Scientific Journal of Computer Science

futuristech
PT. Teknologi Futuristik Indonesia

Received: May 29, 2026; Revised: August 2, 2026; Accepted: August 5, 2026; Published: August 12, 2026; Current version: August 12, 2026. e-ISSN: 3110-3170
Digital Object Identifier 10.64539/sjcs.v2i2.2026.520

Article

SMOTETomek-DNN: A Machine Learning Framework for Credit Risk Prediction with an Imbalanced Dataset

Tiruneh Kebede Dubale[1,*], Siraj Sebhatu Seyoum[2]

[1] Department of Information Technology, Wolaita Sodo University, Wolaita Sodo, Ethiopia; e-mail: tirunehkebede6@gmail.com (T. K. Dubale).
[2] Department of Information System, Wolaita Sodo University, Wolaita Sodo, Ethiopia; e-mail: siraj.sebhatu@wsu.edu.et (S. S. Seyoum).
* Correspondence

The authors received no financial support for the research, authorship, and/or publication of this article.

Abstract: Accurate credit risk prediction plays a critical role in strengthening the financial stability of microfinance institutions, especially in developing economies, where increasing loan defaults and imbalanced borrower records create significant challenges for reliable decision-making. Although machine learning approaches have improved credit assessment practices, existing models often favor majority-class borrowers and fail to detect high-risk default cases effectively because of severe class imbalance. This limitation highlights the need for more robust and imbalanced-sensitive predictive frameworks. This study aims to develop an effective machine learning-based credit risk prediction framework by integrating data balancing strategies with ensemble and deep learning models. This study systematically investigates the impact of baseline learning and multiple resampling techniques, including oversampling, undersampling, and hybrid methods, when applied to Random Forest, XGBoost, LightGBM, CatBoost, and Deep Neural Network classifiers. The effectiveness of the proposed models was assessed using imbalance-aware evaluation measures, particularly ROC-AUC and Geometric Mean, along with conventional classification metrics. The experimental findings demonstrate that incorporating resampling techniques substantially improved the default detection performance. The DNN model combined with SMOTETomek achieved the best results, obtaining 94.9% F1-score, ROC-AUC 98%, and 97.2% of G-Mean. CatBoost also exhibited consistent competitiveness across different sampling configurations. These findings suggest that hybrid sampling integrated with advanced learning architectures can provide a reliable and practical solution for managing credit risk in imbalanced microfinance datasets, supporting improved lending decisions and sustainable financial operations in the future.

Keywords: Credit risk prediction; Microfinance institutions; Imbalanced learning; Machine learning; Sampling techniques.

1. Introduction

Financial inclusion in emerging economies depends significantly on the ability of microfinance institutions (MFIs) to provide accessible and sustainable credit to underserved populations. In Ethiopia, licensed MFIs play an important role in expanding access to finance, serving more than 7.6 million borrowers with a loan portfolio of approximately ETB 94.3 billion as of 2023 [1]. Globally, the expansion of microfinance has also been substantial, with approximately 140 million people receiving microloans totaling more than USD 124 billion by 2018 [2]. However, the rapid expansion of lending activities has increased the exposure of financial institutions to credit risk and loan default. In Ethiopia, MFIs reported an average non-performing loan (NPL) ratio of approximately 6.8% in 2023, exceeding regulatory thresholds and creating concerns regarding institutional sustainability [1]. Credit risk assessment is therefore an important component of responsible lending because inaccurate identification of high-risk borrowers can increase financial losses and weaken the sustainability of lending institutions.

The challenge is particularly important in the

Ethiopian microfinance context because credit assessment is affected by limited financial information and institutional constraints. Ethiopia does not have a centralized credit bureau system dedicated to microfinance institutions, and many loan assessments continue to depend on manual evaluation, heuristic judgment, or basic scorecard approaches [3]. Borrowers may have informal sources of income, limited financial histories, incomplete documentation, and irregular economic activities, making conventional credit assessment difficult [3]. These characteristics create a need for predictive approaches that can identify potential defaulters from complex borrower information while remaining suitable for data-constrained financial environments.

Machine learning (ML) provides an opportunity to improve credit risk prediction by learning complex relationships among borrower characteristics and repayment outcomes. Algorithms such as Decision Trees, Random Forest, Support Vector Machines, Gradient Boosting Machines, Extreme Gradient Boosting (XGBoost), LightGBM, CatBoost, and Deep Neural Networks (DNNs) have demonstrated their ability to model nonlinear patterns in financial data. Previous studies have reported that boosted ensemble models can outperform conventional statistical approaches and individual classifiers in loan default prediction [4]-[6]. Similarly, combining deep learning and ensemble-based approaches can improve predictive robustness when dealing with complex and high-dimensional financial datasets [6], [7]. Nevertheless, most existing credit-scoring research has concentrated on commercial banking, peer-to-peer lending, or fintech environments in developed economies where borrower information is generally more structured and digitally available [5]. Consequently, models developed in these settings may not directly address the operational and data characteristics of Ethiopian MFIs.

One of the most important unresolved methodological challenges is class imbalance. In credit datasets, non-default borrowers usually substantially outnumber default borrowers. Consequently, conventional classifiers may become biased toward the majority class and achieve high overall accuracy while failing to identify borrowers who are actually at high risk of default [8]. This problem is particularly serious in credit risk prediction because correctly identifying the minority default class is often more important than obtaining high overall accuracy. Research has shown that imbalanced datasets can systematically bias machine learning models toward majority classes unless appropriate corrective mechanisms are applied [9]. Therefore, evaluation using only accuracy may provide an incomplete or misleading assessment of model performance in credit risk applications.

Several techniques have been proposed to address class imbalance, including minority oversampling, majority undersampling, and synthetic data generation. Among these, the Synthetic Minority Oversampling Technique (SMOTE) and its variants generate synthetic minority observations rather than simply duplicating existing cases [10]-[12]. Advanced approaches such as ADASYN, Borderline-SMOTE, and SVM-SMOTE attempt to improve minority-class learning by emphasizing difficult observations and decision-boundary regions. Algorithm-level methods, including cost-sensitive learning and focal-loss optimization, provide alternative solutions, while hybrid approaches combine resampling with ensemble or deep learning models. Previous evidence indicates that combining imbalance-handling strategies with machine learning classifiers can improve minority-class detection and measures such as recall and ROC-AUC [8]. However, the most appropriate resampling strategy may depend strongly on the characteristics of the dataset and the prediction model, meaning that a single balancing technique cannot necessarily be assumed to be optimal.

Despite these developments, important gaps remain in Ethiopian microfinance credit risk research. Existing local studies have generally emphasized traditional statistical methods or individual machine learning classifiers and have given limited systematic attention to severe class imbalance. In particular, there is limited evidence comparing multiple advanced resampling techniques, including SMOTE, ADASYN, Borderline-SMOTE, and SVM-SMOTE, across a broad range of machine learning models in Ethiopian MFI data. Previous work has also provided limited comparative evidence concerning the performance of deep learning models under imbalance-aware learning conditions. This is important because DNNs can capture nonlinear interactions among demographic, socioeconomic, institutional, and behavioral borrower characteristics that may not be adequately represented by conventional models. Furthermore, most credit risk studies emphasize predictive performance while giving less attention to model interpretability. Explainability is particularly important for financial institutions because credit decisions need to be understandable to loan officers, borrowers, managers, and regulators. Recent research also indicates that class imbalance can affect the stability of model explanations, demonstrating that predictive performance and interpretability should be considered together [13].

The main contributions of this study are as follows:

- A comprehensive imbalance-aware credit risk prediction framework was developed for Ethiopian microfinance institutions by integrating advanced machine learning, deep learning, and data resampling techniques to improve the default prediction performance.
- A systematic evaluation of multiple class-imbalance handling strategies was conducted by comparing oversampling, undersampling, and hybrid

sampling methods across XGBoost, LightGBM, CatBoost, Random Forest, and Deep Neural Network (DNN) models.
- A robust comparative performance benchmarking framework was established using imbalance-sensitive evaluation metrics, including Accuracy, Precision, Recall, F1-score, ROC-AUC, and G-Mean, enabling a reliable assessment of minority-class prediction performance beyond conventional accuracy.
- An SHAP-based explainable artificial intelligence (XAI) framework was incorporated to provide a transparent interpretation of model predictions by identifying the most influential borrower attributes contributing to credit risk decisions.
- A practical decision-support framework is proposed that combines predictive accuracy, imbalance correction, and model interpretability, offering an effective and trustworthy solution for improving credit assessment and sustainable lending decisions in Ethiopian microfinance institutions.

The remainder of this paper is organized as follows. Section 2 reviews previous studies on machine learning-based credit risk prediction and class imbalance handling. Section 3 describes the research methodology, including the dataset, preprocessing procedures, feature engineering, resampling techniques, machine learning and deep learning models, hyperparameter optimization, evaluation metrics, and experimental setup. Section 4 presents the experimental results obtained under different sampling strategies. Section 5 discusses the findings, compares them with previous studies, and highlights their practical implications and limitations for Ethiopian microfinance institutions. Section 6 concludes the study, and Section 7 outlines recommendations for future research.

2. Related Work

Credit risk prediction is vital in microfinance as microfinance institutions (MFIs) grow in emerging economies. Traditional credit-scoring methods, such as logistic regression and LDA, often have problems with complex data. They also address severe class imbalances. There are many non-defaults but few defaults [7], [14]. Previous advances have used machine learning (ML) to identify nonlinear patterns in borrower data. However, they must also directly address the imbalance problem. For example, [8] studied P2P lending. They found that using random forest (RF) with random undersampling helped detect defaulters better. This is demonstrated by the G-mean metric. [15] compared deep learning and ensemble trees using an imbalanced credit-card dataset. Tree ensembles, such as RF and gradient boosting, outperformed the neural networks. This is especially true when researchers have limited and skewed data [14]. [13] reviewed over 50 ML credit-scoring studies. They found that boosting models, such as XGBoost, were the most common. The usual evaluation metrics are AUC, accuracy, recall, precision, and F1 score.

Several Previous studies have focused on microfinance or credit-poor contexts. [16] applied ML to a microlending dataset from Africa. They used SMOTENC to rebalance the classes. This changed the percentage of "poor" risk from 83.7% to 36.2% after resampling [13]. The ensemble trees achieved an accuracy of over 70 %. Handling imbalance is also crucial [13], [15]. [17] proposed a WSMOTE-ensemble, combining bagging and weighted SMOTE, for small business loans. They found that RF with WSMOTE boosted the minority-class accuracy by 15.16% [17]. [17] compared several methods: SMOTE, ADASYN, SMOTE+Tomek, and ClusterCentroids. They used RF and XGBoost with German credit data to predict credit ratings. They found that all models performed better with balancing. In addition, oversampling and hybrid methods performed better than simple undersampling [18].

Other studies have highlighted hybrid and novel sampling methods. [19] introduced the weighted hybrid sampling boost (WHSBoost), which combines weighted SMOTE and weighted under-sampling in a boosting framework. WHSBoost outperformed standard SMOTE and SMOTEBoost on the benchmark and real credit datasets. It exhibited better recall, F1, and AUC. [7] studied federated learning for credit risk. Imbalance-aware FL models boosted minority-class prediction by 17.9% compared to local models. In agricultural microfinance, [20] combined environmental indicators, such as rainfall and soil, with borrower data in an XGBoost model reached 99% accuracy. It also achieved 84% recall and an F1 score of approximately 86%. This shows that contextual variables can enhance rural MFI scoring.

Research shows that ensemble tree models are popular for credit scoring applications. The models included XGBoost, LightGBM, CatBoost, and Random Forest. Deep neural networks have also been used. Ensembles often perform best on imbalanced data [13], [15]. A study [21] found that LightGBM, Random Forest, and CatBoost work well for credit risk. These models are typically assessed based on several metrics. These include accuracy, precision, recall, F1-score, ROC-AUC, and G-mean for imbalance [8], [15]. Researchers have used various resampling techniques to rebalance the classes. These range from no sampling (none) to random oversampling. They also use synthetic methods such as SMOTE, Borderline-SMOTE, SVMSMOTE, and ADASYN. Some hybrids include SMOTE+ENN and SMOTE+Tomek [16], [18]. Oversampling or hybrid methods usually improve minority recall without significantly reducing precision [18]. In contrast, pure undersampling can result in the loss of majority-class information.

Table 1. Annual Distribution of Loan Records.

year	2017	2018	2019	2020	2021	2022	2023	2024
Records	1,180	1,320	1,410	1,460	1,520	1,640	1,700	1,770

Table 2. Class Distribution of the Dataset.

Loan Status	Code	Frequency	Percentage (%)
Non-default	0	10,800	90.0
Default	1	1,200	10.0
Total	—	12,000	100.0

Ensemble learning with balance correction is the dominant approach. XGBoost+SMOTE or similar combos frequently emerge as top performers [22]. Balancing the training set (via SMOTE, ADASYN, etc.) consistently improves the detection of defaulters [17], [18]. Surveys suggest using several metrics to test skewed credit data. These include accuracy, AUC, recall, precision, F1, and G-mean [8], [13]. Importantly, microfinance/EME contexts are understudied: most prior work is on bank or P2P data. Ethiopian MFIs face unique challenges, such as informal income and limited records. Therefore, methods that have been successful in other developing areas, such as those in [16] and [20], can provide helpful insights. This study fills this gap by using ML models and sampling strategies on Ethiopian MFI data. Based on these findings, a suitable pipeline for Ethiopian MFI data would include data cleaning and feature selection, followed by training many classifiers (especially RF, XGBoost, LightGBM, CatBoost, and a DNN) under different sampling schemes (None, RUS, SMOTE variants, ADASYN, hybrid) [18], [21]. Accuracy, precision, recall, F1, ROC-AUC, and G-mean were used to evaluate the imbalance performance [8], [13].

3. Materials and Methods

3.1. Research Design

This study used an experimental design to create and test the machine learning models. The goal of this study was to predict credit defaults using imbalanced microfinance loan data. Experimental designs are effective for predictive modeling. They allow us to compare the algorithms in a controlled manner. We focused on preprocessing, sampling, and validation. In past financial risk analytics studies, ensemble learning and nonlinear machine learning methods have shown better prediction abilities than traditional scoring models. This is especially true for complex and imbalanced classification problems [21], [23]. This study compares the following top machine learning algorithms: Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), Categorical Boosting (CatBoost), Random Forest (RF), and Deep Neural Networks (DNN). Maintaining the same data sources, preprocessing steps, sampling methods, and validation processes in all experiments allows us to link differences in predictive performance directly to model features. In this way, we avoided the influence of external factors on the results.

3.2. Data Collection

This study utilized borrower loan records obtained from selected microfinance institutions (MFIs) operating under the regulatory supervision of the National Bank of Ethiopia. The dataset is not publicly available online because it contains confidential financial and personal borrower information protected by institutional data-sharing agreements and ethical research requirements. Access to the dataset is therefore restricted and requires authorization from the respective MFIs and relevant regulatory bodies. Historical loan records covering the period from 2017 to 2024 (see Table 1) were collected from participating institutions after obtaining the necessary permissions. All personally identifiable information was removed before analysis to ensure borrower privacy and compliance with ethical standards. The original dataset contained approximately 12,400 loan records. After data cleaning, duplicate removal, consistency verification, and exclusion of incomplete entries, 12,000 valid observations remained. The target variable was loan repayment status, where 0 denotes non-default and 1 denotes default (see Table 2).

A borrower is in default if repayment is more than 90 days late. This follows international banking and microfinance risk management standards [24].

3.3. Dataset Description

The final analytical dataset consisted of 12,000 instances and 17 predictor variables (Table 3). Preliminary analyzed features and used correlation and mutual information. This helped identify the most predictive variables, which were guided by domain expertise. We retained important demographic features such as age, education, and marital status. We also included economic indicators, such as income, occupation, and household size. They also retained loan-specific details, such as the amount, term, interest rate, collateral, and purpose. Finally, we examined repayment behavior metrics, including the number of past loans, repayment frequency, and any prior defaults. We removed features with low variance and high redundancy. This simplification was beneficial for the model. This selection process matches studies that find that borrower income, loan size, and repayment history are key predictors of default risk. Borrower capacity is important in microfinance credit models. It compares income to the debt. The past repayment history also plays a key role.

Table 3. Credit risk prediction dataset features.

Attribute Name	Attribute Description
Gender	Gender of the client (Male / Female)
Age	Age of the client in years
MaritalStatus	Marital status of the client (Married, Single, Widowed, etc.)
Education	Educational level of the client (Primary, Secondary, Higher Education, etc.)
Residence	Area of residence of the client (Urban / Rural)
Occupation	Occupation category of the client (Agriculture, Handicraft, Business, Other, etc.)
MonthlyIncome	Monthly income earned by the client
HouseholdSize	Number of family members in the household
LoanAmount	Total amount of loan received by the client
LoanTermMonths	Duration of the loan repayment period in months
LoanPurpose	Purpose for taking the loan (Working Capital, Expansion, Personal Use, Other, etc.)
Collateral	Type of collateral or guarantee provided for the loan
PastLoans	Number of previous loans taken by the client
RepaymentFreq	Loan repayment frequency (Weekly, Monthly, etc.)
InterestRate	Interest rate charged on the loan
IncomeToPaymentRatio (IPR)	the ratio of the borrower's monthly income to the expected monthly loan payment.
DebtToIncomeRatio (DTI)	the ratio of total debt obligations to monthly income.
Default	Loan repayment status where Default = 1 and Non-default = 0

3.4. Feature Engineering

The dataset contained 12,000 loan records and 17 predictor variables. Several preprocessing and feature engineering steps were performed to improve data quality and model performance. First, records containing missing values were removed. Next, outliers were identified using the Interquartile Range (IQR) method [8], which helps reduce data noise and improve model generalization by minimizing the influence of extreme observations. The acceptable range for outlier detection was computed as:

$$Q_1 - 1.5 \times (Q_3 - Q_1) < x < Q_3 + 1.5 \times (Q3 - Q_1) \quad (1)$$

where Q1 and Q3 denote the first and third quartiles, respectively.

After outlier treatment, variables that could introduce data leakage were excluded because they would not be available at the time of loan approval and could therefore lead to overly optimistic prediction performance. Categorical variables were converted into numerical representations using one-hot encoding.

For numerical features, variables exhibiting substantial positive skewness were first transformed using the logarithmic transformation:

$$x' = log(x + 1) \quad (2)$$

to reduce skewness and stabilize the variance.

Following the log transformation, the numerical features were scaled according to the requirements of the machine learning algorithms. Min-Max normalization was applied to tree-based models (Decision Tree, Random Forest, and XGBoost) to map feature values to the range [0,1], using:

$$x' = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (3)$$

where X_{min} and X_{max} represent the minimum and maximum values of each feature, respectively.

For algorithms that are sensitive to feature scale, specifically Support Vector Machine (SVM) and Artificial Neural Network (ANN), the numerical features were standardized using the StandardScaler (z-score normalization) instead of Min-Max normalization. Standardization transformed each feature to have zero mean and unit variance, thereby improving optimization and convergence during model training [25]. The two scaling methods were not applied sequentially to the same features; rather, the appropriate scaler was selected according to the requirements of the corresponding machine learning algorithm.

Finally, correlation analysis was performed to identify redundant variables, and low-variance features were removed. Two additional ratio-based features—income-to-payment ratio and an updated debt-to-income ratio—were engineered to improve predictive capability, following approaches reported in previous studies [3].

$$DTI = \frac{Total\ Debt}{Monthly\ Income} \quad (4)$$

These financial ratios highlight the important borrower traits. They also improve the discrimination power. This disciplined, domain-informed feature engineering aligns with the best practices in credit-risk modelling [1], [3].

3.5. Machine Learning Models

We trained five classification models chosen for their success in credit risk tasks. Each of these is briefly described below.

3.5.1. Extreme Gradient Boosting (XGBoost)

A high-performance gradient-boosted tree ensemble [26] was used. XGBoost builds decision trees sequentially using gradient descent to minimize the loss function. This includes regularization to prevent overfitting. In practice, XGBoost has repeatedly outperformed older methods in credit-scoring benchmarks [24], [27]. We tuned parameters such as the learning rate, tree depth, and number of estimators via a grid search.

3.5.2. LightGBM

A gradient boosting framework by Microsoft, optimized for efficiency on large tabular data [24]. LightGBM grows trees leaf-wise and uses histogram-based splitting, making it extremely fast. It often matches or exceeds the speed and accuracy of XGBoost. LightGBM showed strong results, similar to those of previous studies. One study noted that "LightGBM outperformed other models" and had "outstanding performance" on credit data [21].

3.5.3. CatBoost

A boosting algorithm by Yandex that natively handles categorical features. CatBoost uses an "ordered boosting" scheme to prevent target leakage during category encoding. Previous reviews have noted that CatBoost performs at least as well as XGBoost and LightGBM on imbalanced financial datasets [24], [27]. We treated CatBoost similarly by tuning its learning rate, depth, and other hyperparameters.

3.5.4. Random Forest

Random Forest (RF) is an ensemble learning algorithm that constructs multiple decision trees using bootstrap sampling and random feature selection, reducing variance while improving predictive accuracy and robustness to overfitting [25], [28]. Owing to its ability to capture complex nonlinear relationships and handle heterogeneous tabular data, RF remains a reliable benchmark for credit risk prediction [29]. In this study, key hyperparameters, including the number of trees and maximum features, were optimized for comparison with boosting models.

3.5.5. Deep Neural Network (DNN)

A multilayer perceptron has many hidden layers. It uses ReLU for activation and a sigmoid output for default probability. In principle, a DNN can learn complex nonlinear interactions [30]. We used two to three hidden layers with dropout to regularize and trained using the Adam optimizer. However, deep networks require more data and tuning. Many studies have shown that tree ensembles perform better than deep nets on tabular credit data [31], [32]. Therefore, the DNN acts as a helpful benchmark.

3.6. Handling Class Imbalance

Class imbalance is a significant challenge in credit risk prediction, particularly for microfinance institutions (MFIs). In these cases, borrowers who default are a small part of all borrowers. The loan repayment behavior in this study was extremely uneven. Non-default borrowers were much more common than default borrowers. An imbalance can skew the classification algorithms toward the majority class. This may result in missing risky borrowers, even if the accuracy is high. Classifiers trained on imbalanced credit datasets often struggle to identify minority defaults. This occurs because the majority class patterns dominate the learning process. Previous studies have highlighted this issue [20], [24], [33]. To address this challenge, we used resampling strategies. This helped to rebalance the training dataset before developing the model.

3.6.1. Oversampling Techniques

In this study, we used oversampling techniques. This helped us address the severe class imbalance in the credit risk datasets. In credit risk prediction, most borrowers do not default on their loans. Only a small number of borrowers fall into the default category. This imbalance can skew machine learning predictions toward the majority class. As a result, this may lead to poor identification of risky clients. Oversampling methods were preferred. They retain the original majority class data and boost minority class representation by creating synthetic samples. Oversampling techniques enhance the predictive power of machine learning models. This is particularly true for financial risk analytics and imbalanced credit scoring systems [17], [34], [35].

Undersampling methods can drop important financial data by removing majority class observations. This limitation is a major issue in microfinance datasets. Each borrower record holds key socioeconomic and behavioral traits that affect the repayment performance. Oversampling methods boost minority class instances without losing existing data. This improves class separation and helps the model learn complex default patterns. Oversampling methods, such as SMOTE and ADASYN, boost recall, F1-score, and AUC in imbalanced credit prediction tasks [17], [34], [35].

In this study, four oversampling approaches were employed: SMOTE, ADASYN, Borderline-SMOTE, and SVM-SMOTE. These techniques were chosen for their success with nonlinear and imbalanced data. The SMOTE is a popular method for handling imbalanced learning issues. SMOTE generates artificial minority class samples by

interpolating between existing minority observations and their nearest neighbors. This method boosts the diversity of minority instances. It also reduces overfitting linked to random duplication. The mathematical representation of SMOTE is as follows:

$$x_{new} = x_i + \lambda(x_{nn} - x_i), \lambda \in [0,1] \quad (5)$$

where x_i is a sample from the minority class. x_{nn} is one of its closest minority neighbors. λ is a random number between 0 and 1. The new synthetic sample, x_{new}, is between two minority observations. Research by [34] found that SMOTE resampling increased classification accuracy. It also helps identify minority classes in credit risk prediction systems. This is important when using imbalanced financial data.

ADASYN (Adaptive Synthetic Sampling) builds on SMOTE. This generates more synthetic samples for hard-to-learn minority instances. ADASYN differs from the traditional SMOTE. It targets areas with more error classification. This approach helps to improve the decision boundary learning. This adaptive mechanism boosts the classifier's sensitivity to high-risk borrowers. It also strengthens the predictive accuracy in complex financial settings. ADASYN improves the predictive performance in skewed credit-scoring datasets, as shown in earlier studies [17], [36].

Borderline-SMOTE is a unique method. This creates synthetic data near the class decision boundaries. Borderline-SMOTE does not oversample all minority instances equally. Instead, it focuses on samples from risky borderline areas. These are the spots where misclassification is more likely to occur. This targeted sampling strategy helps distinguish between default and non-default borrowers. It also boosts the ability of the classifier to generalize. Borderline-SMOTE helps improve the precision of minority classes in imbalanced financial datasets. It also reduces the classification bias [37].

SVM-SMOTE combines Support Vector Machine (SVM) concepts with synthetic oversampling techniques. This method identifies support vectors that are close to the decision boundary. Then, it creates synthetic minority samples in these areas. Consequently, SVM-SMOTE improves the boundary definition and increases the classifier sensitivity to minority observations. SVM-SMOTE improves the prediction of credit defaults by neural networks and ensemble methods. This is especially true for imbalanced datasets [37], [38].

3.6.2. Under-sampling

This study used undersampling techniques to address the serious class imbalance in credit datasets. In these datasets, non-default borrowers greatly outnumber the default borrowers. Undersampling methods attempt to balance the dataset by reducing the majority class instances. This differs from oversampling, which creates new samples for the minority class. This strategy is effective because many credit risk datasets contain redundant or closely related majority-class data. This can skew machine learning algorithms to predict the dominant class. Consequently, financially risky borrowers may be overlooked. Undersampling techniques reduce computational complexity and training time. This is key in resource-limited financial settings and real-time credit-scoring systems [39].

In this part of the study, we used three key undersampling techniques: Random Under Sampling (RUS), Cluster Centroids (CC), and edited nearest neighbors (ENN). These methods were chosen because they exhibit different types of undersampling strategies. These include random elimination, prototype generation, and noise-cleaning approaches. Their combined application allowed a full evaluation of how various data reduction methods affect credit risk prediction in imbalanced financial data sets.

Random undersampling is one of the simplest and most widely used imbalance-handling techniques. This method randomly removes samples from the majority class. This process continues until the class distribution is balanced. Let Nm represent the number of minority class samples and NM represent the number of majority class samples, where NM > Nm. RUS reduces the majority class such that:

$$N_M{}' = N_m \quad (6)$$

where $N_M{}'$ denotes the new reduced majority class size after the sampling. RUS balances the dataset and reduces the training costs. It may miss key information by randomly dropping helpful majority-class observations [40]. Nevertheless, previous studies have demonstrated that RUS can improve minority-class sensitivity and enhance classifier fairness in financial risk prediction problems.

Cluster Centroids is a smarter undersampling method. It replaces groups of majority-class instances with their corresponding centroids. This is better than removing observations randomly. This method uses the K-means clustering algorithm for the majority class. It finds centroid points that maintain the overall distribution structure. Suppose that the majority class is partitioned into K clusters *C1, C2, ..., Ck*. The centroid of each cluster is calculated as follows:

$$\mu_j = \frac{1}{|C_j|} \sum_{x_i \epsilon C_j} x_i \quad (7)$$

where μ_j represents the centroid of cluster C_j and x_i denotes the feature vectors within the cluster. This method reduces redundancy by replacing many majority samples with centroid representatives. This retains important

Table 4. hyperparameter selected values for the DNN.

Hyperparameter	Selected Value
Random state	42
Input layer neurons	128
Dropout rate	0.30, 0.20
Batch size	32
Optimizer	Adam
Loss function	Binary cross-entropy
Output activation	Sigmoid
Activation function	ReLU

structural information [4]. In credit scoring, Cluster Centroids can boost performance. They maintain the traits of low-risk borrowers and help balance the dataset.

Edited nearest neighbors (ENN) is a method for cleaning data and reducing noise. It removes most samples from the majority class, which the nearest neighbors are likely to misclassify. For each sample x_i, the algorithm finds its K-nearest neighbors. The sample is removed if its class label is different from the majority label of those neighbors. A team formally removes a sample when:

$$y_i \neq mode(NN_k(x_i)) \quad (8)$$

where y_i is the class label of sample x_i, and $NN_k(x_i)$ denotes the set of K-nearest neighbors. ENN enhances class separability by removing noisy and borderline majority-class instances. This can confuse machine learning models. This trait is particularly useful in microfinance credit datasets. Here, noisy borrower records and overlapping financial profiles often occur.

Overall, the selected undersampling methods provide complementary strategies for addressing class imbalances in Ethiopian microfinance credit risk prediction. Their comparison helps identify preprocessing techniques that can boost the predictive performance and fairness of credit scoring systems in new financial settings.

3.6.3. Hybrid Sampling Methods for Imbalanced Credit Risk Prediction

Credit risk prediction faces several challenges. Non-default borrowers greatly outnumber the borrowers who default. This imbalance causes classifiers to favor the majority class. Consequently, they struggle to predict high-risk borrowers effectively. In microfinance, these limits can harm loan quality, hurt sustainability, and impact financial inclusion. To address this challenge, we used hybrid resampling techniques. We focused on SMOTETomek and SMOTE-ENN methods. These methods effectively balance the datasets. They improve class separation and reduce noise in the credit data.

Among the hybrid approaches, SMOTETomek combines SMOTE oversampling with Tomek Link undersampling. After SMOTE generates synthetic minority samples, Tomek Links identifies pairs of nearest-neighbor instances from different classes. We removed these unclear majority-class instances. This creates a cleaner and clearer decision boundary. This hybrid strategy enhances class separability. This reduces the overlap between defaulters and non-defaulters. This is key in credit scoring, especially with noisy financial records and varying borrower behaviors. According to previous studies, SMOTETomek improves important metrics such as F1-score, Recall, and ROC-AUC for credit default prediction systems [21], [41]. Similarly, SMOTE-ENN integrates SMOTE with the Edited Nearest Neighbours (ENN) algorithm. Following oversampling, ENN removes misclassified and noisy samples by analyzing neighborhood consistency. SMOTE-ENN creates a better-balanced dataset than traditional oversampling methods. This is achieved by removing borderline and inconsistent observations. This method is particularly suitable for Ethiopian MFIs, where borrower records often contain informal income information, incomplete financial histories, and manually recorded data. These include informal income, missing financial history, and manual records. SMOTE-ENN enhances the strength and reliability of models in imbalanced financial datasets. Evidence supports this claim [21], [42]. These hybrid sampling methods provide an effective framework for addressing class imbalance in credit risk prediction for MFIs. They reduce classification bias and improve predictive performance.

3.7. Model Configuration and Hyperparameter Optimization

To ensure optimal performance and eliminate structural bias across imbalanced evaluation scenarios, hyperparameter optimization was systematically executed using a stratified grid search paired with 10-fold cross-validation (k=10) on the training subset (70). Stratified sampling was strictly applied during the data partitioning step (70/30 train–test ratio) to maintain the original empirical class distribution (approximately 1:9 ratio of minority to majority classes) across all folds.

The hyperparameter search space was designed around the core regularization parameters, structural depth constraints, learning rates, and tree ensemble behaviors to mitigate potential overfitting in the minority class. The model performance during the grid search was directly guided by the Area Under the Precision-Recall Curve (PR-AUC) and Macro-F1 score, ensuring that parameter selection prioritized non-trivial classification of the high-risk minority class rather than overall accuracy.

3.7.1. Deep Neural Network (DNN) Architecture and Hyperparameter Search Space

A Deep Neural Network (DNN) architecture was designed to handle high-dimensional feature interactions without succumbing to internal covariate shifts or rapid overfitting. The grid search for the DNN evaluated

Table 5. Final hyperparameter values for tree-based ensemble algorithms.

XGBoost		**LightGBM**	
Hyperparameter	**Selected Value**	**Hyperparameter**	**Selected Value**
n_estimators	300	n_estimators	300
learning_rate (η)	0.05	learning_rate	0.03
max_depth	5	num_leaves	31
Subsample	0.80	max_depth	7
colsample_bytree	0.80	min_child_samples	20
scale_pos_weight	8.5	subsample bagging)	0.80
gamma (γ)	0.10	scale_pos_weight	8.5
Random Forest		**CatBoost**	
Hyperparameter	**Selected Value**	**Hyperparameter**	**Selected Value**
n_estimators	300	iterations	600
max_depth	20	learning_rate	0.03
min_samples_split	5	depth	6
min_samples_leaf	2	l2_leaf_reg	5
max_features	'sqrt'	auto_class_weights	Balanced
class_weight	'balanced_subsample'		

Table 6. Evaluation Metrics.

Metric	**Formula**	**Definition**
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Overall correctness (useful but can be misleading under imbalance).
Precision	$\frac{TP}{TP+FP}$	(for the default class): Of all predicted defaulters, the fraction truly defaulted.
Recall	$\frac{TP}{TP+FN}$	Sensitivity: The fraction of actual defaulters correctly identified.
F1-Score	$\frac{2\ X\ Precision\ X\ Recall}{Precision\ +\ Recall}$	Harmonic mean of precision and recall, balancing both.
ROC-AUC	$\frac{1+\ TPR-FPR}{2}$	Area under the Receiver Operating Characteristic curve, measuring overall discrimination ability.
G-Mean	$\sqrt{recall\ +specificity}$	Geometric mean of sensitivity and specificity, which rewards models that perform well on both classes.

combinations of structural depth, layer width, dropout probability, and optimization parameters (see Table 4).

The study methods build on established credit-scoring research and are adapted to the microfinance context. The severe imbalance in defaults (approximately 10%) is a well-known challenge [22], [27]. We apply hybrid resampling (SMOTE-ENN) [24], [43] and use powerful ensemble classifiers [24]. As past reviews have indicated, boosting algorithms perform better with proper weighting or resampling [26]. Ref [31] shows that methods such as XGBoost, LightGBM, and CatBoost can focus on the minority class. Using resampling with these models leads to "heightened robustness" [26]. The study design followed the best practices. This includes careful data cleaning, preparing features, managing imbalance, and using metrics focused on minorities [4], [26]. This approach ensures that our findings reliably show which models work best for predicting microfinance defaults in real-world situations.

3.7.2. Grid Search Space and Selected Parameters for Tree Ensembles

For gradient boosted decision tree (GBDT) algorithms, specifically XGBoost, LightGBM, and CatBoost early, stopping based on validation loss was incorporated with a patience threshold of 50 rounds to prevent overfitting. For Random Forest, tree diversity was enforced using bootstrap sampling combined with square-root feature subsampling at each split. Table 5 summarizes the defined search bounds and the exact hyperparameter configurations selected via the grid search process.

3.8. Experiment tools

All experiments were performed using Google Collaboratory. It offers free computing resources, such as CPUs, GPUs, and TPUs. In our case, we used a T4 GPU owing to its free access, fast runtime, and performance. We

Table 7. Classification Performance of Models on the Original Imbalanced Dataset.

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	G-Mean
XGBoost	0.968	0.921	0.918	0.919	0.979	0.945
LightGBM	0.965	0.909	0.905	0.907	0.978	0.938
CatBoost	0.971	0.934	0.934	0.934	0.979	0.954
Random Forest	0.962	0.899	0.876	0.887	0.977	0.922
DNN	0.973	0.935	0.959	0.947	0.980	0.967

used an HP laptop with an Intel(R) Core(TM) i5-4210M CPU @ 2.60GHz. It has a 1TB HD and 8GB RAM. The laptop ran on a 64-bit Microsoft Windows 10 operating system to tool the model. Data manipulation was performed using Pandas and NumPy. For preprocessing and classical models, we relied on the scikit-learn library. The boosting models were XGBoost, LightGBM, and CatBoost. Finally, we built a neural network using Keras and TensorFlow. Resampling methods used the imbalanced-learn package. We fixed the random seeds and recorded the library versions to ensure reproducibility. This software stack is standard in financial ML and supports transparent and replicable research.

4. Results

This section analyzes the accuracy of the predictions made by the different machine learning models. We considered XGBoost, LightGBM, CatBoost, Random Forest, and Deep Neural Network (DNN). These models were chosen for their strong ability to predict financial data. They are also increasingly used in credit-scoring research. We evaluated the performance using several metrics: Accuracy, Precision, Recall, F1-score, ROC-AUC, and Geometric Mean (G-Mean) (see Table 6). These metrics provide a better understanding of the effectiveness of classification in imbalanced datasets. Accuracy alone does not capture the full picture [44], [45].

4.1. Experiment results with original imbalanced dataset (No Sampling)

Table 7 shows the performance of the classifier on the original imbalanced dataset. Here, default borrowers comprise approximately 9–10% of all loan records. All five models showed high accuracy, even with class imbalances. They effectively used patterns in borrower demographics, finances and loan details.

Table 7 presents the baseline performances of the five classification models evaluated on the original imbalanced credit risk dataset without applying any sampling technique. Overall, the findings indicate that the Deep Neural Network (DNN) achieved the most robust predictive performance across nearly all evaluation metrics. Specifically, the DNN attained the highest ROC-AUC (0.980) and G-Mean (0.967), with an accuracy of 97.3% and an F1-score of 0.947. These results demonstrate its superior ability to discriminate between default and non-default borrowers while maintaining a balanced classification performance, despite the inherent class imbalance. The effectiveness of the DNN can be attributed to its capacity to learn complex nonlinear interactions among heterogeneous financial and demographic variables that characterize credit risk.

CatBoost ranked as the second-best baseline model, achieving an ROC-AUC of 0.979 and a G-Mean of 0.954, with an overall accuracy of 97.1% and an F1-score of 0.934. Its competitive performance highlights the effectiveness of gradient boosting with ordered boosting and efficient handling of categorical attributes, making it particularly suitable for structured credit data sets.

XGBoost and LightGBM also demonstrated strong discriminative capabilities, producing ROC-AUC values of 0.979 and 0.978, respectively. However, their comparatively lower G-Mean values (0.945 and 0.938) indicate reduced sensitivity to the minority default class under imbalanced conditions. Random Forest exhibited the weakest baseline performance, recording the lowest F1-score (0.887) and G-Mean (0.922), suggesting a greater tendency to favor the majority class. Overall, the results establish the DNN as the strongest baseline classifier without sampling, with CatBoost emerging as the closest competing method.

4.2. Experiment results with Oversampling Strategies

Four oversampling techniques were used on the training data before fitting the model: SMOTE, Adaptive Synthetic Sampling (ADASYN), Borderline-SMOTE, and SVM-SMOTE. Table 8 presents the comprehensive results for all sampling-model combinations.

The application of synthetic oversampling techniques substantially enhanced the ability of all classifiers to identify minority class instances, leading to consistent improvements in recall, F1-score, G-mean, and ROC-AUC compared with learning on the original imbalanced dataset. Among the evaluated oversampling methods, DNN with Borderline-SMOTE achieved the most balanced and robust predictive performance. As presented in Table 8, this model achieved an accuracy of 0.973, recall of 0.972, F1-score of 0.946, ROC-AUC of 0.980, and the highest G-Mean of 0.972, indicating excellent discrimination between default and non-default borrowers while maintaining balanced sensitivity across both classes. These findings suggest that generating synthetic samples near the decision boundary effectively improves the classification in regions where class overlap is most pronounced.

Table 8. Classification Performance of Models Oversampling.

Sampling	Model	Accuracy	Precision	Recall	F1- Score	ROC-AUC	G-Mean
SMOTE	XGBoost	0.966	0.899	0.926	0.912	0.979	0.948
SMOTE	LightGBM	0.965	0.898	0.918	0.908	0.978	0.944
SMOTE	CatBoost	0.970	0.919	0.938	0.928	0.979	0.956
SMOTE	Random Forest	0.959	0.851	0.909	0.879	0.976	0.937
SMOTE	DNN	0.975	0.955	0.959	0.957	0.980	0.968
ADASYN	XGBoost	0.967	0.899	0.930	0.914	0.979	0.950
ADASYN	LightGBM	0.966	0.902	0.918	0.910	0.978	0.944
ADASYN	CatBoost	0.970	0.915	0.947	0.931	0.979	0.960
ADASYN	Random Forest	0.960	0.858	0.913	0.885	0.976	0.939
ADASYN	DNN	0.974	0.959	0.942	0.951	0.980	0.960
Borderline-SMOTE	XGBoost	0.966	0.899	0.926	0.912	0.979	0.948
Borderline-SMOTE	LightGBM	0.965	0.898	0.913	0.906	0.978	0.942
Borderline-SMOTE	CatBoost	0.968	0.911	0.934	0.922	0.979	0.953
Borderline-SMOTE	Random Forest	0.958	0.849	0.892	0.870	0.976	0.928
Borderline-SMOTE	DNN	0.973	0.921	0.972	0.946	0.980	0.972
SVM-SMOTE	XGBoost	0.964	0.881	0.926	0.903	0.978	0.947
SVM-SMOTE	LightGBM	0.965	0.885	0.934	0.909	0.978	0.951
SVM-SMOTE	CatBoost	0.968	0.904	0.942	0.923	0.979	0.957
SVM-SMOTE	Random Forest	0.960	0.835	0.938	0.884	0.976	0.950
SVM-SMOTE	DNN	0.967	0.871	0.972	0.919	0.980	0.969

Table 9. Classification Performance of Models Undersampling.

Sampling	Model	Accuracy	Precision	Recall	F1- Score	ROC-AUC	G-Mean
RUS	XGBoost	0.950	0.753	0.976	0.851	0.977	0.962
RUS	LightGBM	0.950	0.748	0.976	0.848	0.977	0.961
RUS	CatBoost	0.946	0.728	0.972	0.833	0.977	0.957
RUS	Random Forest	0.928	0.640	0.968	0.771	0.974	0.945
RUS	DNN	0.958	0.811	0.963	0.881	0.978	0.961
Cluster Centroids	XGBoost	0.920	0.606	0.976	0.748	0.975	0.944
Cluster Centroids	LightGBM	0.914	0.583	0.972	0.730	0.974	0.939
Cluster Centroids	CatBoost	0.922	0.614	0.976	0.755	0.975	0.946
Cluster Centroids	Random Forest	0.905	0.551	0.968	0.704	0.971	0.932
Cluster Centroids	DNN	0.958	0.801	0.972	0.878	0.979	0.964
ENN	XGBoost	0.956	0.803	0.947	0.869	0.978	0.952
ENN	LightGBM	0.957	0.809	0.947	0.872	0.977	0.952
ENN	CatBoost	0.957	0.804	0.955	0.873	0.978	0.956
ENN	Random Forest	0.950	0.759	0.951	0.845	0.975	0.950
ENN	DNN	0.968	0.875	0.972	0.921	0.979	0.969

Table 10. Classification Performance of Models Hybrid Sampling.

Sampling	Model	Accuracy	Precision	Recall	F1- Score	ROC-AUC	G-Mean
SMOTETomek	XGBoost	0.966	0.899	0.926	0.912	0.979	0.948
SMOTETomek	LightGBM	0.965	0.898	0.918	0.908	0.978	0.944
SMOTETomek	CatBoost	0.970	0.919	0.938	0.928	0.979	0.956
SMOTETomek	Random Forest	0.959	0.851	0.909	0.879	0.976	0.937
SMOTETomek	DNN	0.974	0.951	0.947	0.949	0.980	0.972
SMOTE-ENN	XGBoost	0.951	0.761	0.963	0.851	0.978	0.956
SMOTE-ENN	LightGBM	0.949	0.754	0.951	0.841	0.977	0.950
SMOTE-ENN	CatBoost	0.950	0.756	0.963	0.848	0.978	0.956
SMOTE-ENN	Random Forest	0.941	0.705	0.959	0.813	0.974	0.949
SMOTE-ENN	DNN	0.959	0.809	0.972	0.883	0.979	0.964

The conventional SMOTE and ADASYN strategies also produced substantial performance improvements. With SMOTE, the DNN attained the highest F1-score (0.957) and an ROC-AUC of 0.980, whereas ADASYN achieved the highest precision (0.959) while maintaining an ROC-AUC of 0.980 and an F1-score of 0.951. Although

ADASYN yielded a slightly lower recall than borderline-SMOTE, its adaptive generation of minority samples contributed to a more conservative yet precise decision boundary.

Among the ensemble learning algorithms, CatBoost consistently delivered the strongest performance under all oversampling methods, recording F1-scores between 0.922 and 0.931 and achieving the highest ensemble G-Mean of 0.960 under ADASYN. XGBoost and LightGBM followed closely, demonstrating stable predictive behavior with ROC-AUC values of approximately 0.978–0.979 across all sampling strategies. In contrast, Random Forest exhibited comparatively lower F1-scores (0.870–0.885) and G-Mean values (0.928–0.950), indicating a reduced capacity to accurately characterize the minority class despite synthetic resampling. Overall, the findings demonstrate that combining advanced oversampling techniques with deep learning substantially improves credit risk prediction in highly imbalanced datasets, with the DNN–Borderline-SMOTE framework providing the most reliable and well-balanced classification.

4.3. Experiment results with Undersampling Strategies

We examined three undersampling methods: random under-sampler (RUS), Cluster Centroids (CC), and edited nearest neighbors (ENN). Table 9 presents the results. The performance of the undersampling techniques revealed distinct trade-offs across all classification models. Aggressive undersampling methods, particularly Random Under-Sampling (RUS) and Cluster Centroids, consistently improved the identification of minority-class instances by increasing recall. For example, the RUS–XGBoost model achieved a recall of 0.976, 0.95 accuracy, an F1-score of 0.851, an ROC-AUC of 0.977, and a G-Mean of 0.962. However, this improvement in sensitivity was accompanied by a substantial reduction in precision, ranging from 0.551 to 0.753, indicating a higher rate of false positive predictions. Similar behavior was observed for LightGBM and CatBoost, confirming that the aggressive removal of majority-class samples favors minority detection at the expense of classification specificity.

Among the evaluated approaches, the Cluster Centroids produced the least favorable outcomes. Replacing majority-class observations with centroid representations results in excessive information loss, weakening the class boundary definition, and reducing the discriminative capability. Consequently, the Random Forest classifier recorded the lowest performance, with 90.5% accuracy, an F1-score of 0.704, precision of 0.551, and a G-mean of 0.932, despite maintaining a relatively high recall.

In contrast, the Edited Nearest Neighbors (ENN) demonstrated the most balanced and robust performance across all classifiers. By eliminating ambiguous majority-class samples while preserving informative observations, the ENN generates clearer decision boundaries and improves the overall classification stability. The DNN–ENN model achieved the strongest overall performance, with an accuracy, 0.875 precision, 0.972 recall, F1-score, ROC-AUC, and G-Mean of 96.8 %, 0.875, 0.972, 0.921, 0.979, and 0.969, respectively. These findings indicate that although aggressive undersampling enhances recall, ENN provides superior precision–recall stability, making it the most appropriate strategy for credit risk prediction, where accurate default detection must be balanced against minimizing false-positive classifications.

4.4. Experiment results with Hybrid Sampling Strategies

We evaluated two hybrid resampling techniques: SMOTETomek and SMOTE-ENN. We wanted to determine whether using synthetic minority oversampling with targeted majority-class cleaning is better than using each method alone. Table 10 presents these results. The results demonstrated distinct performance characteristics between the SMOTETomek and SMOTE-ENN hybrid resampling strategies. Based on Table 10, the Deep Neural Network (DNN) combined with SMOTETomek achieved the most balanced overall performance, attaining an accuracy of 0.974, precision of 0.951, recall of 0.947, F1-score of 0.949, ROC-AUC of 0.980, and G-Mean of 0.972. These findings indicate that integrating synthetic minority oversampling with Tomek Links effectively improves class separability while preserving the optimal balance between sensitivity and specificity. Notably, the ROC-AUC value of 0.980 represents the highest empirical discrimination performance observed across all the experimental models (Tables 7–10). Among the ensemble learning algorithms, CatBoost consistently delivered competitive predictive performance, whereas XGBoost and LightGBM achieved comparable classification effectiveness with slightly lower overall discriminations.

In contrast, SMOTE-ENN prioritized minority-class detection, with the DNN producing the highest recall (0.972) but at the expense of reduced precision (0.809) and overall accuracy (0.959), reflecting the more aggressive removal of borderline majority-class samples by Edited Nearest Neighbors. Overall, SMOTETomek provided the most robust and well-balanced predictive framework for credit risk classification, offering superior generalization and reliable discrimination across both majority and minority classes.

Figures 1 and 2 show that all the machine learning models have high ROC-AUC values. This was true for the different sampling techniques. This means that they can effectively distinguish between default and non-default borrowers. The Deep Neural Network (DNN) stood out among the classifiers. It reached the top ROC-AUC values, hitting 98% with SMOTE, ADASYN, Borderline-SMOTE, and SMOTETomek. This demonstrates its strength in

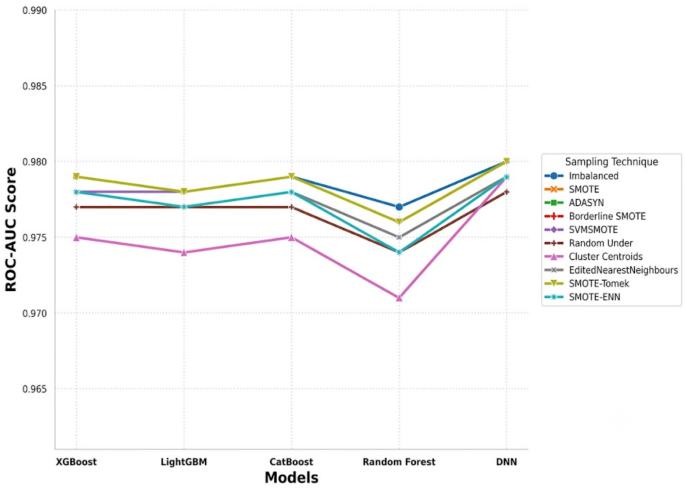

Figure 1. ROC_AUC Comparison Across and sampling.

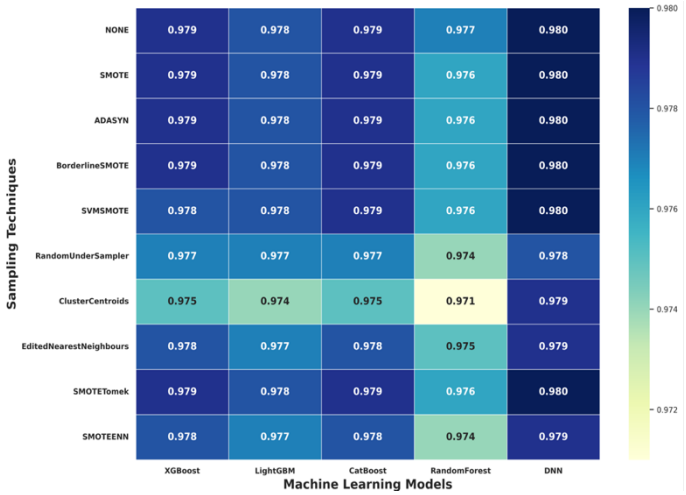

Figure 2. ROC_AUC Heatmap analysis Models with sampling Techniques.

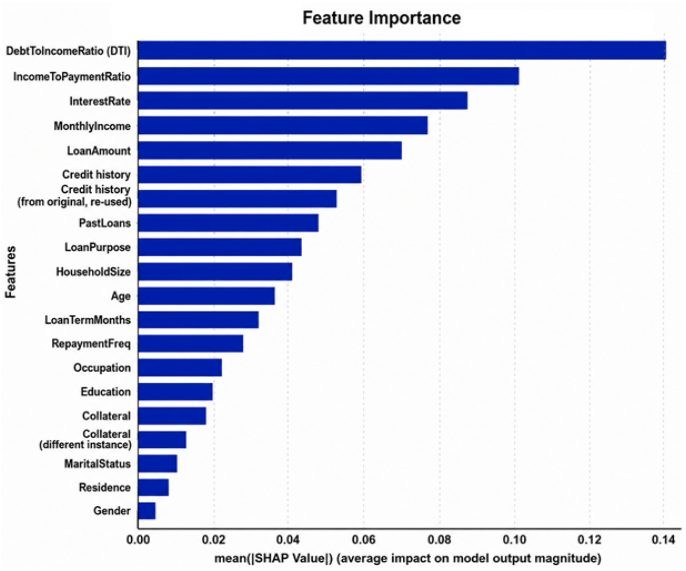

Figure 3. Feature importance analysis with SHAP.

capturing complex nonlinear relationships in imbalanced microfinance credit data. CatBoost and XGBoost showed strong performance, with ROC-AUC values ranging from 0.979 to 0.980 across most sampling methods. This highlights the robustness of gradient boosting algorithms in structured financial datasets. Random Forest had lower ROC-AUC scores, especially with ClusterCentroids (0.971). This suggests that it is more sensitive to information loss from aggressive undersampling. Oversampling and hybrid methods often perform better than undersampling. Examples include SMOTE, Adaptive Synthetic Sampling (ADASYN), Borderline-SMOTE, and SMOTETomek. The findings show that using resampling techniques with deep learning improves credit risk prediction in Ethiopian microfinance institutions.

4.5. Model Interpretability Using SHAP

After attaining a robust predictive performance while reducing the adverse effects of class imbalance, the proposed credit risk prediction framework demonstrated a balanced classification across both low- and high-risk borrower groups. To further enhance model transparency, Explainable Artificial Intelligence (XAI) was incorporated through SHAP (SHapley Additive exPlanations), enabling a reliable interpretation of individual feature contributions without introducing bias toward the majority class. By assigning consistent Shapley values to each predictor, SHAP provides a comprehensive explanation of how the input variables influence the model's decisions, thereby improving confidence in the proposed machine learning approach.

Figure 3 presents the SHAP-based global feature importance analysis of the proposed credit risk prediction model using the imbalanced microfinance dataset. The results indicate that the Debt-to-Income Ratio (DTI) is the most influential predictor of credit risk, with a mean SHAP value of approximately 0.14, demonstrating its dominant contribution to the model's decision-making process. This feature is followed by the IPR and Interest Rate, with SHAP values of approximately 0.10 and 0.09, respectively, highlighting the critical role of borrowers' repayment capacity and loan pricing in determining default risk. Other highly influential variables include Monthly Income, Loan Amount, and PastLoans, all of which have a substantial impact on the prediction outcomes. Variables such as Past Loans, Loan Purpose, Household Size, and Age contribute moderately to the model, indicating that both financial and demographic characteristics influence borrower creditworthiness. In contrast, Marital Status, Residence, and Gender showed relatively low SHAP values, suggesting a limited influence on the model predictions. The findings demonstrate that the proposed machine learning framework primarily relies on economically meaningful indicators associated with repayment ability and borrowing behavior. Furthermore, SHAP analysis enhances model transparency by providing clear

explanations of feature contributions, thereby improving the interpretability, reliability, and trustworthiness of credit risk assessments in microfinance lending environments.

5. Discussion

The findings in this section show that combining machine learning with data-level imbalance correction improves the accuracy of credit risk prediction. This approach is better than traditional statistical scoring methods. The next section covers the findings of credit risk research. They examine why some performance patterns occur. They also highlight the limitations of their methods and discuss the implications for practice and future research.

5.1. Class Imbalance Correction

The experimental results demonstrate that addressing class imbalance is essential for improving the reliability of credit default prediction in microfinance institution (MFI) datasets. Among the evaluated resampling strategies, hybrid sampling methods consistently produced a more balanced classification performance than conventional oversampling or aggressive undersampling techniques. In particular, the Deep Neural Network (DNN) integrated with SMOTETomek achieved the best overall performance, recording an accuracy of 97.4%, precision of 95.1%, recall of 94.7%, an F1-score of 94.9%, an ROC-AUC of 0.980, and a G-Mean of 0.972. These results indicate that combining synthetic minority oversampling with Tomek Link cleaning enables the model to effectively distinguish between default and non-default borrowers while maintaining a strong balance between sensitivity and specificity.

The superior performance of the SMOTETomek–DNN framework can be attributed to its ability to generate informative minority class samples while simultaneously removing ambiguous observations located near the class boundaries. This process improves class separation, reduces the overlap between borrower groups, and enables the model to learn complex credit risk patterns more effectively. Although other oversampling techniques, including SMOTE, ADASYN, and Borderline-SMOTE, also enhanced minority class detection, their overall performance was slightly lower than that of the hybrid approach. Conversely, undersampling methods, such as Random UnderSampling and Cluster Centroids, increased recall but reduced precision because valuable majority-class information was discarded, leading to more false-positive predictions. Overall, the findings suggest that the DNN combined with SMOTETomek provides the most accurate and operationally reliable framework for handling class imbalances in microfinance credit risk prediction.

5.2. Deep Neural Networks versus Gradient Boosting Ensembles

The comparative evaluation revealed that the Deep Neural Network (DNN) consistently achieved the strongest predictive performance across most sampling strategies, although the performance gap relative to gradient boosting ensembles remained modest. This finding differs from many previous credit risk studies, where boosting algorithms, particularly XGBoost, LightGBM, and CatBoost, are frequently reported as the most effective classifiers [6], [24], [27]. A plausible explanation lies in the heterogeneous nature of the dataset, which combines demographic, socioeconomic, behavioral, and institutional variables. Such data may exhibit complex nonlinear relationships that deep neural architectures can capture more effectively through hierarchical feature learning.

The integration of advanced resampling techniques further strengthened the DNN. In particular, pairing the DNN with SMOTETomek produced the highest overall discrimination (ROC-AUC = 0.980) and an excellent balance between sensitivity and specificity (G-Mean = 0.972). An identical G-Mean (0.972) was also obtained with Borderline-SMOTE, indicating that the DNN effectively learned decision boundaries from synthetically balanced training data while maintaining robust generalization through dropout and batch normalization.

Despite these advantages, CatBoost demonstrated highly competitive and stable performance across multiple sampling configurations. It consistently achieved a ROC-AUC of 0.979 under the unresampled, SMOTE, ADASYN, Borderline-SMOTE, and SMOTETomek datasets, while attaining a maximum G-Mean of 0.960 with ADASYN. Given its inherent ability to process categorical variables efficiently, lower computational complexity, and greater interpretability, CatBoost remains an attractive alternative for microfinance institutions and other resource-constrained financial organizations where transparent and operationally efficient credit risk assessment models are essential.

5.3. Selection of Optimal Resampling Strategy

Choosing a classifier is important, but comparing resampling strategies is also crucial. This study offers useful insights for practitioners. The following hierarchy of resampling effectiveness emerged from the experimental evidence: SMOTE-based oversampling methods, such as SMOTE, Borderline-SMOTE (BR-SMOTE), adaptive synthetic sampling (ADASYN), and SVM-SMOTE, produced the best precision-recall profiles. Borderline-SMOTE had the highest G-Mean of 0.972. SVM-SMOTE achieved the highest recall among all oversampling methods. SMOTE works well in this case because it creates synthetic minority instances. This is achieved by blending existing default cases with their k-nearest neighbors. This improves the

Table 11. The comparison of previous studies with the proposed model.

Reference	Proposed Method / Model	Recall (%)	F1-Score (%)	ROC-AUC (%)	G-mean (%)
[8]	Random Forest + Random Under-Sampling (RUS)	76.54	—	85.34	78.41
[16]	XGBoost Multi-class Classifier	68.00	74.00	85.00	—
[17]	WSMOTE-Ensemble (Random Forest)	—	—	84.10	83.50
[19]	WHSBoost (Weighted-Hybrid-Sampling-Boost)	93.28	82.98	—	90.60
[15]	Optimized Deep MLP (with focal loss / class weights)	76.50	61.20	87.90	—
[46]	SMOTE + XGBoost Ensemble Framework	71.43	65.22	81.00	—
Proposed Study	**SMOTETomek +DNN**	94.70	94.90	98.00	97.20

decision boundaries. It also helps avoid the loss of important information, which can occur with undersampling. ENN undersampling was the best variant. It approaches the oversampling-level G-Mean (DNN-ENN: 0.969). This method removes ambiguous majority class observations with precision. This study focuses on people with inconsistent neighborhood labeling. It does not target random or centroid-replacing majority instance. This selectivity maintains the important structure of non-defaulted borrower records. It also helps to better separate different classes.

Hybrid methods, such as SMOTETomek and SMOTE-ENN, often matched the results of SMOTE alone. This means that the extra boundary cleaning of Tomek Links or ENN is small. SMOTE already significantly improved the boundaries. SMOTETomek offers a lower level of precision than SMOTE-ENN. It is best for cases where cutting down on false-positive loan denials is important but still spotting high defaults.

5.4. Comparison with Prior Literature

This study fills this gap in the credit risk research. It uses a detailed multi-model and multisampling approach. This study focuses on data from a less explored area, unlike bank credit or peer-to-peer lending. Previous studies on African microfinance credit scoring offer key benchmarks for this study. For example, [17] used SMOTENC with ensemble trees on a micro-lending dataset. Meanwhile, [21] achieved 99% accuracy using XGBoost on agricultural microfinance data. This includes environmental factors. The DNN-SMOTE setup in this study showed excellent results: accuracy was 97.5%, recall was 95.9%, and F1 was 95.7%. This meets or beats the benchmarks in Ethiopia, proving that the method works well.

The superior performance of the DNN can be attributed to its ability to model complex nonlinear relationships among the engineered financial, demographic, and behavioural features. When combined with the SMOTETomek resampling strategy, the network learned more balanced decision boundaries for minority-class borrowers, while dropout regularization and early stopping improved generalization by reducing overfitting. Although the DNN achieved the best overall results, the performance gap compared with CatBoost and XGBoost remained modest, indicating that both deep learning and gradient-boosting approaches provide reliable solutions for practical credit risk prediction in microfinance institutions.

Table 11 presents a comparative evaluation of the proposed SMOTETomek + Deep Neural Network (DNN) framework against previously reported credit risk prediction approaches. The results demonstrate that the proposed model provides a highly effective solution for addressing class imbalance challenges in microfinance credit datasets. The proposed approach achieved a Recall of 94.7%, F1-score of 94.9%, ROC-AUC of 98.0%, and G-Mean of 97.2%, indicating superior capability in detecting high-risk borrowers while maintaining strong classification reliability.

Compared with conventional machine learning and ensemble-based techniques, including Random Forest with Random Under-Sampling, XGBoost-based classifiers, WSMOTE-Ensemble, WHSBoost, optimized Deep MLP models, and SMOTE-XGBoost frameworks, the proposed method demonstrates improved balance between minority-class recognition and overall predictive performance. Previous studies often achieved competitive ROC-AUC values but showed limitations in recall and G-Mean, reflecting difficulties in identifying default cases within highly imbalanced financial datasets. In contrast, the integration of SMOTETomek resampling with DNN enables effective learning of minority-class patterns while reducing potential bias toward the majority non-default class.

The superior performance of the proposed framework highlights the importance of combining advanced deep learning architectures with appropriate imbalance-handling strategies for credit risk assessment. Furthermore, the findings emphasize the predictive relevance of borrower characteristics, including repayment behaviour, collateral information, loan attributes, and socioeconomic factors, particularly in developing financial environments with limited credit history availability. Overall, the proposed SMOTETomek + DNN model provides a robust and scalable approach for improving automated credit decision systems in microfinance institutions operating under imbalanced and data-constrained conditions.

6. Conclusion

This study presents a comprehensive evaluation of machine learning and deep learning frameworks combined with advanced data resampling techniques for addressing severe class imbalance in credit risk prediction using Ethiopian microfinance institution (MFI) loan datasets. The research investigates the effectiveness of nine imbalance-handling strategies integrated with multiple predictive models to improve the identification of high-risk borrowers, particularly the minority default class that is often overlooked by conventional learning approaches. The experimental framework involved data preprocessing, feature engineering, model development, and comparative performance analysis using key evaluation metrics, including Accuracy, Precision, Recall, F1-score, ROC-AUC, and Geometric Mean (G-Mean). The empirical findings demonstrate that incorporating imbalance-aware learning mechanisms significantly enhances default prediction capability. Among the evaluated approaches, the Deep Neural Network combined with SMOTETomek achieved the most reliable performance, obtaining 97.4% Accuracy, 95.1% Precision, 94.7% Recall, 94.9% F1-score, 0.980 ROC-AUC, and 0.972 G-Mean. These results indicate that hybrid resampling methods effectively improve minority-class representation and enable models to capture complex patterns associated with borrower default behavior. Furthermore, CatBoost demonstrated competitive predictive performance while maintaining computational efficiency, making it a practical alternative for real-world credit assessment applications. The proposed framework provides Ethiopian MFIs with a data-driven approach to automate borrower evaluation, strengthen lending decisions, reduce non-performing loans, and enhance financial sustainability. Future research should explore explainable artificial intelligence methods, real-time financial behavior data, and macroeconomic factors to improve model transparency, adaptability, and generalization under evolving economic conditions.

7. Recommendations for Future Research

Future research should investigate cost-sensitive optimization techniques and advanced loss functions as alternatives to conventional resampling strategies for highly imbalanced credit datasets. Greater attention should also be given to attention-enhanced and transformer-based architectures, federated learning, and temporal modeling to strengthen predictive performance while preserving data privacy. Moreover, validating these approaches across diverse financial institutions and international datasets, alongside richer explainability methods and adaptive online learning, will improve model robustness, interpretability, and practical deployment in evolving credit risk environments.

8. Declarations

8.1. Author Contributions

Tiruneh Kebede Dubale: Conceptualization, Methodology, Formal analysis, Writing – Original Draft; **Siraj Sebhatu Seyoum:** Validation, Investigation, Writing - Review & Editing; Supervision.

8.2. Institutional Review Board Statement

Not applicable.

8.3. Informed Consent Statement

Not applicable.

8.4. Data Availability Statement

These data are available from the corresponding author upon reasonable request.

8.5. Acknowledgment

Not applicable.

8.6. Conflicts of Interest

The author declares no conflicts of interest.

9. References

[1] M. T. Molla, "Association of risk management practices and financial performance of microfinance institutions in Ethiopia : A two-step system generalized method of moments approach," *PLOS One*, pp. 1–18, 2025. https://doi.org/10.1371/journal.pone.0321415.

[2] R. Sifrain, "Factors Influencing Loan Portfolio Quality of Microfinance Institutions in Haiti," *Journal of Financial Risk Management*, vol. 11, no. 1, pp. 95–115, 2022. https://doi.org/10.4236/jfrm.2022.111005.

[3] H. A. Legass and H. A. Roba, "The Impact of Credit Risk Management on Financial Performance of Commercial Banks in Ethiopia," *Journal of Finance and Accounting*, vol. 12, no. 6, pp. 142–155, 2024. https://doi.org/10.11648/j.jfa.20241206.11.

[4] V. García, A. I. Marqués, and J. S. Sánchez, "Exploring the synergetic effects of sample types on the performance of ensembles for credit risk and corporate bankruptcy prediction," *Inf. Fusion*, vol. 47, pp. 88–101, 2019. https://doi.org/10.1016/j.inffus.2018.07.004.

[5] F. Barboza, H. Kimura, and E. Altman, "Machine learning models and bankruptcy prediction," *Expert Syst. Appl.*, vol. 83, pp. 405–417, 2017. https://doi.org/10.1016/j.eswa.2017.04.006.

[6] S. Shi, R. Tse, W. Luo, S. D'Addona, and G. Pau, "Machine learning-driven credit risk: a systemic review," *Neural Comput. Appl.*, vol. 34, no. 17, pp. 14327–14339, 2022. https://doi.org/10.1007/s00521-022-07472-2.

[7] S. Zhang, J. Tay, and P. Baiz, "The Effects of Data Imbalance Under a Federated Learning Approach for Credit Risk Forecasting," *arXiv preprint arXiv:2401.07234*, 2024. https://doi.org/10.48550/arXiv.2401.07234.

[8] A. Namvar, M. Siami, F. Rabhi, and M. Naderpour, "Credit risk prediction in an imbalanced social lending environment," *International Journal of Computational Intelligence Systems*, vol. 11, pp. 925–935, 2018. https://doi.org/10.2991/ijcis.11.1.70.

[9] E. B. Abdelghafour, C. Mohamed, A. Noura, and B. Abdelhamid, "Enhancing Credit Card Fraud Detection Using a Stacking Model Approach and Hyperparameter Optimization," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 15, no. 10, 2024. https://doi.org/10.14569/ijacsa.2024.01510110.

[10] H. Wang, "Credit Risk Management of Consumer Finance Based on Big Data," *Mobile Information Systems*, vol. 2021, 2021. https://doi.org/10.1155/2021/8189255.

[11] L. Gao and J. Xiao, "Big Data Credit Report in Credit Risk Management of Consumer Finance," *Wireless Communications and Mobile Computing*, vol. 2021, 2021. https://doi.org/10.1155/2021/4811086.

[12] A. Niu, B. Cai, and S. Cai, "Big Data Analytics for Complex Credit Risk Assessment of Network Lending Based on SMOTE Algorithm," *Complexity*, vol. 2020, 2020. https://doi.org/10.1155/2020/8563030.

[13] J. P. Noriega, L. A. Rivera, and J. A. Herrera, "Machine Learning for Credit Risk Prediction: A Systematic Literature Review," *Data*, vol. 8, no. 11, 2023. https://doi.org/10.3390/data8110169.

[14] N. Chai, M. Zoynul, L. Yang, and B. Shi, "Farmers' credit risk evaluation with an explainable hybrid ensemble approach : A closer look in microfinance," *Pacific-Basin Financ. J.*, vol. 89, no. June 2024, p. 102612, 2025. https://doi.org/10.1016/j.pacfin.2024.102612.

[15] L. Marceau, L. Qiu, N. Vandewiele, E. Charton, "A comparison of Deep Learning performances with other machine learning algorithms on credit scoring unbalanced data," *arXiv preprint arXiv:1907.12363*, 2019. https://doi.org/10.48550/arXiv.1907.12363.

[16] A. Ampountolas, T. N. Nde, P. Date, and C. Constantinescu, "A Machine Learning Approach for Micro-Credit Scoring," *Risks*, vol. 9, no. 3, 2021. https://doi.org/10.3390/risks9030050.

[17] M. Z. Abedin, C. Guotai, P. Hajek, and T. Zhang, "Combining weighted SMOTE with ensemble learning for the class-imbalanced prediction of small business credit risk," *Complex Intell. Syst.*, vol. 9, no. 4, pp. 3559–3579, 2023. https://doi.org/10.1007/s40747-021-00614-4.

[18] I. Aruleba and Y. Sun, "Effective Credit Risk Prediction Using Ensemble Classifiers With Model Explanation," *IEEE Access*, pp. 115015 – 115025, vol. 12, 2024. https://doi.org/10.1109/ACCESS.2024.3445308.

[19] X. Liu, Z. Zhang, and D. Wang, "Classification of Imbalanced Credit scoring data sets Based on Ensemble Method with the Weighted-Hybrid-Sampling," *arXiv preprint arXiv:2102.04721*, 2021. https://doi.org/10.48550/arXiv.2102.04721.

[20] E. Alfredo, S. Alvarez, M. Angel, and C. Lengua, "Design of a machine learning model for predicting credit risk in microfinance using environmental data," *Bulletin of Electrical Engineering and Informatics (BEEI)*, vol. 14, no. 6, pp. 4578–4589, 2025. https://doi.org/10.11591/eei.v14i6.9848.

[21] C. Yu, Y. Jin, Q. Xing, Y. Zhang, S. Guo, S. Meng, "Advanced User Credit Risk Prediction Model Using LightGBM, XGBoost and Tabnet with SMOTEENN," in *2024 IEEE 6th International Conference on Power, Intelligent Computing and Systems (ICPICS)*, 2024. https://doi.org/10.1109/ICPICS62053.2024.10796247.

[22] N. A. Siagian, S. P. Sipayung, A. Rikki, and N. Marbun, "Integrating SMOTE with XGBoost for Robust Classification on Imbalanced Datasets: A Dual-Domain Evaluation," *Sinkron: Jurnal*

dan Penelitian Teknik Informatika, vol. 9, no. 3, pp. 1094–1107, 2025. https://doi.org/10.33395/sinkron.v9i3.15029.

[23] T. Sodnomdavaa and M. Sandagsuren, "Temporal and Cost-Sensitive Evaluation Framework for Credit Risk Modeling Under Distributional Shifts," *Risks*, vol. 14, no. 4, art. No. 95, 2026. https://doi.org/10.3390/risks14040095.

[24] Z. Zhao, T. Cui, S. Ding, J. Li, and A. G. Bellotti, "Resampling Techniques Study on Class Imbalance Problem in Credit Risk Prediction," *Mathematics*, vol. 12, no. 5, art. No. 701, 2024. https://doi.org/10.3390/math12050701.

[25] A. B. Shaik and S. Srinivasan, "A Brief Survey on Random Forest Ensembles in Classification Model," in *International Conference on Innovative Computing and Communications*, pp. 253–260, 2018. https://doi.org/10.1007/978-981-13-2354-6_27.

[26] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016. https://doi.org/10.1145/2939672.2939785.

[27] A. H. Ginting, R. W. Sembiring, and E. M. Zamzami, "Comparison Analysis: Logistic Regression, Random Forest, XGBoost, and CatBoost in Credit Scoring," in *Proceedings of the 2024 Brawijaya International Conference (BIC 2024)*, 2025. https://doi.org/10.2991/978-94-6463-854-7_16.

[28] H. A. Salman, A. Kalakech, and A. Steiti, "Random Forest Algorithm Overview," *Babylonian Journal of Machine Learning*, vol. 2024, pp. 69–79, 2024. https://doi.org/10.58496/BJML/2024/007.

[29] A. O. Kuyoro, O. A. Ogunyolu, T. G. Ayanwola, and F. Yetunde, "Dynamic Effectiveness of Random Forest Algorithm in Financial Credit Risk Management for Improving Output Accuracy and Loan Classification Prediction," *Ingénierie des Systèmes d'Information*, vol. 27, no. 5, pp. 815–821, 2022. https://doi.org/10.18280/isi.270515.

[30] C.-H. Chen, J.-P. Lai, Y.-M. Chang, C.-J. Lai, and P.-F. Pai, "A Study of Optimization in Deep Neural Networks for Regression," *Electronics*, vol. 12, no. 14, art. No. 3071, 2023. https://doi.org/10.3390/electronics12143071.

[31] M. Schmitt, "Deep Learning vs. Gradient Boosting: Benchmarking state-of-the-art machine learning algorithms for credit scoring," *arXiv preprint arXiv:2205.10535*, 2022. https://doi.org/10.48550/arXiv.2205.10535.

[32] O. Radović, S. Marinković, and J. Radojičić, "Credit Scoring with an Ensemble Deep Learning Classification Methods – Comparison with Traditional Methods," *Facta Universitatis, Series: Economics and Organization*, vol. 18, no. 1, pp. 29–43, 2021. https://doi.org/10.22190/FUEO201028001R.

[33] Y. Chen, R. Calabrese, and B. Martin-barragan, "Interpretable machine learning for imbalanced credit scoring datasets," *Eur. J. Oper. Res.*, vol. 312, no. 1, pp. 357–372, 2024. https://doi.org/10.1016/j.ejor.2023.06.036.

[34] L. Wang, "Imbalanced credit risk prediction based on SMOTE and multi-kernel FCM improved by particle swarm optimization," *Appl. Soft Comput.*, vol. 114, p. 108153, 2022. https://doi.org/10.1016/j.asoc.2021.108153.

[35] M. Song, H. Ma, Y. Zhu, M. Zhang, "Credit Risk Prediction Based on Improved ADASYN Sampling and Optimized LightGBM," *Journal of Social Computing*, vol. 5, no. 3, pp. 232-241, 2024. https://doi.org/10.23919/JSC.2024.0019.

[36] L. H. Chia, "Finding the Sweet Spot: Optimal Data Augmentation Ratio for Imbalanced Credit Scoring Using ADASYN," *arXiv preprint arXiv:2510.18252*, 2025. https://doi.org/10.48550/arXiv.2510.18252.

[37] J. Guo, H. Wu, X. Chen, and W. Lin, "Adaptive SV-Borderline SMOTE-SVM algorithm for imbalanced data classification," *Appl. Soft Comput.*, vol. 150, p. 110986, 2024. https://doi.org/10.1016/j.asoc.2023.110986.

[38] A. S. Almajid and R. Arifudin, "Multilayer Perceptron Optimization on Imbalanced Data Using SVM-SMOTE and One-Hot Encoding for Credit Card Default Prediction," *Journal of Advances in Information Systems and Technology*, vol. 3, no. 2, pp. 67–74, 2021. https://doi.org/10.15294/jaist.v3i2.57061.

[39] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Prog. Artif. Intell.*, vol. 5, no. 4, pp. 221–232, 2016. https://doi.org/10.1007/s13748-016-0094-0.

[40] G. Hou, D. L. Tong, S. Y. Liew, and P. Y. Choo "Comparative Analysis of Resampling Techniques for Class Imbalance in Financial Distress Prediction Using XGBoost," *Mathematics*, vol. 13, no. 13, art. No. 2186, 2025. https://doi.org/10.3390/math13132186.

[41] O.-A. Ampomah, E. Agyemang, K. Acheampong, and L. Agyekum, "Enhancing Credit Default Prediction Using Boruta Feature Selection and DBSCAN Algorithm with Different

Resampling Techniques," *arXiv preprint arXiv:2509.19408*, 2025. https://doi.org/10.48550/arXiv.2509.19408.

[42] C. Vairetti, J. L. Assadi, and S. Maldonado, "Efficient hybrid oversampling and intelligent undersampling for imbalanced big data classification," *Expert Systems with Applications*, vol. 246, 2024. https://doi.org/10.1016/j.eswa.2024.123149.

[43] I. Aruleba and Y. Sun, "Machine Learning with Applications Enhanced credit risk prediction using deep learning and SMOTE-ENN resampling," *Mach. Learn. with Appl.*, vol. 21, no. July 2024, p. 100692, 2025. https://doi.org/10.1016/j.mlwa.2025.100692.

[44] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 20 – 29, 2004. https://doi.org/10.1145/1007730.1007735.

[45] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research (JAIR)*, vol. 16, pp. 321–357, 2002. https://doi.org/10.1613/jair.953.

[46] Z. Zhao and V. Aumeboonsuke, "Imbalanced Credit Risk Prediction in Ensemble Learning Classifiers: A Comparative Analysis of SMOTE, ADASYN, SMOTETomek, and Cluster Centroids," *Arts of Management Journal*, vol. 7, no. 3, pp. 959–984, 2023. https://so02.tci-thaijo.org/index.php/jam/article/view/264093.