 10244–10250, Sep. 2012, doi: 10.1016/j.eswa.2012.02.092.

[10] J. Han, J. Choi, M. Kim, and J. Jeong, "Developing a Risk Group Predictive Model for Korean Students Falling into Bad Debt," *Asian Econ. J.*, vol. 32, no. 1, pp. 3–14, Mar. 2018, doi: 10.1111/asej.12139.

[11] S. Kakkar, "Explainable AI Models for Credit Risk Scoring in Banking: Balancing Accuracy and Regulatory Transparency," *Int. J. Financ. Data Sci.*, vol. 3, no. 2, pp. 1–6, Aug. 2025, doi: 10.34218/IJFDS_03_02_001.

[12] X. Ma, J. Sha, D. Wang, Y. Yu, Q. Yang, and X. Niu, "Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning," *Electron. Commer. Res. Appl.*, vol. 31, pp. 24–39, Sep. 2018, doi: 10.1016/j.elerap.2018.08.002.

[13] H. Kohli, R. Rajamani, M. R. Reddy Deva, and S. A. Pahune, "Improving Prediction of Credit Risks Based on Advanced Machine Learning and Feature Engineering Techniques in Banking Sector," in *2025 5th International Conference on Artificial Intelligence and Signal Processing (AISP)*, IEEE, Nov. 2025, pp. 1–5. doi: 10.1109/AISP68263.2025.11396116.

[14] S. H. D. Kolagani, M. Bhandar, and R. Altounjy, "Integrating AI Predictive Analytics into Financial CRM for Retention," in *2025 7th International Conference on Innovative Data Communication Technologies and Application (ICIDCA)*, IEEE, Oct. 2025, pp. 1508–1513. doi: 10.1109/ICIDCA66325.2025.11280429.

[15] S. Z. H. Shoumo, M. I. M. Dhruba, S. Hossain, N. H. Ghani, H. Arif, and S. Islam, "Application of Machine Learning in Credit Risk Assessment: A Prelude to Smart Banking," in *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)*, IEEE, Oct. 2019, pp. 2023–2028. doi: 10.1109/TENCON.2019.8929527.

[16] F. Shen, X. Zhao, Z. Li, K. Li, and Z. Meng, "A novel ensemble classification model based on neural networks and a classifier optimisation technique for imbalanced credit risk evaluation," *Phys. A Stat. Mech. its Appl.*, vol. 526, p. 121073, Jul. 2019, doi: 10.1016/j.physa.2019.121073.

[17] M. G. Kavussanos and D. A. Tsouknidis, "The determinants of credit spreads changes in global shipping bonds," *Transp. Res. Part E Logist. Transp. Rev.*, vol. 70, pp. 55–75, Oct. 2014, doi: 10.1016/j.tre.2014.06.001.

[18] H. R, M. S. Patel, and O. P. P G, "Predictive Analysis for Loan Defaults: A Deep Learning Approach," in *2025*

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 4, January 2026 Impact Factor: 7.67

Global Conference in Emerging Technology (GINOTECH), IEEE, May 2025, pp. 1–5. doi: 10.1109/GINOTECH63460.2025.11076632.

[19] G. Kaur, H. Kaur, G. Ranjana Panigrahi, and N. Shelke, "Quantum-Enhanced Predictive Analytics for Loan Default and Repayment Behavior," in *2025 International Conference on Innovations and Emerging Technologies In AI & Communication Systems (IETACS)*, IEEE, Nov. 2025, pp. 608–613. doi: 10.1109/IETACS68750.2025.11385671.

[20] B. Gao, "Financial Loan Default Risk Prediction Based on Big Data Analysis," in *2024 IEEE 4th International Conference on Power, Electronics and Computer Applications (ICPECA)*, IEEE, Jan. 2024, pp. 398–403. doi: 10.1109/ICPECA60615.2024.10471076.

[21] S. K. C, M. S, P. M. Reddy, and K. Gopal, "Analyzing the Performance of Ensemble Machine Learning Algorithms for Predicting Loan Eligibility," in *2024 9th International Conference on Communication and Electronics Systems (ICCES)*, 2024, pp. 1362–1367. doi: 10.1109/ICCES63552.2024.10859945.

[22] R. N. Deborah, S. A. Rajiv, A. Vinora, C. M. Devi, and S. M. Arif, "An Efficient Loan Approval Status Prediction Using Machine Learning," in *Proceedings of 3rd International Conference on Advanced Computing Technologies and Applications, ICACTA 2023*, 2023. doi: 10.1109/ICACTA58201.2023.10392691.

[23] P. Pathak, A. Jain, M. Bansal, and P. S. Rana, "SentiNet: Empowering Robust Loan Default Prediction through Ensemble Modeling," in *2023 IEEE International Conference on Computer Vision and Machine Intelligence (CVMI)*, 2023, pp. 1–6. doi: 10.1109/CVMI59935.2023.10464518.

[24] U. E. Orji, C. H. Ugwuishiwu, J. C. N. Nguemaleu, and P. N. Ugwuanyi, "Machine Learning Models for Predicting Bank Loan Eligibility," in *Proceedings of the 2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development, NIGERCON 2022*, 2022. doi: 10.1109/NIGERCON54645.2022.9803172.

[25] A. Akinjole, O. Shobayo, J. Popoola, O. Okoyeigbo, and B. Ogunleye, "Ensemble-Based Machine Learning Algorithm for Loan Default Risk Prediction," *Mathematics*, vol. 12, no. 21, p. 3423, Oct. 2024, doi: 10.3390/math12213423.

[26] R. Yang, "Machine Learning-Based Loan Default Prediction in Peer-to-Peer Lending," *Highlights Sci. Eng. Technol.*, vol. 94, pp. 310–318, 2024, doi: 10.54097/qdjd8r65.

[27] N. Upadhyay, "Machine Learning-Based Default Loan Prediction for Financial Risk Assessment in Digital Lending," no. 2, pp. 9–16, 2026.

[28] X. Zhang *et al.*, "Data-Driven Loan Default Prediction: A Machine Learning Approach for Enhancing Business Process Management," *Systems*, vol. 13, no. 7, p. 581, Jul. 2025, doi: 10.3390/systems13070581.

[29] V. Chang, S. Sivakulasingam, H. Wang, S. T. Wong, M. A. Ganatra, and J. Luo, "Credit Risk Prediction Using Machine Learning and Deep Learning: A Study on Credit Card Customers," *Risks*, vol. 12, no. 11, p. 174, Nov. 2024, doi: 10.3390/risks12110174.

[30] G. Liu, "Research on Personal Loan Default Risk Assessment Based on Machine Learning," *ITM Web Conf.*, vol. 70, p. 01012, Jan. 2025, doi: 10.1051/itmconf/20257001012.