IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Loan Default Prediction In Microfinance Institutions Using Machine Learning And Deep Learning Techniques

Sumedha Arya

Abstract— Microfinance Institutions (MFIs) have proved to be the best support for low-income people in achieving their financial goals. However, there has been always a risk for loan defaults which can threaten the stability of MFIs. Previous techniques have less emphasized on this problem. They were much more into data analytics. Therefore, in this study we used machine learning and deep learning techniques to improve loan default prediction using a large FinTech dataset. Four machine learning and one deep learning technique was developed and compared. The results show that Logistic Regression performed best overall with a ROC-AUC of 0.7186 and recall of 0.65. XGBoost achieved high accuracy of 84% but performed poorly in identifying defaulters. The Neural Network showed competitive performance with a ROC-AUC of 0.7055 and recall of 0.58. The study concludes that Logistic Regression remains a strong and explainable baseline model for MFIs, while advanced models like Neural Networks have good potential for loan default classification.

Index Terms— Microfinance, Loan Default Prediction, Machine Learning, FinTech, SMOTE, Neural Networks, Logistic Regression.

I. Introduction

Microfinance Institutions (MFIs) are the important financial organizations. They have developed over the past few years to provide financial support to the people of low source income. The operation of MFIs is not only limited to cities but they also work in under privileges areas. It helps support economic growth, reduce poverty, and create jobs. They offer small loans, savings, insurance, and other basic financial services with simple procedures and little or no paper work [1].

Poverty remains a serious problem in many underdeveloped countries. In these countries, people face so many challenges such as many of them die from hunger even though there is enough food worldwide. Therefore, MFIs proved to be a lifeline for them, helping poor people by giving them access to small loans and financial services. Still, issues are there which hinders the working of MFIs properly like weak policies, poor financial system support and lack of education. Also, how to overcome the challenges of defaulters who took loans from the MFIs and unable to repay [2, 3].

The emergence of Financial Technology (FinTech) has revolutionized the microfinance sector by introducing automated processes and data-driven decision-making. However, the risk of loan defaulters is always been there, which hurdles the growth of MFIs. Machine learning (ML) and deep learning (DL) techniques can be used to identify the loan defaulters making MFIs stronger with automation. Unlike traditional statistical methods, advanced algorithms can identify complex, non-linear patterns within vast datasets. But, one of the most significant obstacles is "class imbalance," where the number of non-defaulters significantly outweighs the number of defaulters. If not addressed, this leads to models that are biased toward the majority class, failing to identify the very "risky" customers. Furthermore, in the financial world, "interpretability" is as important as "accuracy"; lenders must be able to explain why a loan was rejected.

This research aims to explore the intersection of finance with technology. By implementing and comparing a variety of models, this study evaluates how ML can enhance the sustainability of MFIs. The paper is divided into following sections; review of the literature, research methodology, results analysis, conclusion and future work.

II. Review Of The Literature

The use of technology in microfinance (MF) has benefitted in the working of institutions. With its help, accountability, governance, and transparency, has been improved to build trust among stakeholders. The smart cards innovation, has made microfinance borrowers to access financial services easily. This has reduced poverty by improving access to education, healthcare, government services, and finance [4].

With the advancement in internet technology, mobile phones, and ATMs, enabled MFIs to serve their clients more conveniently and efficiently [5]. The use of smartphones has further improved profitability by enabling mobile banking and digital services. These apps allow MFIs to reach people in rural and underserved areas. Therefore, financial and economic development has boost up in emerging economies [6][7][8]. But to make it more successful, good infrastructure and digital skills are required among users.

Lower operational costs in MFIs with automation reduces manual work and errors, helping them to save money. With effective pricing strategies, and well management further provides scope of good services to clients at lower costs [9]. Further, collecting and analyzing personalized customer data, can help MFIs better understand clients behavior, risks, and needs. It is suitable to design better financial products for long-term financial well-being of the clients [10][11].

The growth of technology in MFIs serves multiple benefits, but still in digitalization there are challenges. It is not only about access to technology but also about trust, culture, and skills. Even where technology is available, some users hesitate to adopt it [12][13]. In addition, some economic barriers can affect the scaling of MFIs. For the low-income users, cost of smartphones, internet, and service fees can prevent accessing digital services [14]. Also, digitalization can weaken traditional microfinance practices such as group lending by reducing face-to-face interaction. Therefore, social bonding is reduced which could negatively affect repayment behavior.

However, empirical studies have shown mixed results. According to some research findings, there is a positive impact of digital adoption on social performance [15][16], while others highlighted the importance of institutional and social context in determining outcomes [17].

FinTech also affects the financial performance of MFIs through several mechanisms. These are as follows:

1. **Efficiency Improvement:** Automation improves the efficiency of administrative work and transaction processing, thereby reducing errors and saving time [18]. However, high initial cost of technology and ongoing maintenance still remains an issue.
2. **Market Expansion:** Digital platforms serve as a market expansion allowing MFIs to reach new customers across the globe without opening physical branches. However, strong infrastructure, marketing investment, and client education, are required to make it successful.
3. **Better Risk Management:** Advanced data analytics and digital credit scoring are required to make MFIs evaluate borrowers more accurately with managing loan accounts more effectively [19][20]. However, high-quality data, skilled staff, and technological investment should be acquired for this process.
4. **Cybersecurity Threats:** It can lead to privacy and security challenges. It can lead to data breaches, fraud, and reputational damage, requiring continuous investment in enhancing security systems [21].
5. **Regulatory Compliance:** It increases operational cost due to data protection and privacy needs [22]. Furthermore, MFIs needs to be continuously, to overcome challenges from fintech companies and technology firms as they can reduce market share and profit margins [23].

Based on above studies, it is clear that some reports show improved efficiency and performance due to technology adoption [24][25], while others find negative effects on operational efficiency [14]. These differences suggest that institutional capacity plays a key role.

Digital infrastructure influences how strongly FinTech affects both social and financial performance.

· For social performance, strong infrastructure enables reliable services such as fast transactions, secure access, and fewer system failures. It also supports financial education through mobile apps, alerts, and interactive tools, improving inclusion and literacy [26]. However, infrastructure is expensive. High costs for maintenance, security, and upgrades can increase service fees, which may reduce affordability and increase exclusion in weaker regions [27].

· For financial performance, good infrastructure improves efficiency through real-time monitoring, automated systems, and advanced analytics. It also supports better risk management. Yet, costs related to bandwidth, licenses, cybersecurity, training, and downtime risks must be considered.

The real impact depends on factors such as implementation quality, organizational readiness, staff skills, and local conditions.

III. RESEARCH METHODOLOGY

This section describes machine learning techniques used on as FinTech dataset to predict whether a loan applicant will **b**e a defaulter or not. The steps include in this process are data acquisition, cleaning, exploration, feature selection, handling data imbalance, building models, and evaluating their performance.

1. The dataset is sourced from Kaggle. It contains 307,511 loan records and 122 features. The target column comprises of 1 and 0 values, where 1 represents defaulter while 0 means non-defaulter.

2. In data preprocessing or cleaning, columns with too many missing values were removed. It has been found that 41 columns had more than 50% missing data, therefore they were dropped. The final dataset comprises of 81 useful columns. The missing values were filled with median for numerical columns while categorical columns, with missing values were filled by mode. The columns were grouped into numerical features (68) and non-numerical features (13).

3. EDA was used to understand patterns in the data. For this univariate and bivariate analysis of the dataset was performed. In, univariate analysis, different graphs such as Countplot and pie plot were used to understand individual features distribution for discrete and categorical features. While histogram, kde and box plots were made for continuous features. Some of them are as follows:

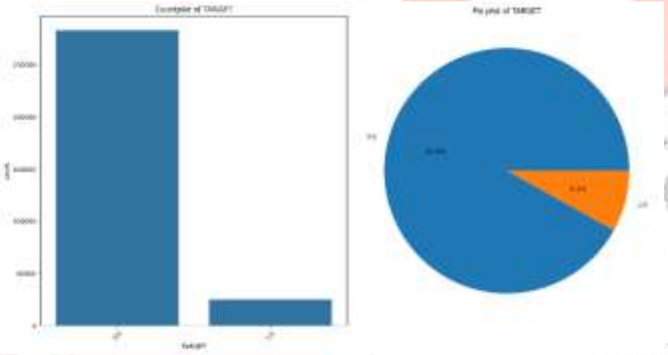


Fig. 1: Countplot and Pie plot for Target

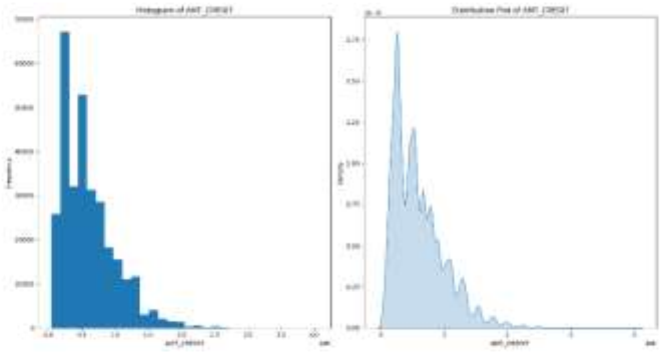


Fig. 2: Histplot and Kde plot for Amount Credit

4. In bivariate Analysis, to understand the relationship between features and the target variable, feature values between defaulters and non-defaulters were compared. Correlation analysis was performed and heatmap, pivot tables were created.

5. Based on this analysis, we found that, features like EXT_SOURCE_2 and EXT_SOURCE_3 were strongly related to loan repayment. Older clients were less likely to default and Clients with lower education levels showed higher default risk.

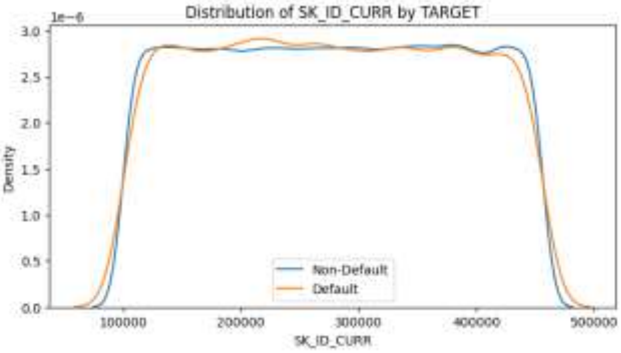


Fig. 3: Distribution of SK_ID_CURR by TARGET

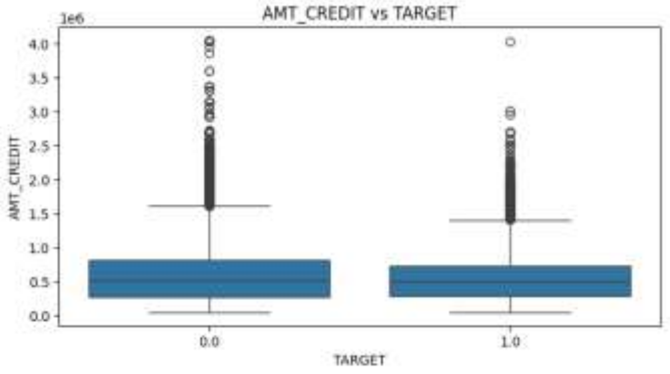


Fig. 4: Amount Credit vs TARGET

6. As, not all features present in the dataset were useful, therefore feature selection technique was applied to find the most important ones from it. Features were ranked based on correlation with the target and the top 40 among them were chosen. Highly correlated features were removed to avoid redundancy. Finally, 15 most useful features were selected.

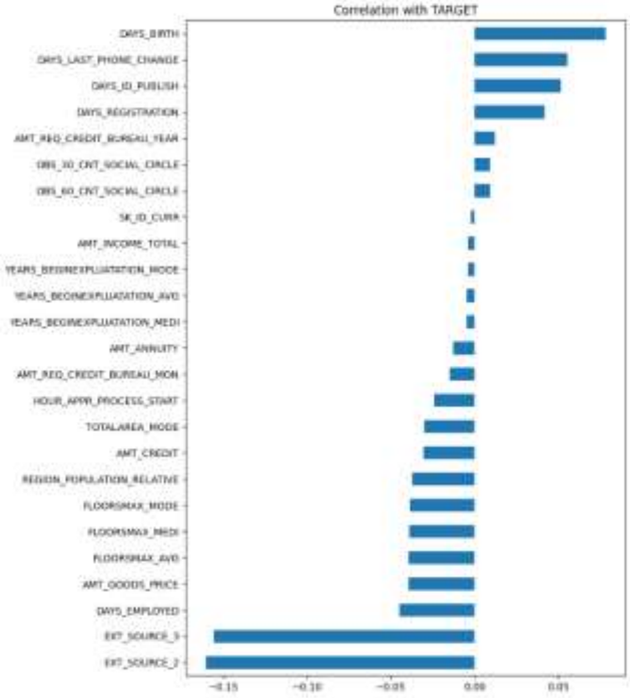


Fig. 5: Correlation with TARGET

7. Inorder for the models to work better, feature scaling is performed so that data should come on the same scale. So, continuous features were standardized using Standard Scaler, which converts values with mean as 0 and standard deviation as 1.

8. Class imbalance is always an issue. In current dataset also, same problem was identified. It shows only about 8% of clients were defaulters. Therefore, to fix this, SMOTE (Synthetic Minority Oversampling Technique) was applied to the training data. It creates synthetic examples of the minority class, making the data more balanced (50:50 ratio).

9. The dataset was divided into two parts with 80% kept for training the models and left 20% for testing the models. Stratified sampling was applied used to make sure both sets had the same proportion of defaulters and non-defaulters.

10. Model Development: Three machine learning models were built to perform the classification based on defaulters and non-defaulters. These models are logistic regression, decision tree and random forest. The features of these models are: Logistic Regression is a simple and interpretable model. It works well as a baseline for binary classification. Decision Tree is used with limited depth to avoid overfitting on the model. It is easy to understand the behavior. Random Forest is an ensemble model that combines many decision trees. It is more powerful and handles complex data patterns better. All models were trained using the balanced dataset created with SMOTE.

11. Model Evaluation: The models were tested using: Classification report with precision, recall, F1-score as the parameters, Confusion matrix to show the visualization for the models ability to detect defaulter and non-defaulters, ROC-AUC score shows the main performance metric of the models. SMOTE improved the model’s ability to identify defaulters, although precision decreased slightly.

12. Tools and Environment: The tools and environment used for this implementation comprises of language as Python 3.12, Libraries are pandas, numpy, matplotlib, seaborn, scikit-learn, imbalanced-learn and the Platform for implementation is Kaggle.

13. This methodology provides a clear and systematic process for predicting loan default risk.

IV. Results Analysis

In this section, detailed analysis of the results is performed. Five machine learning models were tested using unseen test data of 61,503 records. The results achieved after training on balanced data created using SMOTE are as follows:

Table 1: Results Analysis of ML and DL Models

Model	ROC-AUC	Accuracy	Precision (Default)	Recall (Default)	F1-Score (Default)
Logistic Regression	0.7186	0.67	0.15	0.65	0.24
Decision Tree	0.6849	0.62	0.13	0.67	0.22
Random Forest	0.6979	0.72	0.15	0.53	0.24
XGBoost	0.6846	0.84	0.19	0.29	0.23
Neural Network	0.7055	0.70	0.15	0.58	0.24

The analysis of results is performed on the based on the ROC-AUC and classification report metric having recall, precision and accuracy. Our findings based on them are as follows:

ROC-AUC

- Logistic Regression performed best with a score of 0.7186. This means it was the best at separating defaulters from non-defaulter customers.
- Decision Tree performed the worst.
- Random Forest performed better than Decision Tree but poor than Logistic Regression.
- XGBoost showed lower discrimination power.
- Neural Network performed competitively with a ROC-AUC of 0.7055.

Recall (Finding Defaulters)

- Logistic Regression and Decision Tree identified about 65–67% of defaulters correctly.
- This is important because banks prefer to identify faulty customers even if a few good customers are mistakenly rejected.
- Random Forest missed more defaulters with a recall value of 0.53.
- XGBoost had the lowest recall (0.29), meaning many defaulters were missed.
- Neural Network identified 58% of defaulters, showing good ability to detect risky customers.

Precision (Correct Risk Predictions)

- Precision was low for all models, that was around 13–15%.
- This means many customers predicted as risky were actually safe.
- As the dataset was highly imbalanced, therefore, SMOTE was used to balance it. Therefore, due to this process, such results are obtained:
    - We catch more defaulters
    - But we also wrongly mark more safe customers as risky

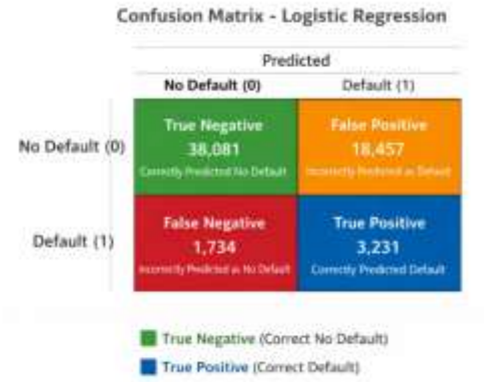
Accuracy

- XGBoost achieved the highest accuracy of 84%.
- Random Forest achieved an accuracy of 72%, while Neural Network achieved 70%.
- But accuracy is not very reliable here because most customers are non-defaulters. And this is due to the dataset and SMOTE balancing technique used.
- That is why ROC-AUC is more meaningful.

Confusion Matrix

Logistic Regression

- Correctly identified 3,231 defaulters avoiding too many false claims showing overall best balance.



Confusion Matrix - Logistic Regression

	Predicted	
	No Default (0)	Default (1)
No Default (0)	True Negative 38,081 Correctly Predicted No Default	False Positive 18,457 Incorrectly Predicted as Default
Default (1)	False Negative 1,734 Incorrectly Predicted as No Default	True Positive 3,231 Correctly Predicted Default

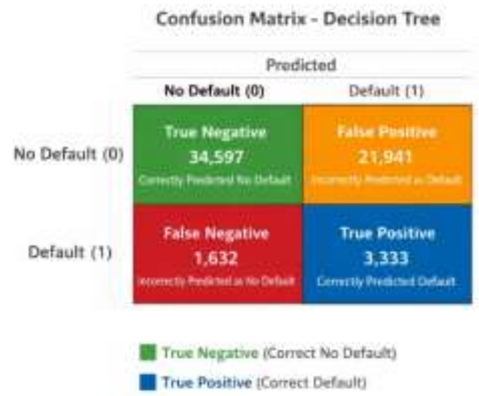
🟩 True Negative (Correct No Default)

🟦 True Positive (Correct Default)

Fig. 6: Confusion Matrix of Logistic Regression

Decision Tree

- Caught the most defaulters as 3,333, showing aggression by wrongly marking too many good customers as defaulters



Confusion Matrix - Decision Tree

	Predicted	
	No Default (0)	Default (1)
No Default (0)	True Negative 34,597 Correctly Predicted No Default	False Positive 21,941 Incorrectly Predicted as Default
Default (1)	False Negative 1,632 Incorrectly Predicted as No Default	True Positive 3,333 Correctly Predicted Default

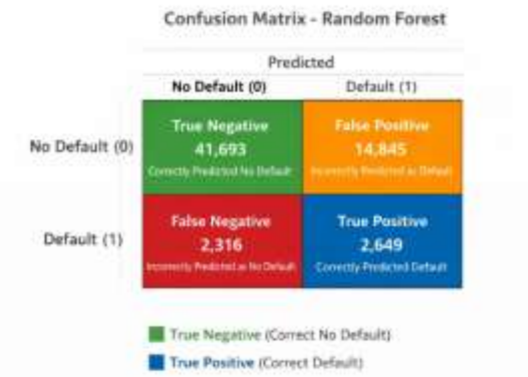
🟩 True Negative (Correct No Default)

🟦 True Positive (Correct Default)

Fig. 7: Confusion Matrix of Decision Tree

Random Forest

- Made the fewest false claims showing to be more conservative by missing many defaulters.



Confusion Matrix - Random Forest

	Predicted	
	No Default (0)	Default (1)
No Default (0)	True Negative 41,693 Correctly Predicted No Default	False Positive 14,845 Incorrectly Predicted as Default
Default (1)	False Negative 2,316 Incorrectly Predicted as No Default	True Positive 2,649 Correctly Predicted Default

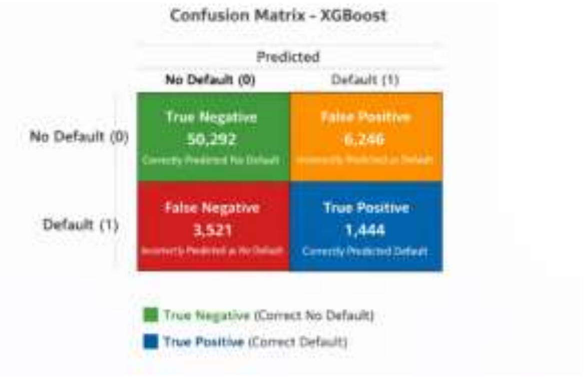
🟩 True Negative (Correct No Default)

🟦 True Positive (Correct Default)

Fig. 8: Confusion Matrix of Random Forest

XGBoost

- Correctly identified 1,444 defaulters, but missed many risky customers, showing weak performance in identifying defaulters.



Confusion Matrix - XGBoost

	Predicted	
	No Default (0)	Default (1)
No Default (0)	True Negative 50,292 Correctly Predicted No Default	False Positive 6,246 Incorrectly Predicted as Default
Default (1)	False Negative 3,521 Incorrectly Predicted as No Default	True Positive 1,444 Correctly Predicted Default

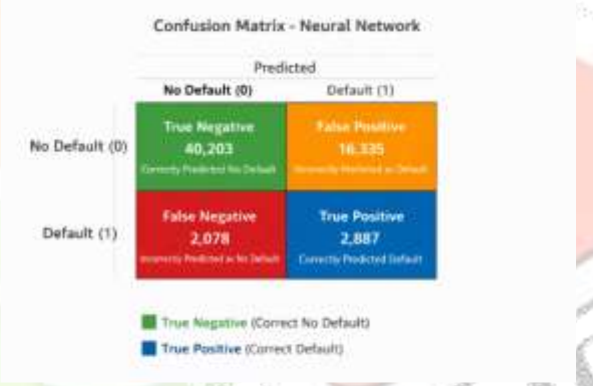
🟩 True Negative (Correct No Default)

🟦 True Positive (Correct Default)

Fig. 9: Confusion Matrix of XGBoost

Neural Network

- Correctly identified 2,887 defaulters, showing better balance between identifying risky and safe customers compared to Random Forest and XGBoost.



Confusion Matrix - Neural Network

	Predicted	
	No Default (0)	Default (1)
No Default (0)	True Negative 40,203 Correctly Predicted No Default	False Positive 16,335 Incorrectly Predicted as Default
Default (1)	False Negative 2,078 Incorrectly Predicted as No Default	True Positive 2,887 Correctly Predicted Default

🟩 True Negative (Correct No Default)

🟦 True Positive (Correct Default)

Fig. 10: Confusion Matrix of Neural Network

The key findings of our research are as follows:

- Data was highly imbalanced.
- SMOTE was used to balance it. It helped in finding the defaulters but reduced precision.
- In feature selection, it was found that, features such as EXT_SOURCE_2, EXT_SOURCE_3, age, and region feature proved to be very useful.
- Logistic Regression is the best baseline model because of its overall performance with explainability as preferred by the financial institutions.
- Neural Network showed strong potential with competitive ROC-AUC and recall.
- XGBoost achieved high accuracy but failed to detect defaulters effectively.
- As a business perspective, a ROC-AUC of 0.72 is a good starting performance.
- MFIs can adjust their policies based on the decision threshold to:
    - Catch more risky customers represented by higher recall
    - Or avoid rejecting good customers represented by higher precision

V. Conclusion And Future Work

This study presented a systematic approach for predicting loan default risk in FinTech dataset of MFIs using ML and DL Models. Logistic Regression, Decision Tree, Random Forest, XGBoost and Neural Network models were evaluated on dataset after handling class imbalance using SMOTE. Among these models, Logistic Regression achieved the best baseline performance with a ROC-AUC score of 0.7186, showing strong ability to distinguish between defaulters and non-defaulters. The XGBoost model achieved an accuracy of 84%, with a ROC-AUC score of 0.6846. However, it showed weaker performance in identifying defaulters, with a recall of 0.29 for the minority class. The Neural Network model achieved a ROC-AUC score of 0.7055, which is comparable to Logistic Regression. It showed better recall for defaulters (0.58) compared to XGBoost, meaning it was more effective in detecting risky customers.

Although advanced models showed promising results, Logistic Regression still remains the most suitable baseline model due to its strong performance, simplicity, and interpretability, which is highly preferred in financial applications where explainability is critical.

In future, the performance of the classification can be further improved by:

- Applying more advanced ensemble models such as LightGBM.
- Performing hyperparameter tuning using Grid Search.
- Using alternative imbalance handling techniques such as class weights.

Overall, this research demonstrates that machine learning techniques, when applied properly, can significantly improve loan default prediction and help MFIs make better and more responsible lending decisions.

References

[1] C. A. Nwigwe, B. T. Omonona, and V. O. Okoruwa, “Microfinance and poverty reduction in Nigeria: A critical assessment,” Australian Journal of Business and Management Research, vol. 2, no. 4, pp. 33–40, 2016. doi: 10.52283/NSWRCA.AJBMR.20120204A05.

[2] R. Cu, “Microfinance in PHL at crossroads,” BusinessMirror, Mar. 1, 2017. [Online]. Available: https://businessmirror.com.ph/microfinance-in-phl-at-crossroads/

[3] J. Cui, J. Sun, and R. Bell, “The impact of entrepreneurship education on the entrepreneurial mindset of college students in China: The mediating role of inspiration and the role of educational attributes,” The International Journal of Management Education, vol. 19, no. 1, Art. no. 100296, 2021. doi: 10.1016/j.ijme.2019.04.001.

[4] S. Cecchini and C. Scott, “Can information and communications technology applications contribute to poverty reduction? Lessons from rural India,” Information Technology for Development, vol. 10, no. 2, pp. 73–84, 2003. doi: 10.1080/02681102.2003.10510138.

[5] G. Ivatury, “Using electronic payments to build inclusive financial systems,” CGAP Focus Note, no. 32, 2005. (Available at: https://www.cgap.org/research/publication/using-technology-build-inclusive-financial-systems).

[6] McKinsey Global Institute, “Digital finance for all: Powering inclusive growth in emerging economies,” McKinsey & Company, September 2016. (Report).

[7] P. Gomber, R. J. Kauffman, C. Parker, and B. W. Weber, “On the FinTech revolution: Interpreting the forces of innovation, disruption, and transformation in financial services,” Journal of Management Information Systems, vol. 35, no. 1, pp. 220–265, 2018. doi: 10.1080/07421222.2018.1440766.

[8] J. Pytkowska and P. Koryński, “Digitalizing microfinance in Europe,” Microfinance Centre (MFC) Research Paper, December 2017. (Available at: https://mfc.org.pl/wp-content/uploads/2018/01/Digitalization-research-paper.pdf).

[9] T. T. Dang and H. Q. Vu, “FinTech in microfinance: A new direction for microfinance institutions in Vietnam,” Journal of Asian Business and Economic Studies (or Asian Journal of Business Environment), vol. 10, no. 3, pp. 13–22, 2020. doi: 10.13106/jidb.2020.vol10.no3.13 (or equivalent journal identifier).

[10] M. F. Sultan, A. Rafiq, R. Ahmad, and M. Asim, “Significance of FinTech for Microfinance Institutions (MFIs): Anatomical linkages of FinTech with value chain of MFIs and its implications,” in Financial Inclusion Across Asia: Bringing Opportunities for Businesses, C.-M. Leong et al. (Eds.), Emerald Publishing Limited, pp. 65–75, 2023. doi: 10.1108/978-1-83753-304-620231005.

[11] M. R. Visconti, “FinTech valuation,” in Startup Valuation: The Interaction Between Corporate Financial Theory and Application, Palgrave Macmillan, Cham, pp. 245–279, 2021. doi: 10.1007/978-3-030-71608-0_10 (or chapter-specific DOI).

[12] R. Mushtaq and C. Bruneau, “Microfinance, financial inclusion, and ICT: Implications for poverty and inequality,” Technology in Society, vol. 59, 101154, 2019. doi: 10.1016/j.techsoc.2019.101154.

[13] S. Anrijs, I. Mariën, L. De Marez, and K. Ponnet, “Excluded from essential internet services: Examining associations between digital exclusion, socio-economic resources, and internet resources,” Technology in Society, vol. 73, 101050, 2023. doi: 10.1016/j.techsoc.2023.101050.

[14] M. Fersi, M. Boujelbène, and F. Arous, “Microfinance’s digital transformation for sustainable inclusion,” European Journal of Management and Business Economics, vol. 32, no. 5, pp. 525–559, 2023. doi: 10.1108/EJMBE-10-2022-0332 (or equivalent).

[15] B. Gutiérrez-Nieto, Y. Fuertes-Callén, and C. Serrano-Cinca, “Internet reporting in microfinance institutions,” Online Information Review, vol. 32, no. 3, pp. 417–436, 2008. doi: 10.1108/14684520810889673.

[16] R. Mersland and R. Ø. Strøm, “What drives the microfinance lending rate?,” Midwest Finance Association 2013 Annual Meeting Paper, 2012. (Working paper or conference version).

[17] G. Dorfleitner, Q. A. Nguyen, and M. Röder, “Microfinance institutions and the provision of mobile financial services: First empirical evidence,” Finance Research Letters, vol. 31, pp. 357–362, 2019. doi: 10.1016/j.frl.2018.12.002.

[18] T. Tanchangya et al., “Financial Technology-Enabled sustainable finance for small- and Medium-Sized enterprises,” Environmental Innovation and Management, vol. 1, 2550006, 2025. (Projected/publication details as cited).

[19] A. Ashta and H. Herrmann, “Artificial intelligence and FinTech: An overview of opportunities and risks for banking, investments, and microfinance,” Strategic Change, vol. 30, no. 3, pp. 211–222, 2021. doi: 10.1002/jsc.2404.

[20] E. Widarwati, I. Y. Fajar, N. Nurmalasari, and E. Wityasminingsih, “Digital finance and microfinance risk level,” Paper presented at the 10th International Conference on Management and Muamalah 2023 (ICoMM 2023), Universiti Islam Selangor (KUIS), Malaysia, 2023.

[21] S. Corbet and C. Gurdgiev, “Financial digital disruptors and cyber-security risks: Paired and systemic,” Journal of Terrorism and Cyber Insurance, vol. 1, no. 2, 2017. (Forthcoming or specific issue details as cited).

[22] M. Bakhoum, B. Gonzalez Otero, J. Hoffmann, and M. Sarr, “Data governance in emerging economies to achieve the sustainable development goals: Senegal country report,” Max Planck Institute for Innovation and Competition Research Paper No. 24, 2024.

[23] G. Mdluli and S. Staschen, “Pitfalls in MFI digitization: Overlooking the regulatory environment,” CGAP Blog, 2022. (Available at: https://www.cgap.org/blog/pitfalls-in-mfi-digitization-overlooking-regulatory-environment).

[24] G. Hishigsuren, “Information and communication technology and microfinance: Options for Mongolia,” ADB Institute Discussion Paper No. 42, 2006.

[25] T. Mora and F. Prior, “The impact of mobile financial services usage on microfinance delinquency,” Applied Economics, vol. 50, no. 50, pp. 5354–5365, 2018. doi: 10.1080/00036846.2018.1488073 (or equivalent).

[26] J. Kacani and G. Shaqiri, “Emerging ICT clusters in the Western Balkans: A convergence model based on financial statement analysis,” Regional Science Policy & Practice, vol. 16, no. 4, 2024. doi: 10.1111/rsp3.12687.

[27] Z. Elouaourti and A. Ibourk, “Financial Technologies for All MENA citizens: Tackling barriers and promoting inclusion,” Regional Science Policy & Practice, vol. 16, no. 6, 2024. doi: 10.1016/j.rspp.2024.100019.