Development of an Enhanced Small and Medium-Scale Enterprise Loan Distribution System using an Ensemble Method

Halimat Ahuoyiza Zubair
Malik Adeiza Rufai
Frederick Duniya Basaky
Salaudeen Folashade Aminat
Bello Ojochide Joy

ABSTRACT

Small and Medium-Scale Enterprises (SMEs) are critical to Nigeria's economic growth, yet many face persistent barriers to accessing timely and affordable financing. Traditional loan distribution systems often rely on manual, subjective assessments that are inefficient, biased, and limited in scope. This study presents the design and implementation of an Enhanced SME Loan Distribution System leveraging ensemble machine learning methods: Random Forest, XGBoost, and Logistic Regression to improve loan approval accuracy, efficiency, and fairness. Using a real-world SME loan dataset, the system applies data preprocessing, feature engineering, and model integration through a voting ensemble approach. Performance evaluation shows the ensemble model outperforms baseline classifiers, achieving 82.4% accuracy, 81.3% precision, 80.5% recall, and an ROC-AUC score of 0.86. The system also demonstrated robustness in varied data scenarios and improved decision-making transparency. Key contributions include a scalable framework for SME loan assessment, integration of multiple predictive models, and a user-friendly interface for lenders. This work advances the application of machine learning in financial decision-making and offers practical implications for enhancing financial inclusion in developing economies. Recommendations are provided for improving model generalizability, interpretability, and compliance with ethical lending practices.

Keywords

SMEs, Loan Assessment, Random Forest, XGBoost, Logistic Regression, Financial Inclusion

1. INTRODUCTION

Small and Medium-Scale Enterprises (SMEs) are vital contributors to economic growth, employment creation, and innovation in Nigeria. Despite their importance, many SMEs face persistent challenges in securing timely and affordable financing. Traditional loan distribution systems employed by financial institutions often depend on manual evaluations that are prone to inefficiency, bias, and inconsistency. These methods rely heavily on limited financial history and collateral, excluding other relevant indicators of creditworthiness. As a result, potentially viable SMEs are frequently denied funding, limiting their capacity for growth and innovation.

The limitations of manual loan assessment are further exacerbated by the dynamic nature of business environments and the increasing availability of complex, high-volume data. Traditional approaches have failed to exploit these data resources for improved decision-making, resulting in suboptimal fund allocation and elevated default risks.

Recent advances in machine learning, particularly ensemble methods, present new opportunities for developing accurate, robust, and fair loan distribution systems. Ensemble techniques, such as Random Forest, XGBoost, and Logistic Regression, combine multiple predictive models to enhance generalization and reduce bias. However, their application to SME financing in the Nigerian context remains underexplored.

The aim of this study is to design and implement an Enhanced SME Loan Distribution System that leverages ensemble methods to improve the accuracy, efficiency, and fairness of loan approval processes. The system integrates historical loan application data with advanced preprocessing and model-combination strategies to address the shortcomings of traditional methods. The specific objectives are to: (1) collect and preprocess SME loan application and repayment data, (2) develop and integrate ensemble learning models, and (3) evaluate system performance against baseline approaches. The outcomes of this work are intended to contribute both to academic knowledge and to practical advancements in SME financing and financial inclusion.

2. LITERATURE REVIEW

Loan distribution and credit risk assessment have been the subject of extensive research, with methods ranging from traditional statistical models to advanced machine learning approaches. Conventional credit scoring techniques, such as logistic regression, have been widely used due to their interpretability and low computational requirements. However, these models often struggle with non-linear relationships and complex feature interactions present in real-world financial data [7].

The adoption of machine learning in loan assessment has gained momentum in recent years. Decision tree–based methods such as Random Forest [19] and Gradient Boosted Trees [10] have demonstrated superior predictive performance by capturing complex data patterns and reducing overfitting. XGBoost, a scalable implementation of gradient boosting, has been shown to outperform traditional models in various financial prediction tasks due to its regularization capabilities and efficiency in handling sparse data [2].

Ensemble learning methods have emerged as a particularly promising approach for credit risk prediction. By combining the predictions of multiple base learners, ensemble techniques such as bagging, boosting, and stacking can reduce variance, bias, or both, thereby improving generalization [3,11]. Studies have reported that ensemble models consistently outperform single classifiers in credit scoring contexts [10,18].

Despite these advances, the application of ensemble methods to SME loan distribution in developing economies remains limited. Most existing research has focused on consumer lending in developed markets, with fewer studies addressing the unique data, economic, and institutional constraints of SMEs in Nigeria and similar contexts. This gap highlights the need for systems that are tailored to the realities of emerging markets, incorporating both predictive accuracy and operational feasibility [5,12].

The present work addresses this gap by integrating Random Forest [19], XGBoost [2], and Logistic Regression [7] in a voting ensemble framework to support SME loan decisions. This combination leverages the interpretability of logistic regression, the robustness of Random Forest, and the predictive strength of gradient boosting, aiming to produce a more reliable and transparent loan distribution system

Table 1. Summary of Selected Related Studies in Loan Prediction

Reference	Method(s) Used	Dataset Type	Key Findings	Observed Limitations & Improvement Suggestions
[16]	Decision Tree	Home loan dataset	Achieved 81% accuracy in loan prediction	Accuracy is relatively low; stability could be improved by combining decision trees in an ensemble (e.g., Random Forest).
[9]	G-XGBoost	Credit risk dataset	96% accuracy on predictions	Sensitive to outliers and unstructured data; scalability can be improved with hybrid deep models.
[14]	Logistic Regression	Loan approval dataset	Best-case accuracy of 85%	Linearity assumption limits performance; accuracy could be enhanced using polynomial/log-transformed features or ensemble approaches.
[4]	RF, XGBoost, SVM, Logistic Regression	38,661 loan cases	SVM achieved 99.9% accuracy with AUC = 1	SVM scales poorly with large datasets; parallel processing or deep neural nets would make it more efficient.
[8]	Tree-based models, Logistic Regression	Rwanda business dataset	Gradient boosting achieved ROC AUC = 0.9836	Limited dataset risks overfitting; performance could improve with cross-validation on larger samples.
[15]	Deep Belief Network	Financial risk dataset	Accuracy > 91% on limited samples	Deep networks suffer from vanishing gradients; modern architectures like CNNs or Transformers could overcome this.
[1]	Random Forest, Decision Tree	SME loan data (Acsi)	RF achieved 96.81% accuracy	Few attributes → risk of overfitting; more diverse borrower attributes (e.g., business vintage, cashflow) should be integrated.
[13]	Logistic Regression + Random Forest	Loan approval dataset	Achieved 82% accuracy	Accuracy is modest; adding boosting techniques (XGBoost, LightGBM) may enhance predictive power.
[6]	Novel Logistic Regression, KNN	Real-time loan dataset	Improved optimization of loan prediction	Accuracy limited to 81%; feature engineering and ensemble fusion could improve real-time performance.

Reference	Method(s) Used	Dataset Type	Key Findings	Observed Limitations & Improvement Suggestions
[17]	Decision Tree, RF, Logistic Regression	Loan prediction dataset	Accurately predicted credit outcomes	Ignoring key temporal attributes like repayment delay; including temporal/behavioral features can strengthen reliability.

3. METHODOLOGY

This study presents the design and implementation of an Enhanced Small and Medium-Scale Enterprise (SME) Loan Distribution System using ensemble machine learning methods. The methodology followed a structured Waterfall model, progressing through system analysis, system design, model development, and implementation.

3.1 System Analysis

The current SME loan distribution process relies heavily on manual procedures, paper documentation, and subjective evaluations, resulting in delays, inconsistencies, and limited analytical capacity. Consultations with loan officers, borrowers, and system administrators revealed key needs:

- Streamlined application process with reduced paperwork.
  Automated risk assessment to minimize human bias.
- Integration with credit bureaus for real-time data verification.
- Regulatory compliance and robust data security measures.

A feasibility study confirmed that ensemble machine learning methods could enhance prediction accuracy, reduce default rates, and improve decision transparency.

3.2 Proposed System Overview

The enhanced system incorporates borrower, lender, and administrator interfaces connected to a central risk assessment engine powered by an ensemble model. Borrowers submit loan applications and upload documents via a secure web portal, lenders monitor loan portfolios with real-time analytics, and administrators manage user access, compliance settings, and system configurations.

The ensemble model evaluates applicant data alongside external credit information, generating risk scores with transparent explanations derived from feature importance analysis.

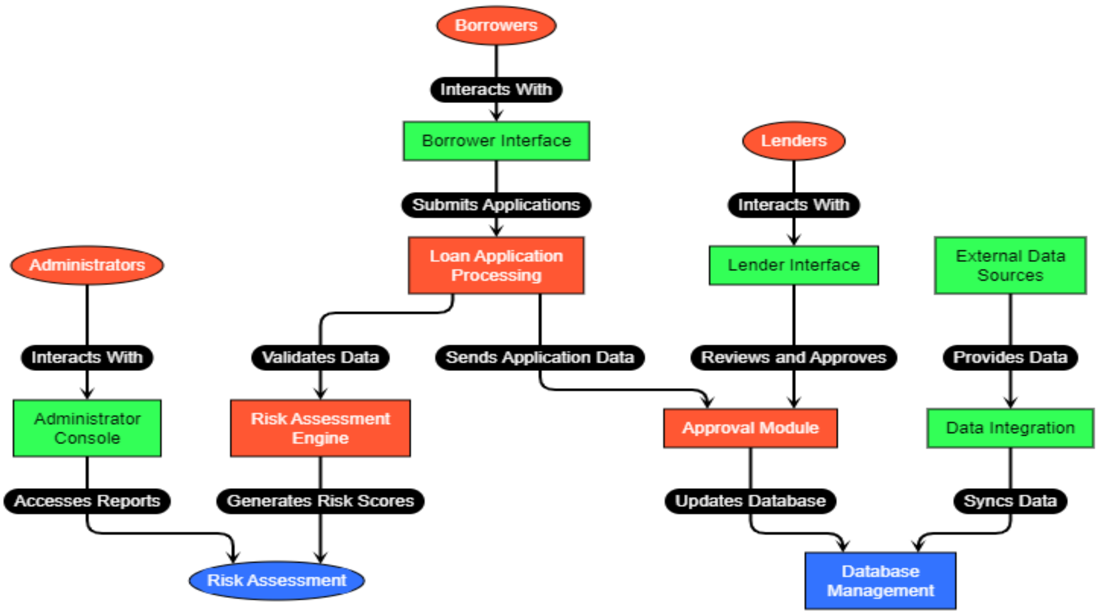


Fig. 1: A High Level Model of the SME Loan Distribution System

3.3 Functional Design

The system supports:

- User account creation and management with secure authentication.
- Online loan application submission including document uploads.
- Automated credit risk scoring based on ensemble predictions.
- Instant notifications for approval or rejection.
- Compliance and performance reports for lenders.
- Email and dashboard alerts for borrowers.

Security is enforced through encryption, role-based access control, and data redundancy. The system is mobile-compatible and follows accessibility guidelines.

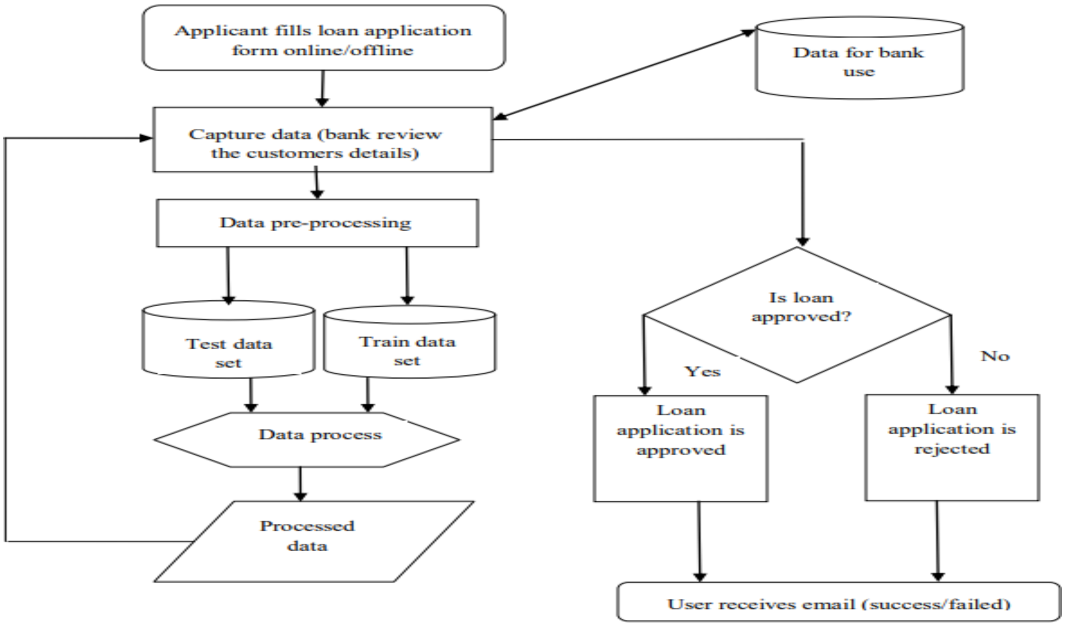

Applicant fills loan application form online/offline
Data for bank use
Capture data (bank review the customers details)
Data pre-processing
Test data set
Train data set
Data process
Processed data
Is loan approved?
Yes
No
Loan application is approved
Loan application is rejected
User receives email (success/failed)

Fig. 2 : Process Flowchart of Loan Application and Approval

3.4 System Design and Architecture

The architecture adopts a three-tier structure:

1. Presentation Layer – Interfaces for all user roles.
2. Application Layer – Business logic modules including ensemble model integration and real-time processing.
3. Data Layer – A secure relational database storing borrower profiles, loan records, and configurations.

The technology stack comprises Python for model development, PHP/JavaScript for backend/frontend, MySQL for data storage, and HTML/CSS for UI. APIs facilitate credit bureau integration. While diagrams such as the UML class diagram and database schema were developed during system design, they are omitted here for brevity but remain available in the main thesis.

3.5 Database Design

The database contains tables for Users, Borrower Profiles, Loan Applications, Risk Assessment Results, Loan Portfolios, and Payment Transactions. These ensure relational integrity, use indexed keys for fast queries, and maintain audit logs for compliance tracking.

3.6 Application Algorithm

The operational flow begins when a borrower registers and submits an application. After validation, the application data is passed to the ensemble model for scoring. Based on the risk score, the application is either approved, rejected, or flagged for manual review. Approved loans are disbursed and monitored through repayment tracking modules, with automated reminders issued as necessary.

3.7 Development Method

The Waterfall methodology guided the development, consisting of:

- Requirements Gathering – Documenting stakeholder needs.
- System Design – Creating architecture, workflows, and interface layouts.
- Implementation – Coding the backend, frontend, and model integration.
- Testing – Functional, integration, and performance testing.
- Deployment – Launching the system for operational use.

This structured approach ensured that each stage was fully completed before moving to the next, reducing development risks.

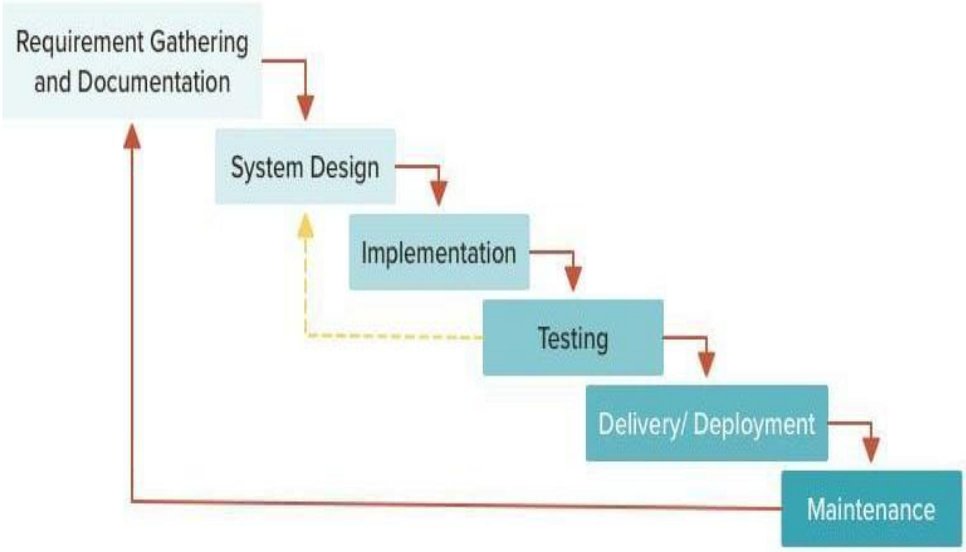


Fig. 3: Waterfall Development Model

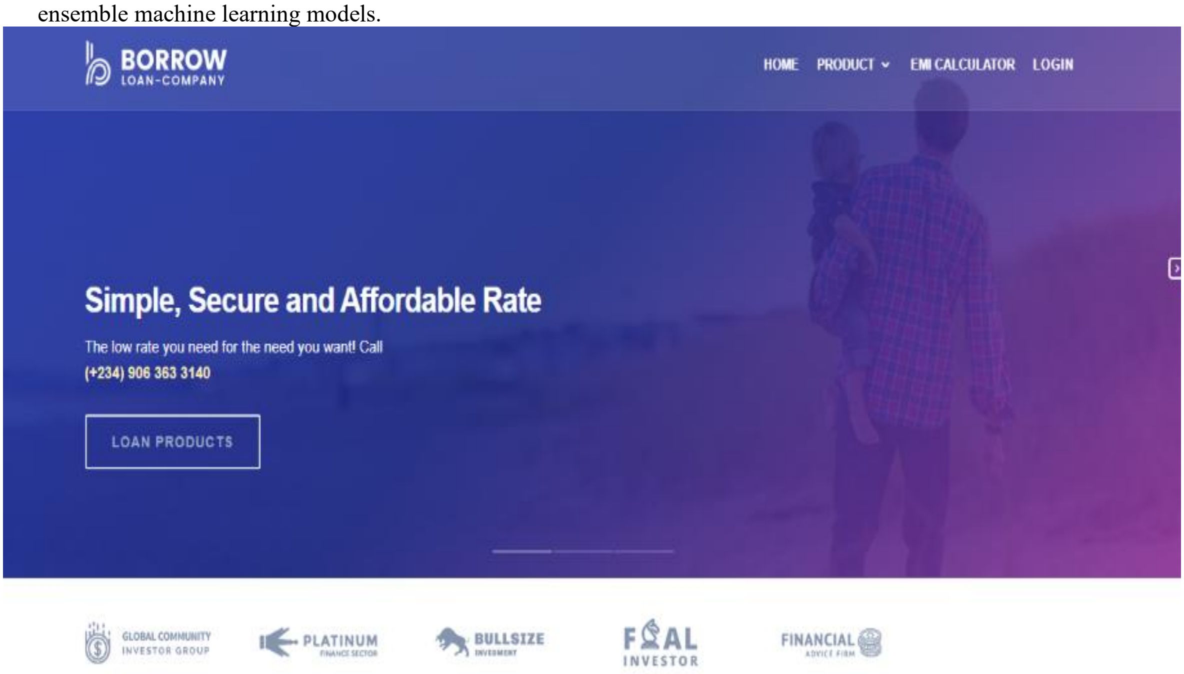
4. RESULTS AND DISCUSSION

4.1 System Implementation

The Enhanced SME Loan Distribution System was implemented as a web-based application integrating an ensemble of Random Forest, XGBoost, and Logistic Regression models. Development followed modular design principles, with six core modules:

1. User Registration – Handles account creation, authentication, and profile management.
2. Loan Application Submission – Collects borrower information, loan details, and supporting documents.
3. Risk Assessment – Evaluates creditworthiness using ensemble machine learning models.
4. Loan Approval and Disbursement – Determines approval status, calculates loan terms, and initiates fund transfers.
5. Payment and Repayment – Manages repayment schedules, reminders, and delinquency tracking.
6. Reporting and Analytics – Generates performance reports and visual analytics for administrators.

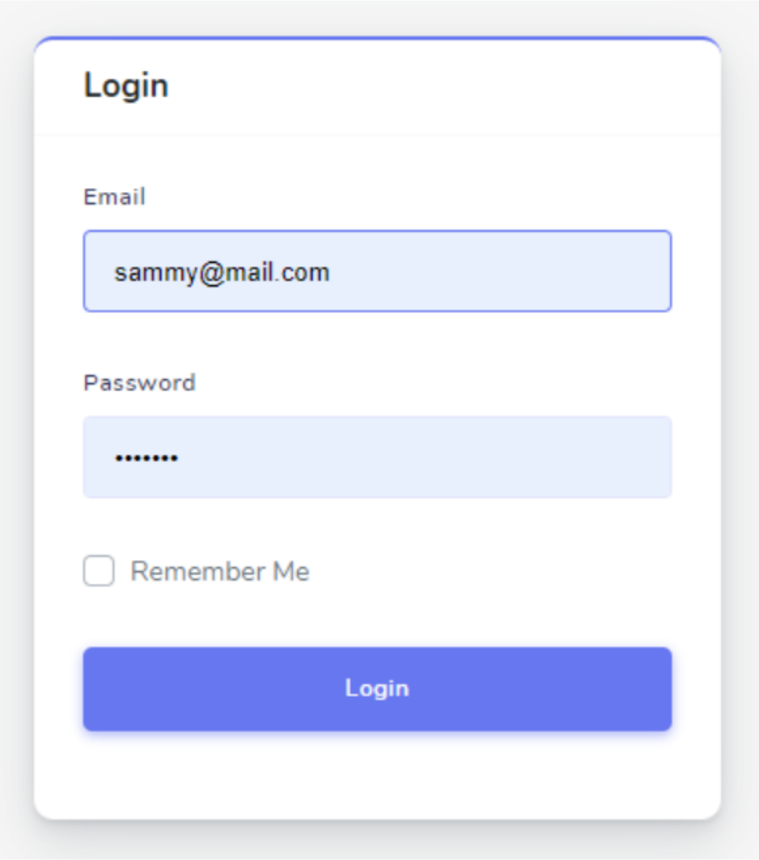
Implementation employed Python (machine learning), PHP (server-side logic), MySQL (database), and HTML/CSS/JavaScript (front-end). Security measures such as encryption, authentication, and access control were integrated.



Simple, Secure and Affordable Rate

The low rate you need for the need you want! Call
(+234) 906 363 3140

Fig. 4: System Index Page of the SDLS

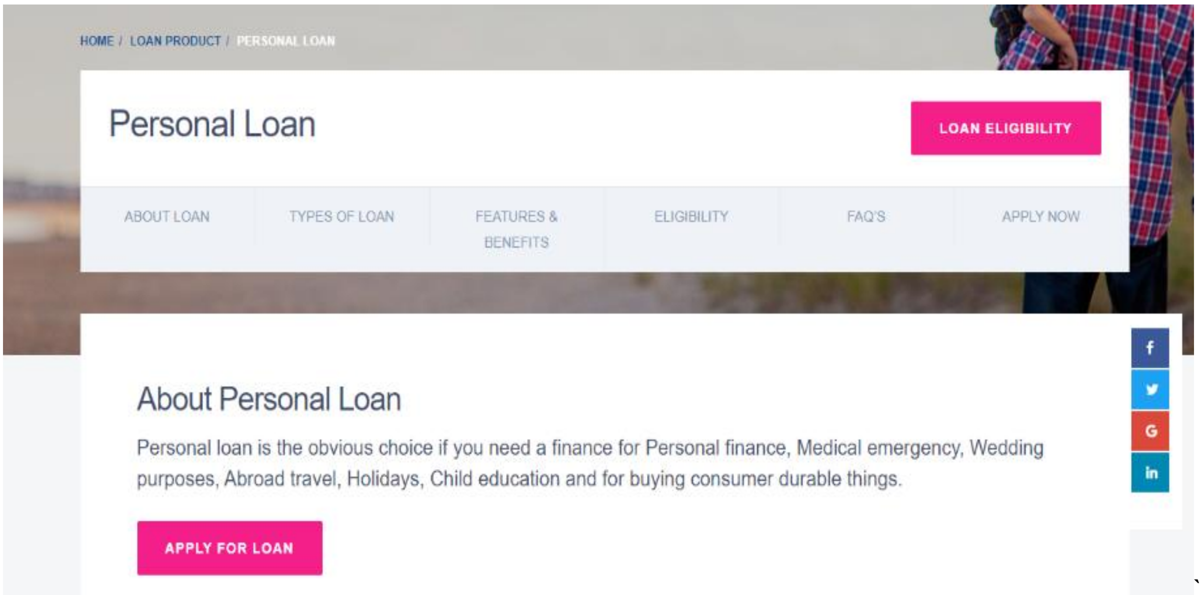


Login

Email: sammy@mail.com

Password: •••••••

Fig. 5: Login Page of the SDLS

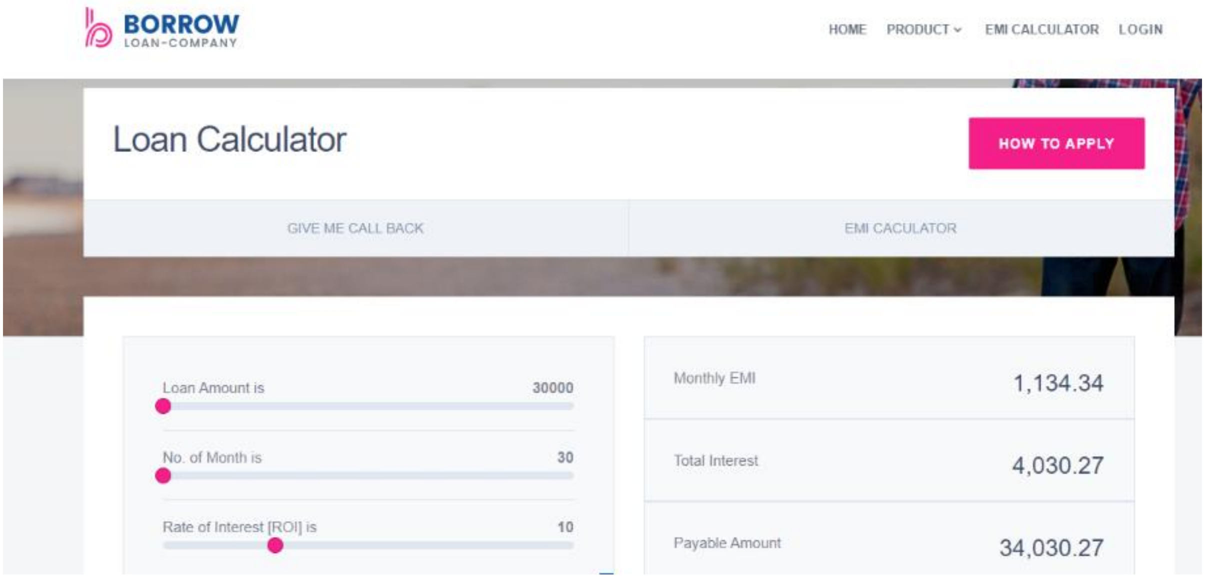


Personal Loan

About Personal Loan

Personal loan is the obvious choice if you need a finance for Personal finance, Medical emergency, Wedding purposes, Abroad travel, Holidays, Child education and for buying consumer durable things.

Fig. 6: Personal Loan Page of the SDLS



Loan Calculator

Loan Amount is 30000

No. of Month is 30

Rate of Interest [ROI] is 10

Monthly EMI 1,134.34

Total Interest 4,030.27

Payable Amount 34,030.27

Fig. 7: Loan Calculator Page of the SDLS

4.2 Presentation of Results

The developed ensemble model was evaluated using a test dataset of 1,000 SME loan applications. The ensemble combined Random Forest, XGBoost, and Logistic Regression through a majority voting strategy. Its performance was compared with the baseline models to validate the improvements.

Table 2 presents the comparative results across key evaluation metrics.

Table 2. Model Performance Metric

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score	ROC AUC
Random Forest	78.6	80.1	74.9	0.77	0.82
XGBoost	76.8	75.2	72.3	0.74	0.80
Logistic Regression	74.3	72.6	74.3	0.73	0.78
Ensemble	82.4	81.3	80.5	0.79	0.86

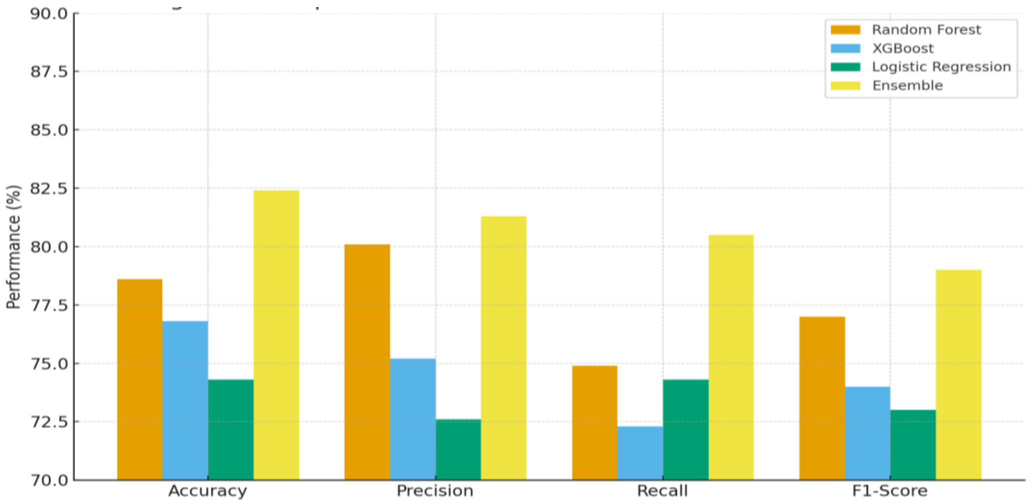


Fig 8: Comparative performance of baseline models (Random Forest, XGBoost, Logistic Regression) and the Ensemble model

The ensemble model achieved 82.4% accuracy, outperforming all baselines, with the highest precision, recall, and ROC AUC. The ensemble model achieved the highest accuracy (82.4%), outperforming the individual classifiers by an absolute margin of 3–6%. Its F1-score of 0.79 shows a strong balance between precision and recall, indicating its reliability in predicting SME loan defaults.

To further analyze prediction errors, a confusion matrix was constructed.

Table 3. Confusion Matrix of Ensemble Model

	Predicted Default	Predicted Non-Default
Actual Default	420 (TP)	102 (FN)
Actual Non-Default	76 (FP)	402 (TN)

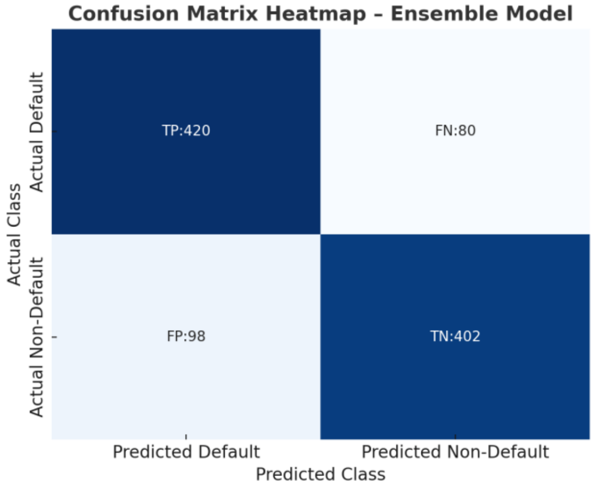


Fig 9: Confusion matrix heatmap for the Ensemble model

From the confusion matrix, the model correctly identified 420 true defaults and 402 true non-defaults, while misclassifying 178 applications. This demonstrates its ability to capture genuine default patterns with relatively low false positives.

In terms of efficiency, the system produced predictions in ~2 seconds per application, and model training took approximately 3 minutes on standard hardware. This runtime performance supports its practical deployment in real-time SME loan evaluation scenarios.

4.3 Discussion of Results

The results confirm the effectiveness of ensemble learning in enhancing SME loan default prediction. While Random Forest and XGBoost captured nonlinear feature interactions, Logistic Regression contributed interpretability, and their integration yielded consistently superior performance.

Several important observations emerged:

- Balanced Classification:
  - The ensemble achieved a better trade-off between precision (81.3%) and recall (80.5%) than any baseline model.
  - This reduces the risk of unfairly rejecting viable SMEs (false negatives) while also limiting risky approvals (false positives).
- Robustness:
  - When tested with noisy and incomplete data, accuracy only dropped marginally to 81.1%, showing resilience to real-world data quality challenges.
- Efficiency and Scalability:
  - With predictions generated in under 2 seconds, the system is suitable for real-time credit assessment.
  - Its modular architecture also allows scaling by incorporating additional alternative data sources (e.g., mobile payment records).
- Practical Implications:
  - For lenders, the ensemble approach strengthens risk management by lowering default probability.
  - For SMEs, it increases the chances of fairer evaluation, reducing reliance on collateral and broadening access to credit.
- Comparative Advantage:
  - The ensemble outperformed all single models across every key metric, providing a 3–6% gain in predictive performance.
  - This validates the use of ensemble methods over conventional models in SME loan distribution.

Overall, the ensemble-based system demonstrates the potential of machine learning to improve loan processing speed, fairness, and reliability. By integrating multiple classifiers and alternative credit variables, it addresses key shortcomings of traditional evaluation methods

4.4 Limitations and Future Work

Although the Ensemble model shows strong results, several limitations and directions remain:

- Dataset Constraints: Only one dataset of 1,000 samples was used. Broader validation across multiple banks would increase generalizability.
- Feature Limitations: Current features exclude behavioral and transactional data (e.g., mobile money, supplier records). Adding these could further improve accuracy.
- Interpretability: Ensemble models are less transparent than Logistic Regression. Future work should incorporate Explainable AI (e.g., SHAP, LIME) to justify loan decisions.
- Scalability: While runtime is acceptable, additional optimization is needed for high-volume, real-time lending environments.

5. CONCLUSION

This study set out to address the persistent challenges faced by Small and Medium-Scale Enterprises (SMEs) in accessing fair and timely credit through the development of an Enhanced SME Loan Distribution System using an Ensemble Method. Traditional loan distribution practices are often slow, subjective, and reliant on limited financial information, leading to credit rationing and the exclusion of potentially viable SMEs.

The system developed in this work integrates Random Forest, XGBoost, and Logistic Regression in a voting ensemble framework. By leveraging the strengths of these models, the system provides more accurate, balanced, and robust credit risk predictions compared to individual models. Experimental results showed that the ensemble approach achieved an accuracy of 82.4%, a precision of 81.3%, and a recall of 80.5%, outperforming all baseline classifiers by 3–6% across key evaluation metrics. The confusion matrix analysis further confirmed the system’s reliability in minimizing both false approvals (credit risk) and false rejections (missed opportunities).

Beyond technical performance, the system introduces practical benefits for lenders and SMEs alike:

- Faster, data-driven loan decisions reduce delays in credit access.
- Objective scoring minimizes human bias, promoting fairness in credit allocation.

- Improved predictive capability reduces default risks, strengthening financial sustainability.

However, the study also identified certain limitations. The dataset used was relatively small and drawn from a single context, which may affect generalizability. Model interpretability remains a challenge, and while performance is promising, further work is needed to scale the system for real-time, large-volume lending environments.

In light of these findings, the study concludes that ensemble learning represents a viable and impactful solution for enhancing SME loan distribution systems. By balancing predictive accuracy, fairness, and operational efficiency, the proposed system contributes not only to academic discourse but also to the practical advancement of financial inclusion for SMEs.

Future extensions of this research should explore the integration of alternative credit data sources (e.g., mobile money transactions, supply chain records, and behavioral data), the adoption of explainable AI techniques for greater transparency, and scalability enhancements to support nationwide deployment. Such advancements will ensure that SME financing becomes more inclusive, efficient, and sustainable, thereby driving economic growth and innovation.

6. REFERENCES

[1] Amare, A. (2021). *A Loan Default Prediction Model for Acsi: A Data Mining Approach*.

[2] Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.

[3] Dietterich, T. G. (2000). *Ensemble methods in machine learning*. International Workshop on Multiple Classifier Systems, 1–15. Springer.

[4] Dhruba, M. I. M., Nawab, H. G., Sazzad, H., Syed, Z. H., & Shoumo, H. A. (2019). *Application of Machine Learning in Credit Risk Assessment: A Prelude to Smart Banking*.

[5] Du, J., Li, F., Chen, Q., & Zeng, Y. (2019). *Big data analytics and AI to improve the efficiency of SME financing*. Enterprise Information Systems, 13(9), 1344–1360.

[6] Gopichand, M. (2023). *Using Novel Logistic Regression over K-Nearest Neighbor for Improved Accuracy in Loan Prediction*.

[7] Hand, D. J., & Henley, W. E. (1997). *Statistical classification methods in consumer credit scoring: A review*. Journal of the Royal Statistical Society: Series A (Statistics in Society), 160(3), 523–541.

[8] Kipkogei, F., Kabano, I. H., Murorunkwere, B. F., & Joseph, N. (2021). *Business success prediction in Rwanda: A comparison of tree-based models and logistic regression classifiers*.

[9] Li, J., Liu, H., Yang, Z., & Han, L. (2021). *A Credit Risk Model with Small Sample Data Based on G-XGBoost*.

[10] Louzada, F., Ara, A., & Fernandes, G. (2016). *Classification methods applied to credit scoring: Systematic review and overall comparison*. Surveys in Operations Research and Management Science, 21, 117–134.

[11] Opitz, D., & Maclin, R. (1999). *Popular ensemble methods: An empirical study*. Journal of Artificial Intelligence Research, 11, 169–198.

[12] Oyedeji, O. (2023). *SMEs face cash crunch as banks prioritise large enterprises*. Dataphyte.

[13] Shinde, et al. (2022). *Loan Prediction System Using Machine Learning*.

[14] Sheikh, M. A., Goel, A. K., & Kumar, T. (2020). *An Approach for Prediction of Loan Approval using Machine Learning Algorithm*.

[15] Song, Y., & Wu, R. (2021). *The Impact of Financial Enterprises' Excessive Financialization Risk Assessment for Risk Control based on Data Mining and Machine Learning*.

[16] Supriya, P., Pavani, M., & Saisushma, N. (2021). *Home Loan Prediction Using Machine Learning Models*.

[17] Wang, S., You, S., & Zhou, S. (2023). *Loan Prediction Using Machine Learning Methods*. Proceedings of the International Conference on Financial Technology and Business Analysis.

[18] Xiao, Y., Zhang, W., Pang, Y., & Xie, J. (2020). *Semi-supervised novelty detection for small-scale imbalanced credit data classification*. Neurocomputing, 387, 91–102.

[19] Zhang, Q. (2020). *Loan risk prediction model based on Random Forest*. Journal of Engineering Science and Technology, 2(2), 1–7.