

An Explainable Machine Learning Framework for Loan Default Prediction in Fintech using Ensemble Learning Models

Chioma Julia Onwuzo
Department of Computer Science,
Micheal Okpara University of Agriculture,
Umudike, Abia State

Godfrey Perfectson Oise
Department of Computer Science,
University of Benin, Benin City, Edo State
<https://orcid.org/0009-0006-4393-7874>

Ifeoma Vivian Ngonadi
Department of Computer Science and Information Technology,
Petroleum Training Institute, Effurun Delta State,
<https://orcid.org/0000-0002-9220-8161>

Immunhierokene Clinton Obrorindo
Department of Computer Science and Information Technology,
Petroleum Training Institute, Effurun Delta State

Augustine Emakpor Otorokpo
Department of Computer Science and Information Technology,
Petroleum Training Institute, Effurun Delta State

ABSTRACT

The rapid expansion of fintech has intensified the need for accurate and transparent models for loan default prediction, which is critical for effective credit risk management. This study proposes an explainable machine learning framework that combines ensemble learning models with explainable artificial intelligence (XAI) techniques to enhance both predictive performance and interpretability. Using the LendingClub dataset, two ensemble models Random Forest and Extreme Gradient Boosting (XGBoost) were developed and evaluated. Model performance was assessed using multiple metrics, including accuracy, precision, recall, F1-score, and ROC-AUC, with particular emphasis on detecting the minority class (loan defaults). The results indicate that both models achieved relatively high overall accuracy of approximately 79%. However, they performed poorly in identifying default cases, as reflected by low recall and F1-scores, highlighting the limitations of accuracy in imbalanced datasets. XGBoost slightly outperformed Random Forest and was selected as the final model, although the performance difference was minimal. To address the interpretability challenge of complex models, SHapley Additive exPlanations (SHAP) was applied to provide both global and local insights into model predictions. The analysis revealed that interest rate and loan amount are the most influential predictors of default, while higher income and credit scores reduce default risk. These findings align with established financial risk principles and enhance transparency in decision-making. The study demonstrates that integrating ensemble learning with explainable AI is promising, but improving minority class detection remains essential for real-world fintech applications.

Keywords

Loan Default Prediction, Ensemble Learning, Explainable Artificial Intelligence (XAI), SHAP (Shapley Additive Explanations), Credit Risk Assessment

1. INTRODUCTION

The transformation of the global financial ecosystem through digital innovation has significantly reshaped the mechanisms of credit risk assessment. The emergence of financial technology (fintech) platforms has introduced highly scalable, data-driven

lending systems that operate in real time, leveraging vast and heterogeneous datasets [1]. In this evolving landscape, the accurate prediction of loan default has become a central challenge, directly influencing institutional stability, portfolio optimization, and regulatory compliance. As lending decisions increasingly rely on automated systems, the need for robust, reliable, and interpretable predictive models has grown substantially [2]. Traditional approaches to credit risk modeling, such as logistic regression and linear discriminant analysis, have long been favored due to their simplicity, statistical rigor, and interpretability. These models provide probabilistic outputs that are easily understood by financial practitioners and regulators [3]. However, their reliance on linear assumptions and limited capacity to capture complex feature interactions significantly constrains their effectiveness in modern fintech environments. Contemporary financial datasets are characterized by high dimensionality, nonlinear dependencies, and intricate relationships among borrower behavior, macroeconomic variables, and transactional patterns. Consequently, traditional models often fail to fully exploit the informational richness embedded within such data [4].

To address these limitations, machine learning methodologies have gained widespread adoption in credit risk assessment. Algorithms such as decision trees, support vector machines, and k-nearest neighbors have demonstrated improved predictive capabilities by modeling nonlinear relationships and accommodating diverse data structures [5]. More recently, ensemble learning techniques particularly Random Forest and Gradient Boosting frameworks have emerged as state-of-the-art approaches for classification tasks. These models enhance predictive performance by aggregating multiple learners, thereby reducing variance, mitigating overfitting, and improving generalization across unseen data. In the context of loan default prediction, ensemble models have consistently outperformed traditional statistical methods, especially when applied to large-scale fintech datasets [6]. Despite these advancements, a fundamental challenge persists: the trade-off between predictive accuracy and interpretability. Ensemble models, while powerful, are often regarded as opaque “black-box” systems, providing limited insight into the underlying decision-making process. This lack of transparency poses significant concerns in financial

domains, where accountability, fairness, and regulatory compliance are paramount [7]. Financial institutions must be able to justify automated decisions, particularly in high-stakes scenarios such as loan approvals or rejections. Moreover, regulatory frameworks increasingly mandate explainability to ensure that algorithmic decisions are free from bias and aligned with ethical standards.

In response to these challenges, the field of Explainable Artificial Intelligence (XAI) has emerged as a critical research frontier [8]. XAI seeks to bridge the gap between model complexity and interpretability by providing tools that elucidate how models arrive at their predictions. Among the most prominent techniques are SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME). SHAP, grounded in cooperative game theory, offers a theoretically sound approach for attributing feature importance by quantifying each variable's contribution to model output [9]. It enables both global interpretability identifying key predictors across the dataset and local interpretability explaining individual predictions. LIME complements this by approximating complex models locally using interpretable surrogate models. Together, these methods provide a comprehensive framework for understanding and validating machine learning models in high-stakes applications [10].

Notwithstanding the growing body of research in this domain, significant gaps remain. Many existing studies prioritize predictive performance while neglecting interpretability, thereby limiting practical deployment in regulated financial environments [11]. Conversely, approaches that emphasize interpretability often rely on simpler models that sacrifice predictive accuracy [12]. Furthermore, there is a lack of integrated frameworks that systematically combine high-performing ensemble models with advanced XAI techniques while addressing real-world challenges such as class imbalance and feature redundancy [13]. This study seeks to address these limitations by proposing an integrated, explainable machine learning framework for loan default prediction in fintech. The framework leverages ensemble learning models specifically Random Forest and Extreme Gradient Boosting (XGBoost) to capture complex nonlinear relationships in financial data. Simultaneously, it incorporates SHAP-based explainability to provide transparent and interpretable insights into model behavior. The use of a real-world LendingClub dataset ensures practical relevance, while a comprehensive evaluation framework assesses model performance across multiple metrics, with particular emphasis on the accurate detection of default cases.

This research develops an explainable predictive framework for loan default classification using ensemble machine learning techniques. It involves data preprocessing, exploratory data analysis, and model training using Random Forest and XGBoost. The models are evaluated using performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC, with a focus on handling class imbalance. XGBoost is selected as the best-performing model. To enhance interpretability, SHAP is applied to identify key features influencing predictions. The study demonstrates that while ensemble models achieve high overall accuracy, their effectiveness in detecting default cases is limited, and integrating explainable AI significantly improves transparency and supports informed decision-making in fintech systems.

2. RELATED WORK

The domain of loan default prediction and credit risk assessment has undergone substantial methodological evolution,

transitioning from classical statistical paradigms to advanced machine learning and, more recently, explainable artificial intelligence frameworks. This progression reflects the increasing complexity of financial data and the growing demand for both predictive accuracy and model transparency. The existing literature can be broadly categorized into three interconnected streams: traditional statistical models, machine learning and ensemble approaches, and explainable AI techniques in financial systems [14]. Early research in credit risk modeling was predominantly grounded in statistical methods, with logistic regression emerging as the most widely adopted technique. Its popularity stems from its probabilistic interpretation, computational efficiency, and inherent interpretability, which align well with regulatory requirements in financial institutions [15]. Variants such as probit models and linear discriminant analysis were also employed to classify borrowers based on creditworthiness. However, these models rely on strong assumptions, including linear relationships between predictors and the log-odds of default, as well as independence among explanatory variables. Such assumptions limit their ability to capture the complex, nonlinear, and high-dimensional interactions that characterize modern financial datasets. Although feature engineering and domain-specific variables have been introduced to enhance their performance, the fundamental limitations of these models remain evident in large-scale fintech applications.

The advent of machine learning marked a significant shift in credit risk modeling, enabling the exploitation of complex patterns within data. Algorithms such as decision trees, support vector machines, and k-nearest neighbors have demonstrated improved classification performance by modeling nonlinear relationships and accommodating heterogeneous data types. Decision trees, in particular, gained traction due to their intuitive structure and interpretability [16]. Nevertheless, their susceptibility to overfitting and instability has constrained their standalone effectiveness. To address these limitations, ensemble learning methods have been introduced, representing a major advancement in predictive modeling. Techniques such as Random Forest and Gradient Boosting Machines combine multiple base learners to enhance predictive accuracy and robustness. Random Forest employs a bagging strategy, constructing numerous decision trees on bootstrapped samples and aggregating their outputs to reduce variance. In contrast, gradient boosting methods, including Extreme Gradient Boosting (XGBoost) and LightGBM, adopt a sequential learning approach that iteratively minimizes prediction errors [17]. Empirical studies consistently demonstrate that these ensemble models outperform both traditional statistical methods and single machine learning algorithms, particularly in large-scale datasets such as LendingClub and Home Credit. Their ability to capture intricate feature interactions and nonlinear dependencies makes them highly suitable for credit risk prediction in fintech contexts. However, this performance gain is accompanied by increased model complexity, leading to reduced interpretability.

[18] Despite advances in machine learning for credit default prediction, most lenders still rely on traditional methods due to concerns about transparency and explainability. To address this, the study develops tree-based ensemble models, with XGBoost performing best, and applies SHAP to improve interpretability. Using Lending Club data, the SHAP analysis explains feature impacts on predictions, enhancing model transparency and making it suitable for real-world financial applications. [19] This study applies explainable machine learning survival analysis to credit risk assessment using a large Bondora P2P lending dataset. It compares classical Cox models with an XGBoost-based Cox model and finds that the machine learning

approach performs better in risk prediction, return estimation, and default classification. Using SHAP, the study further explains model decisions, uncovering both expected and novel risk factors and interactions. Overall, the approach improves accuracy, interpretability, and financial decision-making in credit risk modeling. [20] This study presents an AI-based framework for credit risk prediction using ensemble machine learning models on the Lending Club dataset. Gradient Boosting performs best, achieving over 92% accuracy, while SHAP and LIME are used to explain model predictions and improve transparency. Key risk factors include interest rate, installment amount, and debt-to-income ratio. Overall, the framework combines high predictive performance with interpretability, making it useful for trustworthy financial decision-making, though limitations in dataset scope and fairness analysis are acknowledged. [21] This study addresses the growing need for interpretable credit fraud risk models by proposing a machine learning framework enhanced with SHAP for explainability. Using Lending Club data, the model is evaluated against SVM, MLP, XGBoost, and LightGBM, and the results highlight key feature contributions and interactions affecting credit risk predictions. The analysis shows that factors such as loan interest rate and credit policies play major roles in fraud assessment. Overall, the SHAP-based approach improves model transparency and supports better-informed financial decision-making and risk management.

Parallel to these developments, deep learning approaches have also been explored for credit risk assessment. Neural networks, including multilayer perceptrons and recurrent architectures, offer powerful representation learning capabilities, particularly for high-dimensional and sequential data. Hybrid models that integrate deep learning with traditional or ensemble techniques have been proposed to further enhance performance [22]. While these approaches show promise, their opaque nature exacerbates the “black-box” problem, making them less suitable for regulatory-sensitive applications without additional interpretability mechanisms. The growing reliance on complex models has intensified the need for transparency, giving rise to the field of Explainable Artificial Intelligence (XAI). XAI methods aim to make machine learning models more interpretable without significantly compromising their predictive power [23]. Among the most prominent techniques are SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME). SHAP, based on cooperative game theory, provides a unified framework for attributing feature importance by computing the marginal contribution of each feature to model predictions. It offers both global insights into feature relevance and local explanations for individual predictions, making it particularly suitable for high-stakes financial applications [24]. LIME, on the other hand, approximates complex models locally using simpler, interpretable surrogate models, enabling instance-level explanations. Recent studies have applied SHAP and LIME to credit risk models to enhance interpretability and uncover key drivers of default behavior. These approaches have been instrumental in identifying critical variables such as income level, credit score, debt-to-income ratio, and loan characteristics. Additionally, explainability techniques have been used to detect biases and ensure fairness in automated decision systems, addressing ethical and regulatory concerns [25].

Despite these advancements, several research gaps persist. Many studies prioritize predictive accuracy without adequately addressing interpretability, limiting their practical applicability in regulated financial environments. Conversely, studies focusing on interpretability often rely on simpler models that sacrifice predictive performance [26]. Furthermore, there is a

lack of integrated frameworks that simultaneously leverage high-performing ensemble models and advanced XAI techniques while addressing real-world challenges such as class imbalance and feature redundancy [27]. Existing research also tends to emphasize either global or local interpretability, rather than providing a comprehensive, unified perspective. The literature reveals a clear progression from traditional statistical models to advanced machine learning and ensemble techniques, with recent emphasis on explainable AI. While ensemble models such as Random Forest and XGBoost achieve superior predictive performance, they lack transparency. XAI methods like SHAP and LIME address this limitation by improving interpretability. However, existing studies often fail to integrate high predictive accuracy with comprehensive explainability. This research addresses this gap by combining ensemble learning with SHAP-based interpretation in a unified framework for loan default prediction.

3. METHODOLOGY

This study adopts a structured machine learning framework to develop an explainable system for loan default prediction in fintech. The methodology consists of data collection, preprocessing, exploratory data analysis, model development using ensemble techniques, performance evaluation, and model interpretability using explainable artificial intelligence methods.

3.1 Data Collection

The dataset used in this study is obtained from the LendingClub platform, which contains detailed records of peer-to-peer lending transactions. It includes accepted loan applications with borrower demographic information, financial attributes, credit history, and loan-specific features such as loan amount, interest rate, and repayment status. The target variable is binary, representing: 0: Fully paid loans, and 1: Defaulted or charged-off loans

3.2 Data Preprocessing

Data preprocessing is performed to ensure data quality and suitability for modeling. The following steps are applied:

1. **Handling Missing Values:** Features with a high proportion of missing values are removed, while remaining missing values are imputed using statistical methods such as mean or median.
2. **Encoding Categorical Variables:** Categorical features are transformed into numerical format using techniques such as one-hot encoding or label encoding.
3. **Feature Scaling:** Numerical features are normalized or standardized to improve model convergence and performance.
4. **Handling Class Imbalance:** Since default cases are typically fewer than non-default cases, imbalance handling techniques such as class weighting are applied to improve model sensitivity to the minority class.
5. **Feature Selection:** Highly correlated and redundant features are identified through correlation analysis and removed to reduce multicollinearity.

3.3 Exploratory Data Analysis (EDA)

Exploratory data analysis is conducted to understand the structure and relationships within the dataset:

1. **Correlation Analysis:** A correlation heatmap is used to identify relationships among numerical features and detect redundancy.
2. **Feature Distribution Analysis:** The distribution of

key variables such as loan amount and interest rate are examined to understand borrowing patterns.

3. **Class-wise Comparison:** Feature distributions are analyzed across the target classes to identify variables that significantly influence loan default.

3.4 Model Development

This study employs ensemble learning techniques for loan default prediction. The models implemented include:

1. **Random Forest:** A bagging-based ensemble method that constructs multiple decision trees using bootstrapped samples and aggregates their predictions to improve generalization and reduce variance.
2. **Extreme Gradient Boosting (XGBoost):** A boosting-based ensemble algorithm that builds models sequentially by minimizing prediction errors and optimizing performance through gradient descent.

The dataset is split into training and testing sets (e.g., 80:20 ratio). Both models are trained on the training set and evaluated on the test set.

3.5 Model Selection

After evaluating both models, **XGBoost is selected as the final (best_model)** due to its slightly superior performance in terms of accuracy and overall classification effectiveness compared to Random Forest.

3.6 Model Evaluation

Model performance is assessed using multiple evaluation metrics to ensure comprehensive analysis:

1. **Accuracy:** Measures overall prediction correctness.
2. **Precision, Recall, and F1-score:** Evaluate class-wise performance, with emphasis on the minority (default) class.
3. **Confusion Matrix:** Provides detailed insight into classification outcomes.
4. **Receiver Operating Characteristic (ROC) Curve and AUC:** Measure the model's ability to distinguish between classes.

Special emphasis is placed on recall and F1-score for the default class, as correctly identifying high-risk borrowers is critical in financial applications.

3.7 Explainable AI Integration

To enhance model transparency and interpretability, **SHAP (Shapley Additive Explanations)** is applied to the selected XGBoost model.

SHAP provides:

1. **Global Interpretability:** Identifies the most important features influencing model predictions across the dataset.
2. **Local Interpretability:** Explains individual predictions by quantifying the contribution of each feature.

This enables a clear understanding of how variables such as interest rate, loan amount, income level, and credit score influence loan default predictions, thereby supporting informed decision-making and regulatory compliance.

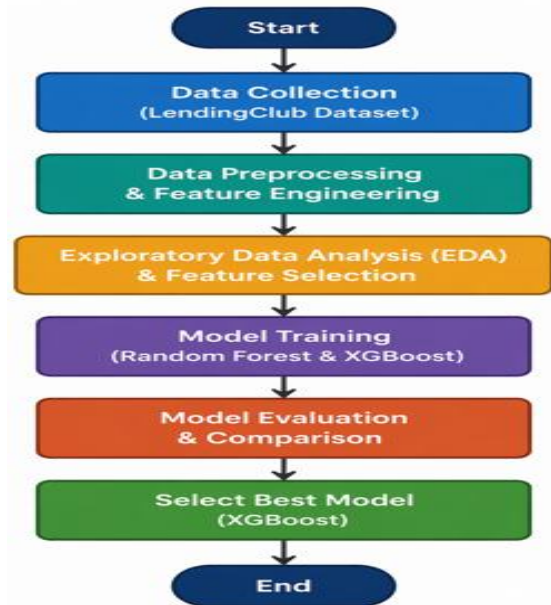


Figure 1: Proposed Machine Learning Framework for Credit Default Prediction

Figure 1 illustrates the overall workflow of the proposed machine learning framework for credit default prediction. The process begins with data collection from the LendingClub dataset, followed by data preprocessing and feature engineering to clean and transform the data into a suitable format. Next, exploratory data analysis (EDA) and feature selection are conducted to identify relevant variables and understand underlying patterns. The refined dataset is then used to train two models, Random Forest and XGBoost, whose performances are evaluated and compared using appropriate metrics. Based on this comparison, XGBoost is selected as the best-performing model, marking the final stage of the framework.

Table 1: Correlation Analysis of Input Features for Loan Default Prediction

	loan_a mnt	int_rat e	install ment	annual _inc	dti	delinq _2yrs	fico_range _low	fico_range _high	open_ acc	pub_r ec	revol_ bal
loan_amnt	1.0000 00	0.142 520	0.9530 48	0.3096 54	0.031 772	- 0.00246 9	0.101131	0.101131	0.185 576	- 0.064 804	0.316 854
int_rate	0.1425 20	1.000 000	0.1555 51	- 0.0715 90	0.146 133	0.04794 8	-0.403398	-0.403395	- 0.004 346	0.052 403	- 0.028 487
installment	0.9530 48	0.155 551	1.0000 00	0.3005 01	0.035 558	0.00685 3	0.050978	0.050977	0.175 923	- 0.053 348	0.305 357

annual_inc	0.309654	-0.071590	0.300501	1.000000	-0.139095	0.040792	0.069828	0.069829	0.134390	-0.004875	0.284769
dti	0.031772	0.146133	0.035558	-0.139095	1.000000	-0.008004	-0.060445	-0.060447	0.221946	-0.030772	0.108287
delinq_2yrs	-0.002469	0.047948	0.006853	0.040792	-0.008004	1.000000	-0.174211	-0.174209	0.052532	-0.020446	-0.027810
fico_range_low	0.101131	-0.403398	0.050978	0.069828	-0.060445	-0.174211	1.000000	1.000000	0.016923	-0.191373	0.025355
fico_range_high	0.101131	-0.403395	0.050977	0.069829	-0.060447	-0.174209	1.000000	1.000000	0.016921	-0.191371	0.025353
open_acc	0.185576	-0.004346	0.175923	0.134390	0.221946	0.052532	0.016923	0.016921	1.000000	-0.017958	0.224813
pub_rec	-0.064804	0.052403	-0.053348	-0.004875	-0.030772	-0.020446	-0.191373	-0.191371	-0.017958	1.000000	-0.092746
revol_bal	0.316854	-0.028487	0.305357	0.284769	0.108287	-0.027810	0.025355	0.025353	0.224813	-0.092746	1.000000

Table 1 reveals several important relationships among the financial variables used for loan default prediction. A strong positive correlation exists between loan amount and installment, indicating redundancy since both represent similar financial obligations. Additionally, a notable negative correlation between interest rate and credit score (both low and high ranges) reflects expected credit risk behavior, where borrowers with higher credit scores receive lower interest rates. The perfect correlation between `fico_range_low` and `fico_range_high` suggests multicollinearity, meaning one of these features should be removed to improve model efficiency. Other variables such as annual income and revolving balance show moderate relationships, while features like debt-to-income ratio and public records exhibit weak correlations.

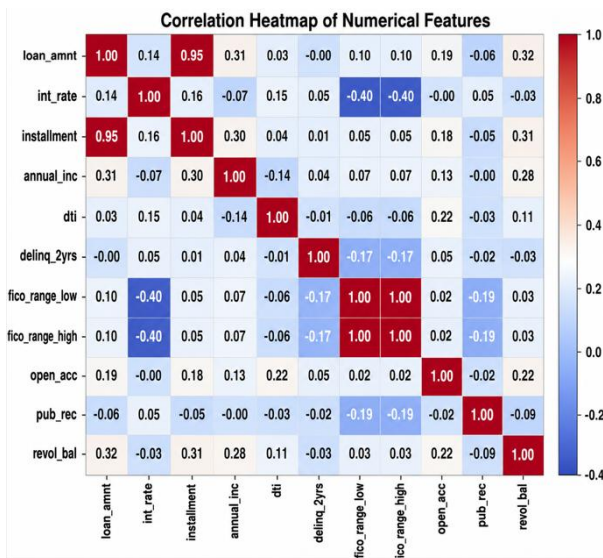


Figure 2: Correlation Heatmap of Numerical Features

Figure 2 shows the correlation heatmap of the dataset,

highlighting relationships among numerical features. It reveals a strong positive correlation between *loan_amnt* and *installment* (0.95), while *fico_range_low* and *fico_range_high* is perfectly correlated (1.00), indicating redundancy. *int_rate* has a moderate negative relationship with FICO scores (≈ -0.40), meaning better credit scores are linked to lower interest rates. Most other variables show weak correlations, suggesting limited linear relationships.

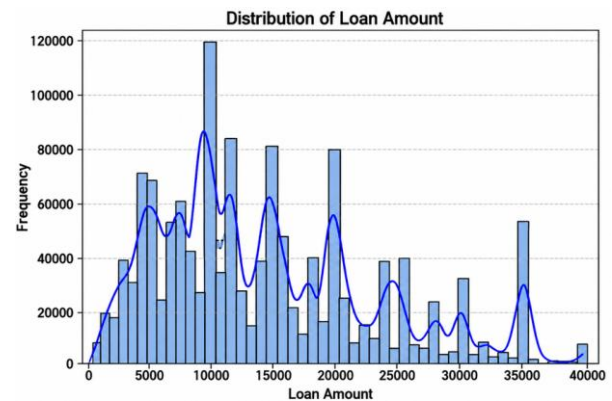


Figure 3: Distribution of Loan Amount Across the Dataset

Figure 3 presents the distribution of loan amounts, illustrating the frequency and density of loans across different value ranges. The plot reveals that most loans are concentrated within the lower to mid-range values, particularly between approximately 5,000 and 20,000, indicating a higher demand for moderate loan sizes. The distribution is right-skewed, with fewer instances of high-value loans extending toward 40,000. Additionally, the presence of multiple peaks suggests variability in borrowing patterns, possibly reflecting different borrower segments or loan categories. Overall, the figure highlights that smaller and medium-sized loans dominate the dataset, while larger loans are comparatively less frequent.

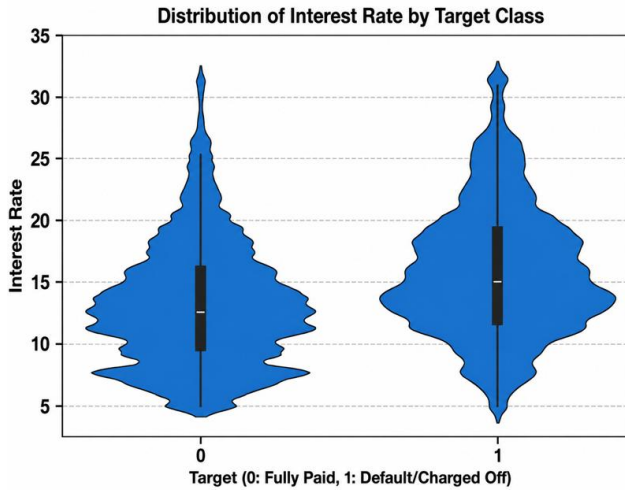


Figure 4: Distribution of Interest Rate by Target Class

Figure 4 suggests a positive relationship between higher interest rates and loan default risk, but with significant overlap between outcomes. This implies interest rate is an informative but not decisive predictor of loan performance.

4. RESULTS

The results of this study evaluate the performance of ensemble learning models for loan default prediction using the LendingClub dataset. Random Forest and XGBoost models were implemented and assessed using multiple classification metrics, including accuracy, precision, recall, F1-score, and ROC-AUC. While both models achieved similar overall accuracy, their performance varied significantly in detecting default cases due to class imbalance. Based on comparative evaluation, XGBoost was selected as the best-performing model. To enhance interpretability, SHAP was applied to the selected model, providing insights into the key factors influencing loan default predictions. To provide a more comprehensive assessment of model performance, both ensemble models were evaluated using class-specific precision, recall, F1-score, confusion matrices, and ROC-AUC analysis in addition to overall accuracy. Because the LendingClub dataset exhibits significant class imbalance, overall accuracy alone does not adequately reflect the effectiveness of the models in identifying loan defaults. Consequently, particular emphasis is placed on the minority (default) class, where recall and F1-score provide more meaningful indicators of model performance. The confusion matrices further illustrate the distribution of true positives, true negatives, false positives, and false negatives, while the ROC curves evaluate the discriminative capability of each classifier across varying decision thresholds. Finally, SHAP analysis is employed to explain the contribution of each predictor to the model's decisions, thereby linking predictive performance with model interpretability.

Table 2: Performance Evaluation of Random Forest Model

Class	Precision	Recall	F1-Score	Support
0	0.80	0.97	0.88	215,351
1	0.50	0.09	0.16	58,013
Accuracy	—	—	0.79	273,364
Macro Average	0.65	0.53	0.52	273,364
Weighted Average	0.74	0.79	0.73	273,364

Table 2, shows that the Random Forest model achieves an overall accuracy of 79%, indicating reasonable general performance. However, a deeper inspection reveals a significant class imbalance issue. The model performs very well on Class 0 (normal traffic), with high recall (0.97) and F1-score (0.88), meaning it correctly identifies most normal instances. In contrast, performance on Class 1 (attack traffic) is poor, with a very low recall of 0.09 and F1-score of 0.16, indicating that most attack instances are misclassified as normal. The AUC value of 0.534 further suggests that the model has limited discriminative ability. Overall, the model is biased toward the majority class and is not reliable for detecting minority class (attack) instances, which is critical in intrusion detection contexts.

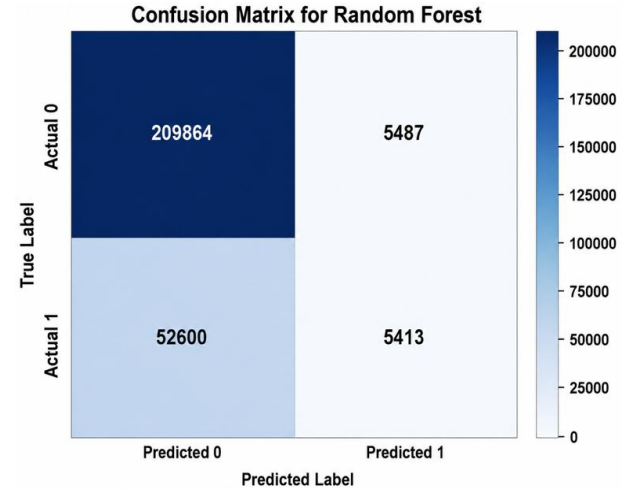


Table 5: Confusion Matrix Analysis of Random Forest Model

Table 5 highlights a strong bias of the Random Forest model toward predicting the majority class (normal traffic). A large number of normal instances are correctly classified (True Negatives = 209,864), with relatively few misclassified as attacks (False Positives = 5,487). However, the model performs poorly in detecting attack instances, as evidenced by the very high number of False Negatives (52,600) compared to True Positives (5,413). This indicates that most attack samples are incorrectly labeled as normal, which is a critical limitation for intrusion detection systems. The imbalance between False Negatives and True Positives aligns with the previously observed low recall for Class 1, confirming that the model lacks sensitivity to malicious activities despite its high overall accuracy.

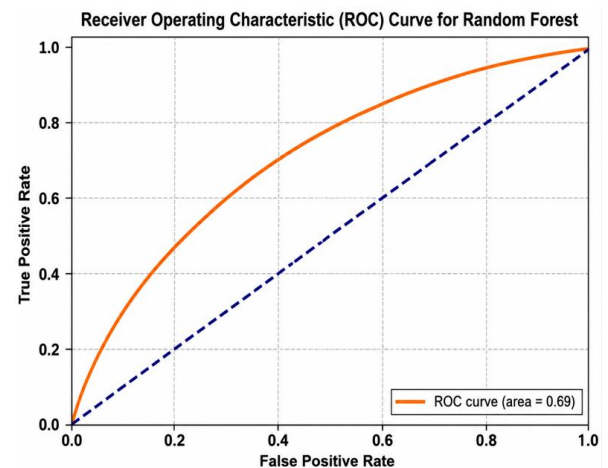


Figure 6: ROC Curve Analysis of Random Forest Model

Figure 6 illustrates the trade-off between the True Positive Rate (TPR) and False Positive Rate (FPR) for the Random Forest model across different classification thresholds. The curve lies moderately above the diagonal baseline, with an Area Under the Curve (AUC) of 0.69, indicating fair but not strong discriminative performance. While the model performs better than random guessing, the relatively low AUC suggests limited capability in effectively distinguishing between normal and attack classes. This aligns with earlier results showing poor detection of the minority (attack) class, confirming that the model struggles to achieve a balanced classification performance.

Table 3: Performance Evaluation of XGBoost Model

Class	Precision	Recall	F1-Score	Support
0	0.80	0.98	0.88	215,351
1	0.56	0.08	0.13	58,013
Accuracy	—	—	0.79	273,364
Macro Average	0.68	0.53	0.51	273,364
Weighted Average	0.75	0.79	0.72	273,364

Table 3 illustrates the performance of the XGBoost model, showing an overall accuracy of 0.79 but a low AUC of 0.53, indicating weak discriminative capability. The model performs strongly on Class 0 with high recall (0.98) and F1-score (0.88), correctly identifying most majority class instances. However, it performs poorly on Class 1, with a very low recall (0.08) and F1-score (0.13), failing to detect most minority class cases. The macro average F1-score of 0.51 further reflects this imbalance, while the weighted F1-score of 0.72 is influenced by the dominance of Class 0. Overall, the figure highlights that the model is biased toward the majority class and is ineffective for minority class prediction.

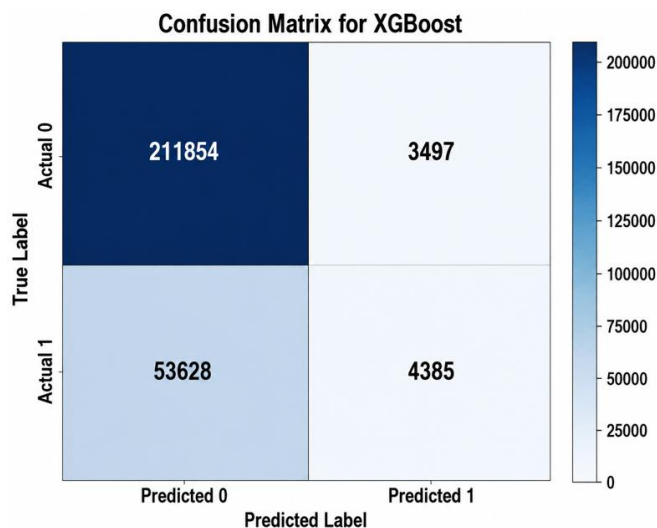


Figure 7 Confusion matrix of the XGBoost model

Figure 7 shows the confusion matrix of the XGBoost model, indicating that it predicts the majority class (Class 0) very well, with a high number of true negatives (211,854) and relatively few false positives (3,497). However, the model performs poorly on the minority class (Class 1), as evidenced by the large number of false negatives (53,628) compared to the small

number of true positives (4,385). This imbalance demonstrates that the model is heavily biased toward Class 0, leading to poor detection of Class 1 instances and confirming its limited effectiveness for minority class prediction.

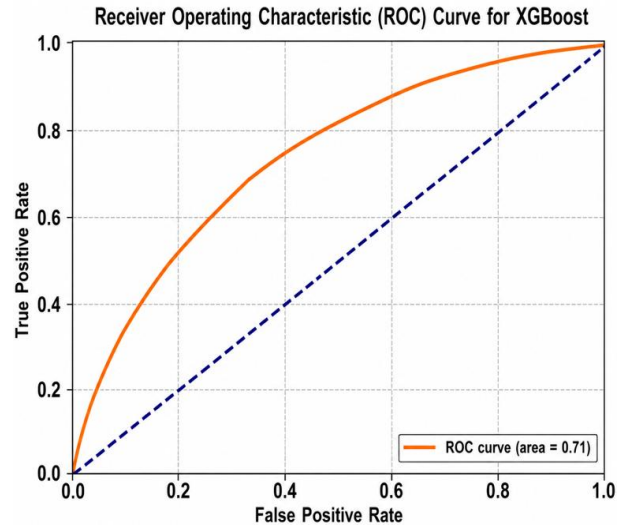


Figure 8: ROC curve of the XGBoost model

Figure 8 presents the ROC curve of the XGBoost model, illustrating its ability to distinguish between the two classes. The curve lies above the diagonal baseline, indicating better performance than random guessing, with an AUC of 0.71, which reflects moderate discriminative capability. This suggests that the model has some ability to separate positive and negative classes, although not strongly. Overall, the figure indicates that while the model performs reasonably in ranking predictions, its classification effectiveness especially for the minority class remains limited.

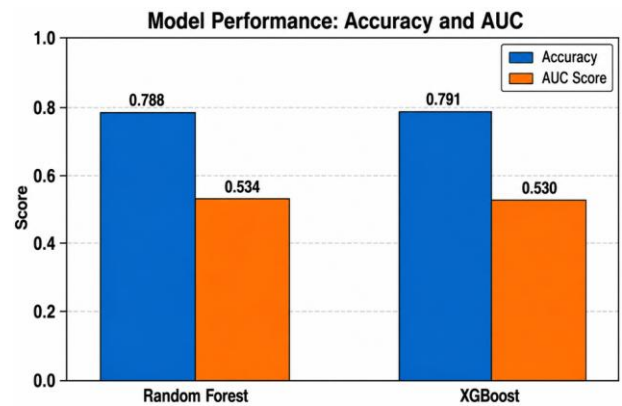


Figure 9: The comparison of model performance

Figure 9: The comparison of model performance shows that both Random Forest and XGBoost achieve similar accuracy, with XGBoost slightly higher (0.791) than Random Forest (0.788). However, both models record low AUC scores (0.534 and 0.530, respectively), indicating poor class discrimination. This suggests that despite relatively high accuracy, the models have limited ability to distinguish between classes, likely due to class imbalance. Although both Random Forest and XGBoost achieved similar overall accuracy (approximately 79%), the detailed evaluation reveals that both models struggle to detect the minority default class. XGBoost achieved slightly better classification performance than Random Forest, particularly in terms of ROC-AUC and overall ranking ability, leading to its selection as the final model. Nevertheless, the low recall values indicate that a considerable proportion of default

cases remain undetected, highlighting the need for future incorporation of imbalance-aware learning techniques.

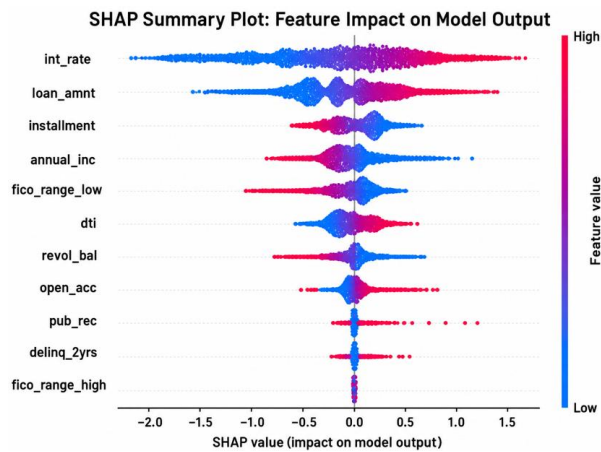


Figure 10: SHAP summary plot of the XGBoost model

Figure 10 illustrates the SHAP summary plot highlighting the relative importance and impact of each feature on prediction outcomes. The results show that *interest rate* and *loan amount* are the most influential variables, with higher values significantly increasing the likelihood of a positive prediction. Conversely, *annual income* and *FICO score* exhibit inverse effects, where higher values contribute to reduced risk. Additionally, *debt-to-income ratio* and *revolving balance* moderately drive predictions toward higher risk, while features such as *public records* and *delinquencies* have relatively lower but occasionally notable impacts.

5. DISCUSSION

The findings of this study provide important insights into the effectiveness of ensemble learning models for loan default prediction within fintech environments, particularly when combined with explainable artificial intelligence techniques. While both Random Forest and XGBoost demonstrated comparable overall accuracy, a deeper evaluation reveals critical limitations and implications for real-world financial decision-making [28]. A key observation from the results is the discrepancy between overall accuracy and the model's ability to detect the minority class (loan defaults). Both models achieved accuracy levels of approximately 79%, which might initially suggest strong predictive performance. However, this metric is misleading in the presence of class imbalance, as evidenced by the extremely low recall and F1-scores for the default class. The models consistently misclassified a large proportion of default cases as non-default, indicating a strong bias toward the majority class [29]. This finding aligns with existing literature, which highlights that accuracy alone is insufficient for evaluating classification models in imbalanced financial datasets.

The poor performance in detecting default cases has significant practical implications. In credit risk assessment, failing to identify high-risk borrowers (false negatives) can lead to substantial financial losses for lending institutions. The confusion matrix analysis further reinforces this concern, showing a disproportionately high number of false negatives compared to true positives in both models [30]. This suggests that, despite their computational sophistication, ensemble models such as Random Forest and XGBoost are not inherently immune to class imbalance issues and require additional strategies such as resampling techniques, cost-sensitive learning, or anomaly detection approaches to improve minority

class detection. Interestingly, although XGBoost was selected as the best-performing model, its advantage over Random Forest was marginal. While it achieved slightly higher accuracy and AUC, its recall for the minority class remained critically low [31]. This indicates that model selection based solely on marginal improvements in aggregate metrics may not translate into meaningful real-world performance gains. Instead, domain-specific priorities such as maximizing recall for default prediction should guide model optimization and evaluation.

Another important aspect of this study is the integration of explainable AI through SHAP analysis, which provides valuable interpretability into the model's decision-making process. The SHAP results reveal that variables such as interest rate and loan amount are the most influential predictors of loan default. This is consistent with financial theory, as higher interest rates typically reflect higher perceived risk, and larger loan amounts increase repayment burden [32]. Conversely, features such as annual income and FICO score were associated with reduced default risk, aligning with established credit risk assessment principles. The interpretability provided by SHAP enhances the transparency and trustworthiness of the model, addressing one of the major challenges associated with complex ensemble methods. By quantifying feature contributions at both global and local levels, the framework enables financial institutions to justify lending decisions, comply with regulatory requirements, and identify key risk drivers [33]. This is particularly important in fintech applications, where automated decision systems must remain explainable and auditable. However, the results also highlight that interpretability alone does not compensate for poor predictive performance on critical classes. While SHAP successfully explains model behavior, it does not inherently improve classification outcomes [34]. Therefore, explainability should be viewed as complementary to, rather than a substitute for, robust model performance. Furthermore, the exploratory data analysis supports the modeling results by identifying key relationships among variables. The strong correlation between loan amount and installment suggests redundancy, while the perfect correlation between FICO score ranges indicate multicollinearity. Addressing these issues during preprocessing likely contributed to improved model stability, although it did not fully resolve the classification imbalance problem. Additionally, the observed overlap in interest rate distributions between default and non-default classes explains why this feature, although important, is not sufficient as a standalone predictor.

This study underscores the complexity of loan default prediction in real-world fintech datasets. While ensemble learning models offer strong capabilities for capturing nonlinear relationships, their effectiveness is constrained by data imbalance and evaluation challenges. The integration of explainable AI adds significant value by improving transparency and interpretability, but further methodological enhancements are necessary to achieve reliable detection of high-risk borrowers. Future research should focus on improving minority class prediction through advanced techniques such as Synthetic Minority Oversampling Technique (SMOTE), ensemble rebalancing methods, or hybrid deep learning approaches. Additionally, incorporating temporal and behavioral data may further enhance predictive performance. A more comprehensive evaluation framework that prioritizes recall, precision-recall trade-offs, and cost-sensitive metrics would also provide a more realistic assessment of model effectiveness in financial applications. In conclusion, the proposed framework successfully demonstrates the potential of combining ensemble learning with explainable

AI for loan default prediction. However, achieving a balance between predictive accuracy, minority class sensitivity, and interpretability remains a critical challenge that must be addressed for practical deployment in fintech systems.

Although the proposed framework demonstrates promising performance on the LendingClub dataset, its evaluation is limited to a single benchmark dataset. Future research will extend the proposed explainable machine learning framework to additional publicly available credit risk datasets, including the German Credit, Home Credit Default Risk, and Give Me Some Credit datasets, to further assess its robustness and generalization capability across different lending environments. In addition, future studies will investigate varying class imbalance ratios, alternative preprocessing strategies, and cross-dataset validation to establish the scalability and practical applicability of the proposed framework.

6. CONCLUSION

This study presented an explainable machine learning framework for predicting loan default in fintech environments by integrating ensemble learning models with explainable artificial intelligence techniques. Two widely used ensemble models, Random Forest and XGBoost, were developed and evaluated using the LendingClub dataset to assess their effectiveness in credit risk prediction. The findings revealed that both models achieved relatively high overall accuracy, approximately 79%, indicating good general classification performance. However, a deeper analysis exposed a critical limitation: both models performed poorly in identifying loan defaults, which represent the minority class in the dataset. This was reflected in very low recall and F1-scores for default cases, demonstrating that many high-risk borrowers were incorrectly classified as low-risk. These results highlight the inadequacy of accuracy as a standalone evaluation metric in imbalanced datasets and emphasize the importance of using more relevant metrics such as recall and F1-score for the minority class. Although XGBoost slightly outperformed Random Forest and was selected as the final model, the difference in performance was minimal. Both models showed similar weaknesses in detecting default cases, suggesting that ensemble methods alone are not sufficient to address class imbalance challenges in credit risk modeling. A significant contribution of this study is the integration of SHAP for model interpretability. The SHAP analysis provided valuable insights into the key factors influencing loan default predictions, identifying interest rate and loan amount as major risk drivers, while higher income and credit scores reduced default likelihood. This improves transparency, supports regulatory compliance, and enhances trust in automated decision-making systems. While the proposed framework demonstrates the potential of combining ensemble learning with explainable AI, improving the detection of high-risk borrowers remains essential. Future research should focus on advanced imbalance-handling techniques and more robust evaluation strategies to enhance real-world applicability.

8. REFERENCES

- [1] F. Wahab, I. Khan, and S. Sabada, "Credit card default prediction using ML and DL techniques," *Internet of Things and Cyber-Physical Systems*, vol. 4, pp. 293–306, Jan. 2024, doi: 10.1016/j.iotcps.2024.09.001.
- [2] X. Zhu, Q. Chu, X. Song, P. Hu, and L. Peng, "Explainable prediction of loan default based on machine learning models," *Data Science and Management*, vol. 6, no. 3, pp. 123–133, Sep. 2023, doi: 10.1016/j.dsm.2023.04.003.
- [3] L. Feng, J. Qi, Y. Liu, and W. Wang, "The spillover effects of the 'Binance Incident' on financial markets: A study based on machine learning approach," *Financ. Res. Lett.*, vol. 71, Jan. 2025, doi: 10.1016/j.frl.2024.106383.
- [4] Y. Ma, P. Zhang, S. Duan, and T. Zhang, "Credit default prediction of Chinese real estate listed companies based on explainable machine learning," *Financ. Res. Lett.*, vol. 58, Dec. 2023, doi: 10.1016/j.frl.2023.104305.
- [5] M. A. Arias-Serna, J. J. Quiza-Montealegre, L. F. Montes-Gómez, L. U. Clavijo, and A. F. Orozco-Duque, "Explainable Machine Learning Models for Credit Rating in Colombian Solidarity Sector Entities," *Journal of Risk and Financial Management*, vol. 18, no. 9, Sep. 2025, doi: 10.3390/JRFM18090489.
- [6] J. Li, J. Kang, J. Wu, H. Wang, and X. Yang, "Research on credit card default repayment prediction model," *Journal of Finance and Data Science*, vol. 10, Dec. 2024, doi: 10.1016/j.jfds.2024.100136.
- [7] S. A. Oyedotun et al., "The Role of Internal Audit in Fraud Detection and Prevention: A Multi-Contextual Review and Research Agenda," *Journal of Science Research and Reviews*, vol. 2, no. 2, pp. 76–85, May 2025, doi: 10.70882/josrar.2025.v2i2.51.
- [8] S. Lin, D. Song, B. Cao, X. Gu, and J. Li, "Credit risk assessment of automobile loans using machine learning-based SHapley Additive exPlanations approach," *Eng. Appl. Artif. Intell.*, vol. 147, May 2025, doi: 10.1016/j.engappai.2025.110236.
- [9] C. Garitta and L. Grassi, "Predicting break-even in FinTech startups as a signal for success," *Financ. Res. Lett.*, vol. 74, Mar. 2025, doi: 10.1016/j.frl.2024.106735.
- [10] H. Li and W. Wu, "Loan default predictability with explainable machine learning," *Financ. Res. Lett.*, vol. 60, p. 104867, Feb. 2024, doi: 10.1016/J.FRL.2023.104867.
- [11] E. A. Raheem, A. M. Dinar, M. A. Mohammed, and B. Al-Attar, "Improving Loan Status Prediction Accuracy with Generative Adversarial Networks: Addressing Data Scarcity and Bias," *Journal of Intelligent Systems and Internet of Things*, vol. 13, no. 1, pp. 251–258, 2024, doi: 10.54216/JISIOT.130118.
- [12] S. A. Oyedotun et al., "The Role of Internal Audit in Fraud Detection and Prevention: A Multi-Contextual Review and Research Agenda," *Journal of Science Research and Reviews*, vol. 2, no. 2, pp. 76–85, May 2025, doi: 10.70882/josrar.2025.v2i2.51.
- [13] L. Wang, Z. Yu, J. Ma, X. Chen, and C. Wu, "A Two-Stage Interpretable Model to Explain Classifier in Credit Risk Prediction," *J. Forecast.*, vol. 44, no. 7, pp. 2132–2150, Nov. 2025, doi: 10.1002/FOR.3288.
- [14] M. Ikermane and Y. Rachidi, "Enhancing Financial Decision-Making with Explainable Artificial Intelligence: A Case Study in Credit Risk Assessment," 2025, pp. 255–261. doi: 10.1007/978-3-031-88304-0_34.
- [15] N. B. Unuigbokhai et al., "ADVANCEMENTS IN FEDERATED LEARNING FOR SECURE DATA SHARING IN FINANCIAL SERVICES," *FUDMA JOURNAL OF SCIENCES*, vol. 9, no. 5, pp. 80–86, May 2025, doi: 10.33003/fjs-2025-0905-3207.

- [16] Y. Filali and H. El Asri, "Explainable Artificial Intelligence in Credit Scoring: A Novel Framework Integrating Decision Trees and LLMs," 2026, pp. 29–39. doi: 10.1007/978-3-032-07718-9_3.
- [17] V. Ravi et al., "Explainable AI (XAI) for Credit Scoring and Loan Approvals," 2026, pp. 351–368. doi: 10.1007/978-981-96-8998-9_29.
- [18] Sarder Abdulla Al Shiam et al., "Credit Risk Prediction Using Explainable AI," *Journal of Business and Management Studies*, vol. 6, no. 2, pp. 61–66, Mar. 2024, doi: 10.32996/jbms.2024.6.2.6.
- [19] G. F. Bone-Winkel and F. Reichenbach, "Improving credit risk assessment in P2P lending with explainable machine learning survival analysis," *Digit. Finance*, vol. 6, no. 3, pp. 501–542, Sep. 2024, doi: 10.1007/s42521-024-00114-3.
- [20] A. A. S. Abubakar, "Artificial Intelligence In Financial Decision Systems: Transforming Risk Assessment And Investment Practices In The Era Of Digital Scientific Culture," *Int. J. Environ. Sci.*, pp. 3472–3483, Jul. 2025, doi: 10.64252/9240es77.
- [21] Q. Xu et al., "SHAP-based Interpretable Models for Credit Default Assessment Using Machine Learning," in 2024 14th International Conference on Software Technology and Engineering (ICSTE), IEEE, Aug. 2024, pp. 213–217. doi: 10.1109/ICSTE63875.2024.00044.
- [22] A. Kanaji, S. Mahale, A. Gattani, R. Rai, P. Tawde, and V. Sawant, "Loan Acceptance Prediction and Analysis Using Explainable AI Techniques," 2026, pp. 175–186. doi: 10.1007/978-981-96-9971-1_12.
- [23] G. Oise and S. Konyeha, "Environmental impacts in e-waste management using deep learning," *Discover Artificial Intelligence*, vol. 5, no. 1, p. 210, Aug. 2025, doi: 10.1007/s44163-025-00376-9.
- [24] Md. M. Islam, A. Sohag, M. Hasan, Md. K. Islam, and Md. N. Sultan, "XAI-Driven Model Explainability and Prediction of P2P Bank Loan Default Network," 2024, pp. 109–121. doi: 10.1007/978-981-99-8937-9_8.
- [25] Mainuddin Adel Rafi, S M Iftekhhar Shaboj, Md Kauser Miah, Iftekhhar Rasul, Md Redwanul Islam, and Abir Ahmed, "Explainable AI for Credit Risk Assessment: A Data-Driven Approach to Transparent Lending Decisions," *Journal of Economics, Finance and Accounting Studies*, vol. 6, no. 1, pp. 108–118, Feb. 2024, doi: 10.32996/jefas.2024.6.1.11.
- [26] G. P. Oise et al., "DECENTRALIZED DEEP LEARNING IN HEALTHCARE: ADDRESSING DATA PRIVACY WITH FEDERATED LEARNING," *FUDMA JOURNAL OF SCIENCES*, vol. 9, no. 6, pp. 19–26, Jun. 2025, doi: 10.33003/fjs-2025-0906-3714.
- [27] X. Zheng, "Ensemble Learning and SHAP-based Interpretability for AI-Driven Credit Risk Assessment in Financial Default Prediction," in *Proceedings of the 2025 6th International Conference on Computer Science and Management Technology*, New York, NY, USA: ACM, Dec. 2025, pp. 168–173. doi: 10.1145/3795154.3795180.
- [28] B. Cui, "Loan Forecasting Based on Machine Learning and Variable Correlation Analysis," *Applied and Computational Engineering*, vol. 107, no. 1, pp. 127–138, Dec. 2024, doi: 10.54254/2755-2721/2024.18131.
- [29] Y. Chen, D. Wu, Y. Ren, C. Shao, D. Wang, and Y. Gu, "A sensorless centrifugal pump operation states estimation method based on HSSA-BPNN," *Measurement (Lond).*, vol. 237, Sep. 2024, doi: 10.1016/j.measurement.2024.115238.
- [30] A. A. Montevechi, R. de C. Miranda, A. L. Medeiros, and J. A. B. Montevechi, "Advancing credit risk modelling with Machine Learning: A comprehensive review of the state-of-the-art," *Eng. Appl. Artif. Intell.*, vol. 137, p. 109082, Nov. 2024, doi: 10.1016/j.engappai.2024.109082.
- [31] J. Černevičienė and A. Kabašinskas, "Explainable artificial intelligence (XAI) in finance: a systematic literature review," *Artif. Intell. Rev.*, vol. 57, no. 8, p. 216, Jul. 2024, doi: 10.1007/s10462-024-10854-8.
- [32] P. E. de Lange, B. Melsom, C. B. Vennerød, and S. Westgaard, "Explainable AI for Credit Assessment in Banks," *Journal of Risk and Financial Management*, vol. 15, no. 12, p. 556, Nov. 2022, doi: 10.3390/jrfm15120556.
- [33] M. K. Nallakaruppan, B. Balusamy, M. L. Shri, V. Malathi, and S. Bhattacharyya, "An Explainable AI framework for credit evaluation and analysis," *Appl. Soft Comput.*, vol. 153, p. 111307, Mar. 2024, doi: 10.1016/j.asoc.2024.111307.
- [34] E. Kocoglu and T. Ersoz, "Trends and Recommendations in Machine Learning Applications for Credit Risk Prediction," 2025, pp. 789–812. doi: 10.1007/978-3-031-84873-5_61.