International Journal of Data and Network Science 10 (2026) 1031–1048

Contents lists available at GrowingScience

International Journal of Data and Network Science

homepage: www.GrowingScience.com/ijds

Explainable ensemble machine learning for disclosure-informed credit risk assessment in peer-to-business lending

Gihan M. Ali[a*]

[a]*Department of Accounting, College of Business Administration in Hawtat Bani Tamim, Prince Sattam bin Abdulaziz University, Saudi Arabia*

C H R O N I C L E

Article history:
Received February 3, 2026
Received in revised format March 5, 2026
Accepted April 25 2026
Available online
April 25 2026

Keywords:
Peer-to-Business (P2B) Lending
Credit Risk Assessment
Explainable Artificial Intelligence (XAI)
Ensemble Machine Learning
Borrower Disclosures
Information Asymmetry
FinTech
SHAP Explainability

A B S T R A C T

This study develops an explainable ensemble machine learning framework for sustainable credit risk assessment in peer-to-business (P2B) lending, a rapidly expanding FinTech model that enhances access to financing for small and medium-sized enterprises (SMEs). The increasing reliance on algorithmic decision-making underscores the need for transparent and interpretable credit evaluation, particularly in environments characterized by information asymmetry and reliance on borrower-provided disclosures. To address these challenges, a heterogeneous ensemble model is proposed, integrating Random Forest (RF), Light Gradient Boosting Machine (LGBM), and deep learning classifiers within a soft-voting architecture. Feature selection and class balancing are guided by RF importance scores and resampling techniques, resulting in a compact and interpretable 12-feature set comprising pricing, contractual, and transaction-level variables derived from borrower disclosures. Using real-world transaction-level data from a UK platform, the proposed model achieves improved predictive performance (ROC-AUC = 0.767) compared to a neural network baseline (ROC-AUC = 0.717) under severe class imbalance. SHAP-based explainability analysis identifies Maturity Days, Annualised Gross Yield, Advance Rate, and Discount Rate as the most influential predictors, highlighting how disclosed information is translated into pricing and contractual terms in digital lending markets. The findings demonstrate that disclosure-informed features can enhance both predictive accuracy and interpretability, supporting more transparent, robust, and accountable credit risk assessment in FinTech lending environments.

© 2026 by the authors; licensee Growing Science, Canada.

1. Introduction

Financial technology (FinTech) refers to technology-driven innovations in the financial sector, where specialized firms leverage artificial intelligence, machine learning (ML), big data, and blockchain to enhance financial access, inclusion, and efficiency across markets (Appiah & Agblewornu, 2025). The evolution of FinTech has transformed traditional lending by enabling digital platforms that directly connect investors with borrowers. A key innovation in this domain is peer-to-business (P2B) lending, which allows small and medium-sized enterprises (SMEs) to obtain financing by selling unpaid invoices through online platforms. Known as invoice trading, this model provides SMEs with short-term working capital by discounting receivables in exchange for immediate liquidity. Unlike conventional loans, P2B lending bypasses banks and financial intermediaries, offering greater speed, transparency, and accessibility for firms often excluded from formal credit channels (Dorfleitner et al., 2017).

Crowdlending platforms assess invoice risk, assign credit ratings, and price financing—typically using fixed-rate mechanisms. Risk is distributed among multiple investors who collectively fund parts of the same invoice. Prominent platforms such as Kriya have gained traction despite offering limited borrower-level disclosures. Consequently, credit risk assessments rely primarily on invoice-level variables (e.g., advance rate, maturity, and total face value), macroeconomic

* Corresponding author
E-mail address g.ali@psau.edu.sa (G.M. Ali)

ISSN 2561-8156 (Online) - ISSN 2561-8148 (Print)
© 2026 by the authors; licensee Growing Science, Canada.
doi: 10.5267/j.ijdns.2026.4.022

signals, and platform-specific behavior. Corrales et al. (2024) highlight a persistent gap between the implicit probability of default (IPD) and observed default probability (OPD), indicating potential pricing inefficiencies driven by borrower opacity and operational risk. Grounded in agency and signaling theory, this study posits that even limited borrower disclosures serve as critical signals that mitigate information asymmetry and are ultimately embedded in pricing and contractual terms used for credit risk assessment (Long et al., 2022).

While P2B lending promotes financial inclusion and supports the Sustainable Development Goals (SDGs) by expanding credit access for underbanked SMEs, it also presents new challenges for credit risk prediction. Compared with traditional financial institutions, P2B platforms operate with sparse, inconsistent, and heterogeneous data. These limitations are further compounded by class imbalance, limited historical depth, and asymmetric information between borrowers and investors, which collectively hinder the performance of standard ML models.

Recent research has also underscored the broader role of digital transformation and financial innovation in advancing sustainable development through peer-to-peer (P2P) lending and fintech. Sayari et al. (2025) demonstrate that digital economy indicators, such as internet access, digital payments, and fintech adoption, significantly promote socioeconomic progress and environmental sustainability, even amid geopolitical uncertainty. Similarly, Avgeri et al. (2023) find that P2P lending enhances financial inclusion by enabling marginal borrowers to access mortgage financing, thereby supporting SDGs through reduced credit inequality. Complementing these macro-level insights, Chulawate & Kiattisin (2023) identify key success factors, spanning risk management, trust, and information transparency, that strengthen P2P lending's contribution to financial innovation and sustainability within developing economies. Collectively, these studies suggest that fintech-driven P2P lending not only fosters inclusive finance but also acts as a catalyst for sustainable economic transformation.

At the methodological level, recent studies have advanced explainable and ensemble-based machine learning approaches for credit risk prediction in FinTech contexts. Nwafor et al. (2024) propose an explainable hybrid deep learning model that integrates a one-dimensional convolutional neural network with XGBoost to improve transparency and fairness in automated credit decisions within P2P lending platforms. Similarly, Xiao et al. (2025) introduce an example-dependent, cost-sensitive selective deep ensemble that addresses class imbalance and enhances interpretability in customer credit scoring. Collectively, these works highlight the growing academic emphasis on explainable and ensemble-based credit modeling within consumer P2P lending. However, credit risk assessment in P2B lending remains largely unexplored. Existing research has primarily concentrated on P2P lending, where explainable artificial intelligence (XAI) techniques are increasingly applied to improve model transparency and interpretability. To date, no study has examined how XAI and heterogeneous ensemble methods can be applied within the P2B context. Moreover, traditional statistical approaches, such as logistic and probit regression, still dominate much of the literature but suffer from strong distributional assumptions, sensitivity to outliers, and limited capacity to capture nonlinear, high-dimensional relationships. Corrales et al. (2024) recently applied ML techniques to predict defaults in P2B lending, incorporating macroeconomic and invoice-level variables while addressing class imbalance through sampling. Yet their framework relied solely on single ensemble methods and did not extend to a heterogeneous ensemble architecture or employ advanced model interpretability techniques for credit risk explanation.

Ensemble learning enhances financial analytics by combining multiple models to improve predictive accuracy and robustness in complex financial environments. Prior studies demonstrate that ensemble frameworks such as Random Forest, LightGBM, and stacking architectures outperform conventional models in financial risk prediction (Huang, 2024; Shan & Gao, 2025; Zhu et al., 2024). Nevertheless, research on heterogeneous ensemble approaches remains limited in P2B lending, where both model diversity and explainability are essential for transparent and reliable credit risk assessment. This study addresses these methodological gaps by developing a robust heterogeneous voting ensemble model that integrates Random Forest (RF), Light Gradient Boosting Machine (LGBM), and Deep Learning (DL) classifiers. The model combines the complementary strengths of diverse learners to improve prediction accuracy and generalization in highly imbalanced settings. To ensure robustness, the analysis employs 10-fold cross-validation, and the statistical significance of performance improvements over individual base learners is assessed using paired t-tests, confirming that the ensemble significantly outperforms all base models ($p < 0.001$). Feature selection was performed using RF importance scores in combination with various sampling strategies (e.g., random oversampling, undersampling, and hybrid SMOTE-based methods) to identify the optimal pairing of resampling technique and feature subset. This process resulted in a 12-feature configuration that achieved the best trade-off between model simplicity and predictive performance.

Empirical validation using real-world, transaction-level data from the Kriya platform demonstrates that, despite relying exclusively on publicly available invoice-level variables, the proposed ensemble achieves superior performance across key evaluation metrics. The model attains an ROC-AUC of 0.767, alongside higher accuracy (0.907) and specificity (0.915), reflecting a more balanced detection of defaults and an enhanced ability to correctly identify non-default cases. These improvements are particularly important in imbalanced FinTech environments, where minimizing false positives is critical for responsible credit risk assessment. Importantly, the ensemble substantially outperforms the best-performing neural network model developed by Corrales et al. (2024), which reported an ROC-AUC of 0.717, accuracy of 0.855, and specificity of 0.860, confirming the robustness and effectiveness of the proposed approach.

Feature importance analysis and interpretability analysis using SHapley Additive exPlanations (SHAP) reveal that the most influential predictors of default are Maturity Days (16.8%), Annualised Gross Yield (16.5%), Advance Rate (15.4%), and Discount Rate (13.7%). Longer maturities, higher yields, and elevated discount rates are associated with increased default risk, while higher advance rates correspond to lower risk, suggesting that platforms tend to grant larger advances to more creditworthy borrowers. These findings are consistent with established credit risk theory and highlight the pivotal role of loan structure and pricing terms in P2B credit risk assessment. This research makes several key contributions to the FinTech and accounting analytics literature. First, it introduces a novel explainable heterogeneous ensemble model for P2B credit risk assessment that enhances both predictive accuracy and interpretability. Second, it systematically addresses the class imbalance problem through integrated resampling and RF-based feature selection. Third, it validates the model using real-world platform data with rigorous cross-validation and statistical testing, ensuring robustness and generalizability. Finally, by integrating SHAP-based explainability, the study bridges the gap between complex ML models and actionable financial insights, contributing to more transparent, inclusive, and data-driven digital lending ecosystems aligned with the SDGs. The remainder of this paper is structured as follows. Section 2 reviews relevant literature. Section 3 outlines the proposed methodology. Section 4 presents empirical results and discussions. Section 5 concludes the study.

2. Related Work

Credit risk assessment in P2P lending has garnered significant attention in recent years, driven by the need to improve investment decisions and predict borrower defaults. P2P lending involves direct borrowing between individual borrowers and investors through digital platforms. Online invoice trading (P2B) represents a specialized segment within the crowdlending market, where businesses can obtain working capital financing through digital platforms. This crowdlending model allows multiple investors to fund discounted invoices in exchange for fees, with the platform assessing transaction risks and assigning ratings. While some platforms initially used auction-based pricing, most have transitioned to fixed-price models (Dorfleitner et al., 2017).

P2P lending platforms have emerged as significant contributors to SDGs. These platforms can promote financial inclusion by providing alternative financing options for marginalized borrowers, potentially circumventing loan-to-value cap policies and increasing access to mortgages (Avgeri & Psillaki, 2023). Sharia-compliant fintech, including P2P lending, has shown growth and supports SDGs related to economic development and poverty reduction (Trimulato et al., 2022). P2P lending can accelerate the transition to a low-carbon economy by allocating investments to sustainable projects, although compliance risks such as money laundering and fraud must be addressed (Teichmann et al., 2024). In prosocial crowdfunding platforms like Kiva, the presence of SDG themes in project descriptions can influence campaign success, with some themes having positive impacts and others negative (Maiolini et al., 2024). The growth of P2P lending contributes to the broader field of sustainable finance, which offers significant business and investment opportunities. Overall, P2P lending demonstrates potential in advancing SDGs, particularly in areas of financial inclusion and sustainable economic growth.

2.1. Disclosure Determinants of Credit Risk in Crowdlending

Prior research revealed that the determinants of default in crowdlending are multifaceted, encompassing borrower, and loan attributes, firm characteristics, and macroeconomic conditions. These studies highlight the complexity of credit risk assessment in crowdlending and emphasize the importance of incorporating diverse variables to improve prediction accuracy.

Borrower-specific factors, such as income, credit history, and indebtedness, are consistently identified as significant predictors of default. Serrano-Cinca et al. (2015) found that loan purpose, income, housing situation, and credit history are key determinants, with small business loans exhibiting higher default rates compared to wedding loans. The grade assigned by P2P platforms was the most predictive factor, underscoring the importance of platform-specific risk assessments. Similarly, Polena & Regner (2018) demonstrated that annual income, debt-to-income ratio, and recent credit inquiries are significant across all loan risk classes, while variables like revolving credit utilization vary in importance depending on the risk category. Lin et al. (2017) expanded on these insights by examining demographic factors in the Chinese P2P lending market. They found that gender, age, marital status, education, work experience, company size, and financial metrics such as debt-to-income ratio and delinquency history significantly influence default risk, highlighting the cultural and regional nuances in credit risk assessment. Weng & Luo (2021) analyzed information asymmetry in China's P2P lending market using a four-party game model and empirical methods. Results show that platform self-guarantees reduce both borrower and investor participation, lowering transaction volume, while third-party guarantees increase investor participation and overall transactions.

Macroeconomic and behavioral factors also shape credit risk dynamics. Avgeri & Psillaki (2023) incorporated macroeconomic indicators, such as the House Price Index, Consumer Sentiment Index, and unemployment rates, into their logistic regression model. Their analysis revealed that higher housing prices and consumer sentiment are associated with lower default rates, while higher unemployment and GDP growth correlate with increased defaults. These findings underscore the interconnectedness of P2P lending markets with broader economic trends. Nigmonov et al. (2022) further validated the impact of macroeconomic factors, demonstrating that higher interest rates and inflation increase the probability of default in the U.S. P2P lending market. Their probit regression analysis revealed that the effect of interest rates is more pronounced for

loans with lower credit ratings, highlighting the differential impact of macroeconomic conditions across risk classes. Complementing these findings, Kuang et al. (2023) employ supervised machine learning and sentiment analysis to examine how media and social media sentiment affect default probability and cost of capital in China's P2P lending market, finding that only positive sentiment changes significantly reduce default risk, confirming an asymmetric influence of investor sentiment on financial stability.

In the context of P2B lending, borrower reputation and firm characteristics are critical determinants of default. Kgoroeadira et al. (2023) analyzed small business loans on the Prosper.com platform and found that borrower reputation, as indicated by credit grade and repayment history, is the dominant factor influencing repayment performance, particularly for established firms. Interestingly, new firms were not more likely to default than established ones, challenging the assumption that newer businesses are inherently riskier. They underline the importance of reputation-based metrics in P2B lending risk assessment. Xu et al. (2021) examined UK P2B lending data and identified credit group, loan purpose, sector type, loan amount, interest rate, and loan term as key determinants of default risk. Their findings revealed that loans for capital needs, higher interest rates, and longer loan terms are associated with increased default probabilities.

2.2. Credit Risk Assessment Methodologies in Crowdlending

Prior research has explored a variety of methodologies, ranging from traditional statistical models to advanced ML techniques, to address the unique challenges of crowdlending, such as information asymmetry, imbalanced datasets, and the dynamic nature of borrower behavior. Guo et al. (2016) developed an instance-based credit risk assessment model in P2P lending. This model evaluates the risk and return of individual loans and integrates a portfolio optimization framework with boundary constraints to enhance investment decisions. Their proposed model demonstrated its superiority over existing methods, highlighting its potential to improve investment performance. Cho et al. (2019) proposed an investment decision model for the P2P lending market using an instance-based entropy fuzzy support vector machine (IEFSVM) to classify loan status. Applied to Lending Club data, the model accurately identifies fully paid loans and, when combined with a multiple regression-based portfolio, outperforms EFSVM and six other state-of-the-art classifiers in both classification accuracy and investment performance, demonstrating its practical value for P2P investment decisions.

Similarly, Chen et al. (2019) introduced regression models, including Linear Regression and RF Regression, to predict the Net Return Rate (NRR) as a proxy for credit risk. Their findings revealed that LR outperformed other models, achieving a 76% lending ratio with low computational cost and root mean squared error (RMSE). To address the complexity of P2P lending markets, Liu et al. (2024) proposed a two-step ML approach that combines network analysis with traditional credit risk factors. By deriving network centrality metrics and integrating them with financial attributes, their model significantly improved prediction accuracy. Their study underscores the importance of considering borrower connectivity and position within the lending network. Yuan et al. (2018) developed a BP neural network model that incorporates data from P2P platforms and commercial banks. Their model demonstrated strong performance in predicting default rates, compared to existing credit risk assessment methods.

Survival analysis has also been applied to predict the timing of loan defaults. Byanjankar (2017) used survival analysis to estimate loan survival probabilities, integrating these results into neural network models for improved classification. Wang et al. (2018) proposed a behavioral scoring model using mixture survival analysis, which outperformed traditional methods in predicting dynamic default probabilities over time. Ensemble learning methods have also emerged as powerful tools for credit risk assessment. Aleksandrova and Parusheva (2021) demonstrated the effectiveness of ensemble models, such as Stacked Ensembles and XGBoost, in optimizing financial outcomes. Their approach reduced default loan ratios from 20% to 12% while improving portfolio structures, highlighting the practical benefits of ensemble techniques in P2P lending. Although the literature on crowdlending has made considerable progress in identifying default determinants and applying predictive methodologies, the majority of studies remain concentrated on P2P lending, with limited examination of P2B contexts (Aleksandrova & Parusheva, 2021; Chen et al., 2019; Guo et al., 2016; Lin et al., 2017; Polena & Regner, 2018; Serrano-Cinca et al., 2015). Most existing research applies statistical and ML techniques, ranging from regression to ensemble methods and neural networks, predominantly within the P2P domain. While the adoption of advanced ML models has significantly improved credit risk assessment in P2P lending, limited attention has been given to the application of heterogeneous voting ensembles in this area. Unlike traditional single-model approaches, heterogeneous ensembles integrate multiple base learners to reduce prediction errors and enhance robustness, offering a more flexible and reliable framework for complex financial prediction tasks. Building on this gap, the present study develops a robust voting ensemble model for credit risk assessment in P2B lending that balances predictive accuracy with interpretability, thereby addressing both industry and regulatory requirements and contributing to the ongoing evolution of the FinTech landscape.

3. Proposed Framework

3.1. General Context

This study develops a credit risk assessment framework for P2B lending using a heterogeneous ensemble voting model. The methodology consists of two stages: training and testing. As shown in Figure 1, preprocessing involves cleaning and

normalizing P2B crowdlending data. Class imbalance is addressed through random oversampling, while feature relevance is determined using RF. The predictive framework integrates RF, LGBM, and DL in a heterogeneous ensemble to enhance accuracy and robustness. Model performance in the testing phase is evaluated using ROC-AUC, F1 score, sensitivity, specificity, and accuracy.

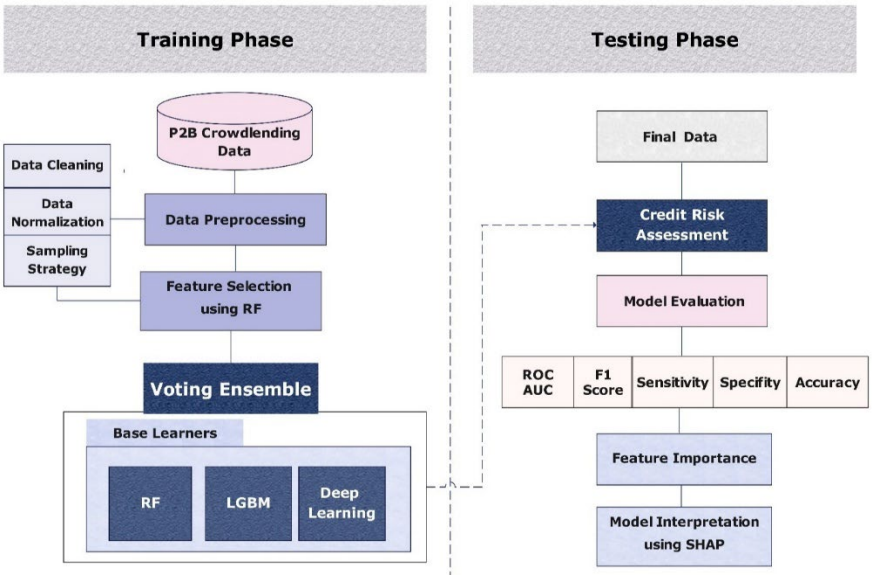

Fig. 1. Block diagram of the proposed credit risk prediction system.

3.2. Phase 1: Data Preparation

Step 1: Dataset Overview

This study utilizes a dataset collected by Corrales et al. (2024) from publicly accessible data provided by Kriya, a UK-based crowdlending platform. Kriya specializes in factoring and P2B lending. As of November 2023, the platform had funded approximately £3.4 billion to SMEs, offering an expected net yield of 4–6% and a loan recovery rate of 78.01%. The dataset includes loans issued between March 2011 and December 2019 (Corrales et al., 2024). The target variable in this research reflected loan status and is binary: classified as "Yes" if the loan was in default (i.e., minority class) and "No" if the loan was not in default (i.e., majority class).

Step 2: Data Preprocessing

Financial datasets often contain missing values, inaccuracies, and inconsistencies that can adversely affect model performance. Therefore, data preprocessing is essential to improve the accuracy and effectiveness of ML models. In this dataset, only the categorical variable TradeBand contains missing values, which are addressed using mode imputation (Kwak & Kim, 2017). The final preprocessed dataset contains 60,226 observations. To address skewness in the data, all continuous variables are subjected to log transformation.

Table 1
Voluntary Disclosures on the P2B Platform

Variable	Description
The Dependent Variable	
Default	A dummy variable that signifies whether the loan resulted in default.
The Independent Variables	
Advance rate	The proportion of an invoice's total amount that a factor pays upfront during purchase.
Annualized gross yield	The effective annual return on factoring, accounting for compound interest.
Total face value	The full monetary value of an invoice, expressed in pounds sterling.
Days/Maturity	The time span, in days, between the advance date and the anticipated repayment date.
Total Gross Advance	The amount disbursed by the P2B platform to the business borrower against an invoice, reflecting the total advance provided.
Trade type	A classification indicating the trade type, such as standard, whole ledger, multi-debtor, or license fee
Trade band	A variable representing the sector or industry of the borrowing party.
Discount Rate	The rate applied to the invoice's face value to calculate the discount (financing cost) charged by the P2B platform in exchange for providing an advance prior to the invoice's maturity.
Unadjusted Total Gross Advance	The full amount advanced by the P2B platform to the business borrower against an invoice, prior to the deduction of any fees, risk adjustments, or holdbacks.

To handle zero values, a constant of one is added to each feature before transformation. One-hot encoding was applied to the categorical variables. Trade type was transformed into four dummy variables: LicenceFee, MultiDebtor, Standard, and WholeLedger. Similarly, Trade band was encoded into ten dummy variables representing Bands 1 through 10. The original dataset contains 10 variables, including the target variable (Default). After applying one-hot encoding to the categorical variables (TradeType and TradeBand), the dataset expanded to 22 variables, including the target variable. Consequently, 21 predictive features were available prior to feature selection. Table 1 presents all the voluntary disclosure variables used in this study, including invoice-related variables, advance rate, trade type, total face value, and maturity, sourced directly from the Kriya database.

Step 3: Imbalance Treatment and Feature Selection

The dataset exhibits a strong class imbalance, which tends to bias classifiers toward predicting the majority class. To mitigate this, several external resampling strategies—independent of any particular learning algorithm—were applied prior to model training (Ali et al., 2022). These included Synthetic Minority Oversampling Technique (SMOTE), which generates synthetic minority class samples by interpolating between existing instances to reduce imbalance (Chawla et al., 2002; Gao et al., 2025); random oversampling, which increases the representation of the minority class by duplicating existing samples; and random undersampling, which balances the dataset by randomly removing examples from the majority class (Carvalho et al., 2025; He & Garcia, 2009). In addition, two hybrid methods were considered: SMOTETomek, which combines SMOTE with the removal of Tomek links—pairs of nearest-neighbor samples from opposite classes that are difficult to separate—to refine class boundaries; and SMOTEENN, which applies SMOTE and then removes noisy or ambiguous instances using the Edited Nearest Neighbors (ENN) rule, resulting in a cleaner dataset. These methods remain among the most widely used in recent imbalanced learning research, with recent surveys emphasizing the growing preference for hybrid and adaptive resampling approaches to enhance classifier performance (Batista et al., 2004; Carvalho et al., 2025; Gao et al., 2025).

Alongside sampling, RF was also used to perform feature selection in combination with each sampling strategy. The goal was to identify the best pairing of resampling method and feature subset, with RF acting as the base classifier. The use of 10-fold cross-validation further helped guard against overfitting during this selection process.

3.3. Phase 2: Model Development

Step 1: Specification of the Heterogeneous Ensemble Model

This study employs ensemble learning to construct a predictive model for loan delinquency classification. Ensemble learning integrates multiple algorithms to improve decision-making in supervised learning tasks. The core principle is that combining predictions from several weak learners yields a stronger and more reliable model, as collective learning mitigates individual biases and errors, a concept often referred to as the “wisdom of the crowd” (Kunapuli, 2023; Mohammed & Kora, 2023). Empirical evidence consistently shows that ensemble methods outperform single models in terms of prediction accuracy and generalization (Ali, 2026b, 2026a, 2026c; Thabet et al., 2024).

Ensemble models are typically classified as either homogeneous or heterogeneous. Homogeneous ensembles use multiple instances of the same algorithm trained on different data subsets. The two most common homogeneous methods are bagging and boosting. Bagging involves parallel training on random subsets of data and aggregating results, as seen in algorithms such as RF and Extra Trees, with RF adopted as a base learner in this study and Extra Trees used for comparative analysis. Boosting, conversely, builds models sequentially, emphasizing previously misclassified instances to improve accuracy. Common boosting algorithms include Adaptive Boosting (ADA), Gradient Boosted Regression Trees (GBR), LGBM, Categorical Boosting (CAT), and Extreme Gradient Boosting (XGB). In this research, LGBM is selected as a base learner, while the others serve as benchmarks for performance comparison (Mienye & Sun, 2022; Mohammed & Kora, 2023).

In contrast, heterogeneous ensembles integrate different types of algorithms, or even homogeneous ensembles, to capture diverse data representations and reduce bias, as illustrated in Fig. 2. This approach offers two major advantages: (1) it combines complementary model perspectives to enhance prediction robustness, and (2) it improves generalization by training all models on the complete dataset (Porwik et al., 2019). Building a heterogeneous ensemble typically involves three stages: (1) training a variety of individual classifiers or regressors on the training data to form a pool of base learners; (2) selecting the most effective subset of these learners based on domain relevance and performance metrics; and (3) aggregating their predictions using a voting mechanism or meta-learner (Bogaert & Delaere, 2023). This multi-model integration enables the proposed ensemble framework to capitalize on the complementary strengths of diverse algorithms, thereby enhancing both predictive accuracy and model stability in forecasting P2P loan delinquencies. Therefore, the proposed ensemble integrates heterogeneous learning paradigms to capture different aspects of the data structure. Tree-based models such as Random Forest (RF) and Light Gradient Boosting Machine (LGBM) are widely recognized for their strong performance on structured tabular datasets, particularly in financial risk prediction, due to their ability to model nonlinear relationships and feature interactions while remaining robust to noise and feature scaling. Combining bagging-based (RF) and boosting-based (LGBM) learners provides complementary strengths in terms of variance reduction and bias correction. Empirical studies have consistently

shown that tree-based ensembles remain highly competitive for tabular prediction tasks across many domains, including financial analytics and credit risk modelling (Ali, 2026a, 2026b; Ali & Yahia, 2025). To increase model diversity and capture additional nonlinear feature representations, a deep learning component is incorporated within the heterogeneous ensemble framework.

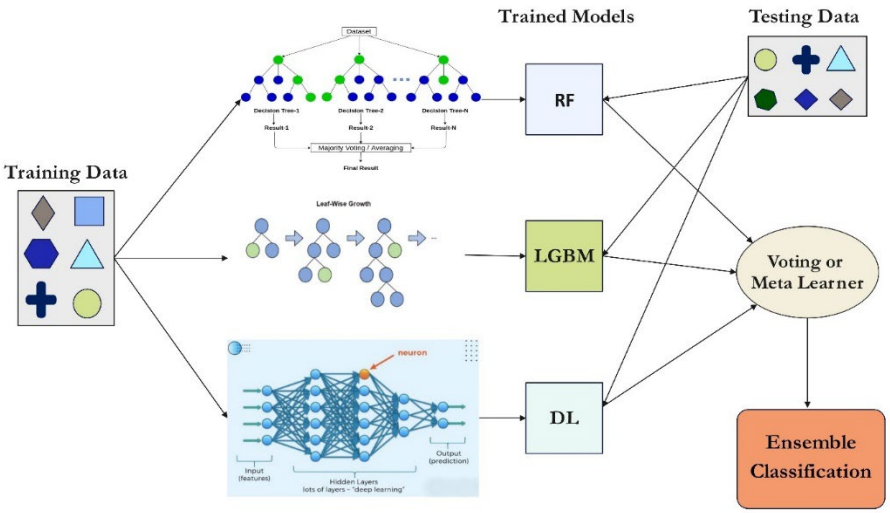

Fig. 2. Architecture of the heterogeneous ensemble model integrating RF, LGBM, and DL models through a voting or meta-learning approach to leverage their complementary strengths and enhance classification performance.

1- Random Forest (RF)

To improve classification accuracy and reduce overfitting, the RF algorithm builds multiple decision trees and combines their predictions through majority voting (Sánchez-Medina et al., 2025). This ensemble approach effectively captures nonlinear relationships and complex feature interactions, making it well suited for high-dimensional FinTech datasets. Additionally, RF demonstrates strong resilience to noise and missing values, ensuring reliable and generalizable performance in credit risk prediction.

2- Light Gradient Boosting Machine (LGBM)

The LGBM is an advanced implementation of the Gradient Boosting Decision Tree (GBDT) framework. Its leaf-wise (best-first) tree growth strategy prioritizes expanding the leaf with the greatest loss reduction, improving predictive accuracy and error minimization (Thai, 2022). LGBM further enhances training efficiency through Gradient-based One-Sided Sampling (GOSS), which focuses on the most informative instances, and Exclusive Feature Bundling (EFB), which reduces data dimensionality (Sarıkaya et al., 2022). Additionally, its histogram-based algorithms and parallel learning strategies optimize computational performance (Machado et al., 2019; Yang et al., 2021). These features make LGBM a scalable, accurate, and efficient base learner in the proposed heterogeneous ensemble framework.

3- Deep Learning (DL)

DL is a subset of ML that employs multilayer neural networks to capture complex nonlinear relationships within data. By automatically learning hierarchical feature representations through gradient-based optimization methods such as Adam, DL eliminates the need for extensive manual feature engineering. Regularization techniques like dropout are applied to prevent overfitting and improve generalization (Mienye & Swart, 2024). In the context of P2B lending, DL is particularly suited for modeling borrower heterogeneity and dynamic financial behaviors that traditional models may overlook. Within the proposed heterogeneous ensemble, the DL model complements tree-based learners such as RF and LGBM by enhancing the system's capacity to detect intricate patterns in financial disclosures, ultimately improving the robustness and predictive accuracy of credit risk assessment. A DL model is developed to predict credit risk using a feedforward neural network trained with stratified 10-fold cross-validation. The architecture comprises an input layer corresponding to the selected transaction-level features, two fully connected hidden layers with ReLU activation functions (64 and 32 neurons, respectively), and a sigmoid output layer for binary classification. Dropout rates of 0.3 and 0.2 are applied to mitigate overfitting, and the model is optimized using the Adam optimizer with a learning rate of 0.001.

Step 2: Training and Validation of Base Classifiers

Prior to model training, all input features are standardized using z-score scaling to ensure consistent data distribution across learners. The base ML models, RF, LGBM, and DL, are trained using a 10-fold cross-validation procedure to ensure robustness and generalization. Specifically, model evaluation was performed using stratified 10-fold cross-validation, where the dataset was divided into ten approximately equal subsets. In each iteration, nine folds were used for training and one fold

was used for testing. This procedure was repeated ten times so that each observation served once as test data and nine times as training data. Stratification ensured that the class distribution of the target variable (Default) remained consistent across all folds. The reported performance metrics correspond to the mean and standard deviation across the ten folds. Model parameters were selected through empirical evaluation ensuring stable performance and mitigating overfitting. This strategy is supported by prior empirical evidence indicating that extensive hyperparameter tuning does not universally improve predictive performance. Probst et al. (2019) introduced the concept of tunability and demonstrated that many algorithms benefit marginally from exhaustive tuning when compared to empirically selected or default configurations. Furthermore, a recent large-scale study by Mantovani et al. (2024) showed that tuning a small subset of hyperparameters, or relying on default settings, can be sufficient to achieve optimal or near-optimal performance for decision tree induction algorithms across a wide range of datasets, reinforcing that exhaustive automated optimization is not always required. The final configurations of the base learners and the proposed voting ensemble are presented in Table 2. Classification thresholds are also adjusted to maximize key metrics such as the F1 score and recall (sensitivity), thereby enhancing the models' ability to accurately identify loan defaults while maintaining overall predictive stability across folds.

Table 2
Tuned parameters of base learners and the proposed voting ensemble.

Model / Parameter	Definition	Selected Value
RF		
n_estimators	Number of trees in the forest	300
max_depth	Maximum depth of each decision tree	8
Threshold	Probability cutoff for classification	0.47
LGBM		
Threshold	Probability cutoff for classification	0.46
DL		
Architecture	3-layer neural network (64 → 32 → 1) with ReLU activations and sigmoid output	
Dropout	Dropout rate for regularization	0.3 (layer 1), 0.2 (layer 2)
Optimizer	Optimization algorithm	Adam
learning_rate	Learning rate for Adam optimizer	0.001
Loss	Objective function	Binary cross-entropy
Epochs	Number of training iterations	30
batch_size	Training batch size	64
Threshold	Probability cutoff for classification	0.50
Proposed Ensemble		
base_estimators	RF, LGBM, and DL (equal weights)	3 equal voters
Voting	Voting strategy (averages predicted probabilities)	soft
Threshold	Final ensemble classification cutoff	0.47

Step 3: Model Evaluation Metrics

Model performance is evaluated using standard classification metrics, with particular emphasis on the confusion matrix for detailed performance assessment (James et al., 2013). Derived metrics include accuracy (correct classifications), specificity (true negatives), recall/sensitivity (true positives), and the F1 score, which balances precision and recall (Altman & Bland, 1994; Donner & Klar, 1996). Since accuracy alone can be misleading in imbalanced datasets, the confusion matrix provides a more comprehensive performance summary (Altman & Bland, 1994). The area under the ROC curve (ROC-AUC) further evaluates performance across decision thresholds, where values above 0.5 indicate acceptable and those approaching 1 reflect high classification accuracy (Guijarro et al., 2022). Collectively, these metrics enable robust comparison across statistical, ML, and ensemble models.

Step 4: Construction of the Ensemble Classifier

In the final modeling stage, the three base classifiers, RF, LGBM, and DL, are integrated into a heterogeneous ensemble voting classifier to enhance predictive accuracy. Each model contributes equally to the ensemble, and predictions are combined using a soft voting strategy that averages class probabilities from all base learners. The final classification outcome is determined by selecting the class with the highest aggregated probability, ensuring a balanced and reliable prediction of loan default.

Step 5: Comparative Performance Analysis

Several comparative analyses are conducted to evaluate the performance of the proposed ensemble classification model in predicting loan delinquencies. This comprehensive assessment examines the model's robustness and highlights potential improvements over its individual base classifiers, traditional statistical models, and other state-of-the-art ML algorithms.

Step 6: Feature Importance Analysis

To identify the key determinants of credit risk in P2B lending, the RF classifier is used as the base model for feature importance evaluation. RF provides a reliable measure of each variable's contribution to risk prediction, enabling a clearer understanding of which factors most strongly influence lending outcomes.

Step 7: Model Explanation

To enhance the interpretability, SHAP analysis is performed using TreeExplainer. This approach quantifies the contribution of each feature to the model's predictions, providing both global and local interpretability. The SHAP framework combines theoretical rigor with practical clarity, illustrating how individual predictors affect risk prediction (Lundberg et al., 2020). Additionally, Partial Dependence Plots (PDPs) are used to visualize the marginal effects of the top four influential variables on credit risk assessment.

4. Experimental Results and Discussion

This section begins by examining the effects of resampling techniques and feature selection on RF performance. It then evaluates the predictive capability of the proposed ensemble model through comparative analysis against various regression-based and ML models using multiple performance metrics. Furthermore, the importance of the selected features is analyzed to enhance the interpretability of the results and support data-driven decision-making. Finally, SHAP values and partial dependence plots are utilized to offer deeper insights into the model's behavior and the relationships between key predictors and credit risk. The section concludes with a discussion of results, highlighting key findings, implications, and their alignment with existing literature.

4.1. Impact of Resampling Strategies and Feature Selection on Model Performance

Table 3 reports the comparative performance of various resampling strategies and feature set sizes on the minority class classification task. Across all methods, random oversampling consistently achieved higher recall than no resampling or SMOTE-based approaches, which either yielded low recall or failed to detect the minority class entirely. Among the random oversampling configurations, the 12-feature model delivered a strong balance between classification performance and model simplicity, with a recall of 0.341 ± 0.033 and competitive ROC-AUC (0.754 ± 0.012). While the 21-feature model offered a slightly higher ROC-AUC, its lower recall (0.328 ± 0.033) indicates reduced sensitivity to minority cases. Since recall is a key metric in this study, prioritizing the detection of rare positive instances, the 12-feature oversampling configuration was selected for subsequent ensemble modeling, as it provided the most favorable balance between minority-class detection and overall model performance. Random oversampling was applied with a minority-to-majority sampling ratio of 0.53 on the training data only to mitigate class imbalance while preventing information leakage into the test set.

4.2. Analysis of Experimental Results

To evaluate the classification effectiveness of the proposed model in forecasting credit risk, its performance is benchmarked against several alternative classifiers. The comparative models are organized into four categories: (1) the model's base learners, (2) regression-based and standalone ML classifiers, (3) alternative ensemble techniques, and (4) a benchmark model from a previous study utilizing the same dataset.

Table 3
Performance comparison of resampling methods and feature configurations.

Sampler	N. Features	ROC-AUC (±)	F1 score (±)	Recall (±)	Accuracy (±)
Random Oversampling	**12**	**0.754 ± 0.012**	**0.109 ± 0.009**	**0.341 ± 0.033**	**0.913 ± 0.003**
Random Oversampling	21	0.757 ± 0.010	0.109 ± 0.008	0.328 ± 0.033	0.916 ± 0.002
Random Oversampling	10	0.756 ± 0.013	0.108 ± 0.008	0.347 ± 0.030	0.911 ± 0.004
Random Undersampling	10	0.747 ± 0.014	0.095 ± 0.011	0.384 ± 0.033	0.885 ± 0.011
Random Undersampling	12	0.747 ± 0.014	0.094 ± 0.012	0.375 ± 0.037	0.887 ± 0.012
Random Undersampling	21	0.751 ± 0.013	0.097 ± 0.011	0.367 ± 0.026	0.893 ± 0.011
SMOTE	21	0.725 ± 0.019	0.090 ± 0.021	0.180 ± 0.041	0.943 ± 0.005
SMOTE	12	0.721 ± 0.018	0.082 ± 0.013	0.271 ± 0.037	0.905 ± 0.005
SMOTE	10	0.723 ± 0.022	0.080 ± 0.013	0.258 ± 0.037	0.908 ± 0.006
SMOTETomek	21	0.721 ± 0.018	0.063 ± 0.008	0.477 ± 0.063	0.780 ± 0.007
SMOTEENN	12	0.720 ± 0.019	0.062 ± 0.008	0.499 ± 0.050	0.763 ± 0.012
SMOTETomek	12	0.719 ± 0.018	0.062 ± 0.008	0.470 ± 0.049	0.778 ± 0.011
SMOTEENN	21	0.721 ± 0.019	0.061 ± 0.006	0.512 ± 0.055	0.756 ± 0.006
SMOTEENN	10	0.718 ± 0.018	0.061 ± 0.006	0.479 ± 0.051	0.770 ± 0.011
SMOTETomek	10	0.717 ± 0.018	0.062 ± 0.006	0.462 ± 0.050	0.784 ± 0.010
No Resampling	21	0.756 ± 0.015	0.000 ± 0.000	0.000 ± 0.000	0.984 ± 0.000
No Resampling	12	0.756 ± 0.013	0.000 ± 0.000	0.000 ± 0.000	0.984 ± 0.000
No Resampling	10	0.755 ± 0.015	0.000 ± 0.000	0.000 ± 0.000	0.984 ± 0.000

4.2.1. Comparative Evaluation of the Proposed Ensemble and Its Base Models

As shown in Table 4, the proposed voting ensemble consistently outperforms its individual base learners across all evaluation metrics. The ensemble achieves the highest ROC-AUC (0.767 ± 0.016), surpassing RF (0.757 ± 0.013), LGBM (0.754 ± 0.021), and the DL model (0.750 ± 0.017), as illustrated in Figure 3. This improvement indicates superior discriminative

capability in distinguishing between positive and negative classes. Additionally, the ensemble delivers the best F1 score (0.111 ± 0.008), reflecting a more balanced trade-off between precision and recall. Although its sensitivity (0.373 ± 0.023) is comparable to the base learners, the ensemble demonstrates slightly higher specificity (0.915 ± 0.006) and overall accuracy (0.907 ± 0.006). The statistical significance of these gains is supported by paired t-tests, which show that the ensemble's ROC-AUC is significantly higher than that of each base model ($p < 0.001$). These results confirm that combining complementary learners through a voting strategy enhances generalization and overall predictive performance.

Table 4
Cross-validation results of the proposed ensemble compared with its base learners.

Model	ROC-AUC	F1 Score	Sensitivity	Specificity	Accuracy	t-test
DL	0.750 ± 0.017	0.096 ± 0.008	0.369 ± 0.046	0.899 ± 0.020	0.890 ± 0.019	5.819***
LGBM	0.754 ± 0.021	0.103 ± 0.012	0.372 ± 0.042	0.907 ± 0.005	0.899 ± 0.006	4.733***
RF	0.757 ± 0.013	0.107 ± 0.006	0.387 ± 0.021	0.907 ± 0.002	0.899 ± 0.002	5.806***
Proposed Ensemble	0.767 ± 0.016	0.111 ± 0.008	0.373 ± 0.023	0.915 ± 0.006	0.907 ± 0.006	_____

Note: Values are reported as mean ± standard deviation. Statistical significance of ROC-AUC improvements is based on paired t-tests (***$p < 0.001$).

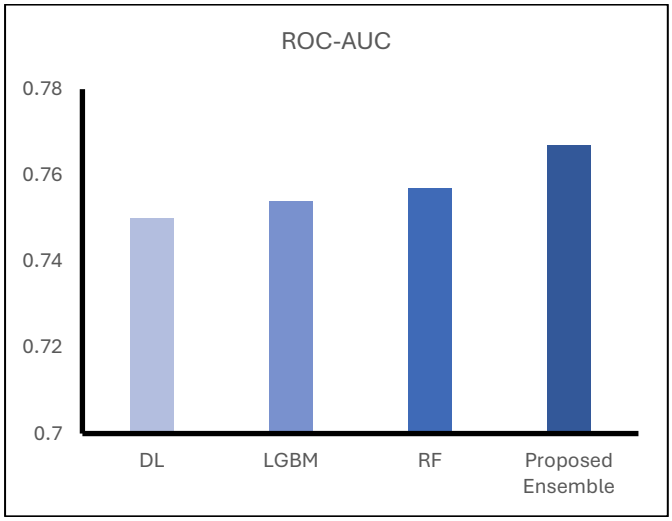

Fig. 3. Bar plots illustrating the ROC-AUC performance of the proposed ensemble compared to its base learners.

4.2.2. Comparative Evaluation of the Proposed Ensemble and Statistical and Individual ML Models

Table 5 presents a comparative evaluation of the proposed voting ensemble against a range of conventional statistical and ML classifiers, including K-Nearest Neighbors (KNN), Gaussian Naïve Bayes (GaussianNB), Decision Tree (DT), Lasso, ElasticNet, Ridge, Support Vector Classifier (SVC Linear), and Logistic Regression (LR).

Table 5
Cross-validation results of the proposed ensemble compared with statistical and ML models.

Model	ROC-AUC	F1 Score	Sensitivity	Specificity	Accuracy
KNN	0.545 ± 0.023	0.053 ± 0.015	0.123 ± 0.035	0.945 ± 0.003	0.932 ± 0.003
GaussianNB	0.676 ± 0.017	0.039 ± 0.001	0.910 ± 0.028	0.291 ± 0.006	0.301 ± 0.006
DT	0.710 ± 0.019	0.082 ± 0.008	0.411 ± 0.070	0.863 ± 0.026	0.856 ± 0.025
Lasso	0.713 ± 0.024	0.077 ± 0.014	0.271 ± 0.044	0.909 ± 0.004	0.899 ± 0.005
ElasticNet	0.713 ± 0.024	0.077 ± 0.014	0.271 ± 0.044	0.909 ± 0.004	0.899 ± 0.005
Ridge	0.716 ± 0.025	0.076 ± 0.014	0.284 ± 0.050	0.902 ± 0.005	0.892 ± 0.005
SVC Linear	0.716 ± 0.029	0.071 ± 0.011	0.273 ± 0.041	0.898 ± 0.007	0.889 ± 0.007
LR	0.717 ± 0.023	0.069 ± 0.010	0.391 ± 0.054	0.842 ± 0.006	0.835 ± 0.006
Proposed Ensemble	0.767 ± 0.016	0.111 ± 0.008	0.373 ± 0.023	0.915 ± 0.006	0.907 ± 0.006

The comparison is based on key performance metrics including ROC-AUC, F1 score, sensitivity, specificity, and overall accuracy. The results show that the proposed ensemble substantially outperforms all individual models across most evaluation metrics. As shown in Fig. 4, the ensemble achieves the highest ROC-AUC of 0.767 ± 0.016, indicating superior discriminative ability compared to classical approaches such as LR (0.717 ± 0.023), Ridge (0.716 ± 0.025), and DT (0.710 ± 0.019). This highlights the effectiveness of the ensemble in capturing complex non-linear relationships that individual models fail to learn. In terms of F1 score, the ensemble also demonstrates a marked improvement (0.111 ± 0.008) over the best-performing single models (e.g., DT: 0.082 ± 0.008, Ridge: 0.076 ± 0.014). Although the absolute F1 score values remain low, a common outcome in highly imbalanced classification tasks, the ensemble's relative improvement reflects better balance between precision and recall. Furthermore, the proposed voting ensemble maintains a competitive sensitivity (0.373 ± 0.023), slightly

lower than LR (0.391 ± 0.054) but considerably higher than KNN (0.123 ± 0.035). Although GaussianNB achieves a much higher sensitivity (0.910 ± 0.028), it does so at the cost of drastically reduced specificity (0.291 ± 0.006) and accuracy (0.301 ± 0.006). Importantly, the proposed ensemble attains a relatively high specificity (0.915 ± 0.006) and accuracy (0.907 ± 0.006), demonstrating its robustness and reliability in correctly identifying negative instances without incurring a major trade-off in positive case detection. Overall, these results confirm that the proposed voting ensemble offers a significant performance gain over conventional single classifiers, achieving a more favorable balance between sensitivity and specificity and improving overall discriminative power. This reinforces the value of combining diverse learners to exploit their complementary strengths, leading to enhanced predictive performance in highly imbalanced default prediction tasks.

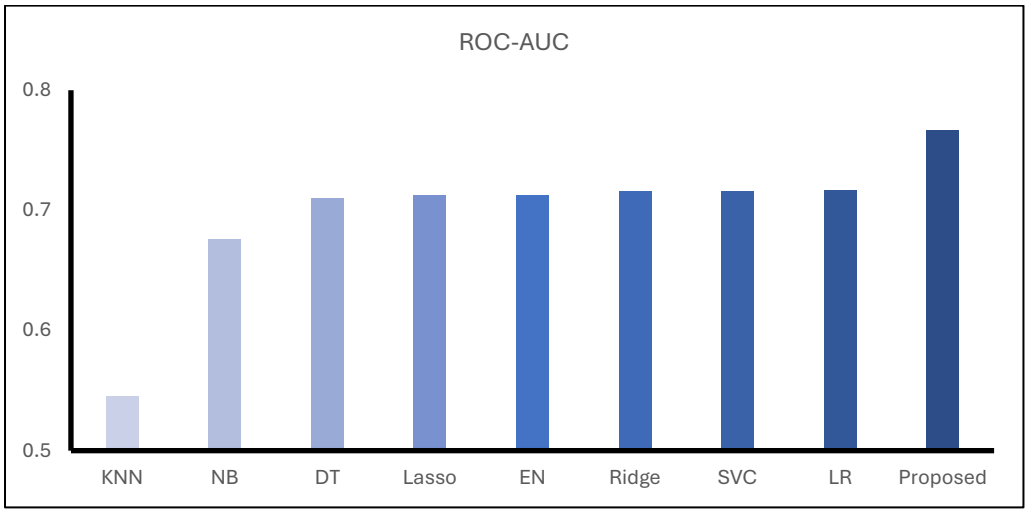

Fig. 4. Bar plots illustrating the ROC-AUC performance of the proposed ensemble compared to statistical and individual ML models.

4.2.3. Comparative Evaluation of the Proposed Ensemble and Various Ensemble Methods

Table 6 presents the comparative performance of the proposed voting ensemble against homogeneous bagging and boosting approaches, as well as heterogeneous ensembles (stacking and weighted voting). Performance was assessed across ROC-AUC, F1 score, sensitivity, specificity, and accuracy. As shown in Fig. 5, the proposed voting ensemble achieved the best overall balance between discrimination and classification performance, obtaining the highest ROC-AUC (0.767 ± 0.016), along with stable F1 score (0.111 ± 0.008) and sensitivity (0.373 ± 0.023). This indicates its ability to leverage the complementary strengths of diverse base learners while avoiding overfitting. In comparison, Weighted Voting (WV) delivered a slightly lower ROC-AUC (0.765 ± 0.017) and nearly identical accuracy (0.907 ± 0.003), confirming the robustness of the majority voting scheme in aggregating model predictions.

Table 6
Cross-validation results of the proposed ensemble compared with alternative ensemble methods.

Model	ROC-AUC	F1 score	Sensitivity	Specificity	Accuracy
Stacking	0.683 ± 0.018	0.052 ± 0.004	0.630 ± 0.047	0.639 ± 0.019	0.639 ± 0.018
Bagging	0.685 ± 0.032	0.087 ± 0.024	0.064 ± 0.021	0.994 ± 0.002	0.980 ± 0.001
ET	0.726 ± 0.015	0.076 ± 0.027	0.083 ± 0.030	0.982 ± 0.003	0.968 ± 0.003
CAT	0.729 ± 0.028	0.112 ± 0.020	0.255 ± 0.051	0.948 ± 0.003	0.937 ± 0.003
XGB	0.733 ± 0.027	0.112 ± 0.025	0.191 ± 0.048	0.965 ± 0.003	0.953 ± 0.003
ADA	0.737 ± 0.023	0.084 ± 0.014	0.387 ± 0.065	0.876 ± 0.006	0.868 ± 0.006
WV Ensemble	0.765 ± 0.017	0.109 ± 0.008	0.367 ± 0.029	0.915 ± 0.003	0.907 ± 0.003
Proposed Ensemble	0.767 ± 0.016	0.111 ± 0.008	0.373 ± 0.023	0.915 ± 0.006	0.907 ± 0.006

Among the boosting methods, AdaBoost (ADA) demonstrated competitive sensitivity (0.387 ± 0.065), but this came at the expense of specificity (0.876 ± 0.006) and accuracy (0.868 ± 0.006), highlighting its bias towards detecting positives. Gradient boosting variants (XGB and CAT) provided balanced performance, with ROC-AUC scores of 0.733 and 0.729, respectively, though they underperformed compared to the proposed voting ensemble. For homogeneous ensembles, Bagging showed extremely high specificity (0.994 ± 0.002) and accuracy (0.980 ± 0.001), but this was coupled with severely reduced sensitivity (0.064 ± 0.021), suggesting it is ill-suited for imbalanced data contexts where recall is critical. Extremely randomized trees (ET) showed slightly better balance (ROC-AUC = 0.726 ± 0.015), though still inferior to boosting- and voting-based approaches.

Finally, Stacking, despite being a heterogeneous ensemble of the same base learners, underperformed with an ROC-AUC of 0.683 ± 0.018 and poor F1 score (0.052 ± 0.004), indicating that the RF meta-learner was not able to exploit the diversity of base models effectively. Overall, the results establish the proposed voting ensemble as the most effective strategy, achieving the best discrimination capability (ROC-AUC), while maintaining strong accuracy and a superior trade-off between sensitivity and specificity compared to both homogeneous and other heterogeneous ensemble models.

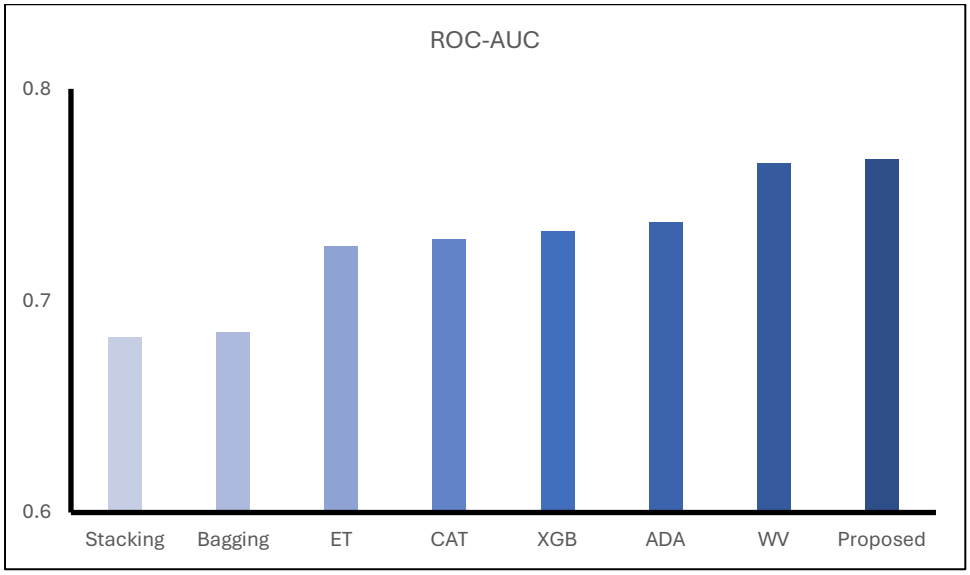

Fig. 5. Bar plots illustrating the ROC-AUC performance of the proposed ensemble compared to alternative ensemble approaches.

4.2.4. Comparative Evaluation of the Proposed Ensemble and a Prior Benchmark

To evaluate the effectiveness of the proposed voting ensemble, its performance was benchmarked against the best-performing model reported by Corrales et al. (2024), namely a neural network classifier. A key distinction between the two approaches lies in the input features: while the neural network leveraged macroeconomic indicators, our proposed model is based exclusively on transaction-level features available through the Kriya platform. Table 7 summarizes the comparative results across several evaluation metrics. As shown, the proposed voting ensemble consistently outperforms the neural network across all metrics, with the most notable improvements observed in AUC (0.767 vs. 0.717) and overall accuracy (0.907 vs. 0.855). Sensitivity also shows a modest increase (0.373 vs. 0.358), while specificity improves substantially (0.915 vs. 0.860). The F1 score, although relatively low in both models due to class imbalance, nearly doubles in the proposed ensemble (0.111 vs. 0.053), reflecting a better balance between precision and recall.

Table 7. Cross-validation results of the proposed ensemble compared with the model of Corrales et al. (Corrales et al., 2024).

	ROC-AUC	F1 score	Sensitivity	Specifity	Accuracy
Neuralnet	0.717	0.053	0.358	0.860	0.855
Proposed Ensemble	0.767	0.111	0.373	0.915	0.907

Although both models exhibit relatively low sensitivity and F1 scores, this outcome is expected given the substantial class imbalance in the dataset, where default events represent a minority of cases. Rather than indicating a model deficiency, these values reflect the inherent difficulty of predicting rare events in credit risk assessment. Importantly, the proposed voting ensemble consistently achieves higher recall and F1 score than the neural network baseline, demonstrating its ability to capture more true default cases while maintaining high performance on overall accuracy, specificity, and AUC. This suggests that transaction-level features provide meaningful discriminatory power, even under challenging conditions of data imbalance.

4.2.5. Feature Importance Analysis

To uncover the main determinants of P2B credit risk, a RF model was applied to assess feature importance. Figure 6 displays the top 12 features ranked by their relative contributions. The results can be grouped into three broad categories: loan structure, yield and exposure metrics, and trade structure characteristics.

1- Loan structure variables.

The analysis shows that MaturityDays (0.168) and AdvanceRate (0.154) are among the most influential predictors. Longer maturities increase uncertainty and default likelihood, while higher advance rates raise the lender's risk exposure. These findings highlight the centrality of loan design in shaping repayment performance.

2- Yield and exposure-related factors.

AnnualisedGrossYield (0.165), DiscountRate (0.137), and UnadjustedTotalGrossAdvance (0.093) emerge as critical drivers. A higher gross yield reflects risk-adjusted pricing, while the discount rate and total advances capture the financial burden and exposure scale. Additional exposure indicators, such as TotalFaceValue (0.072) and TotalGrossAdvance (0.067), further reinforce the role of transaction size in determining credit risk.

3- Trade structure characteristics.

Features such as TradeType_WholeLedger (0.073), TradeBand_1 (0.031), TradeType_Standard (0.021), TradeType_MultiDebtor (0.013), and TradeBand_5 (0.007) have lower relative importance but still contribute to differentiating risk profiles. These variables capture the nature of trade agreements and debtor composition, which may influence repayment stability in specific contexts. Collectively, the feature importance analysis confirms that loan maturity, yield, and exposure-related measures dominate the risk assessment process, while trade structure factors provide complementary but secondary insights. This emphasizes the primacy of financial contract design and exposure management in predicting P2B credit risk.

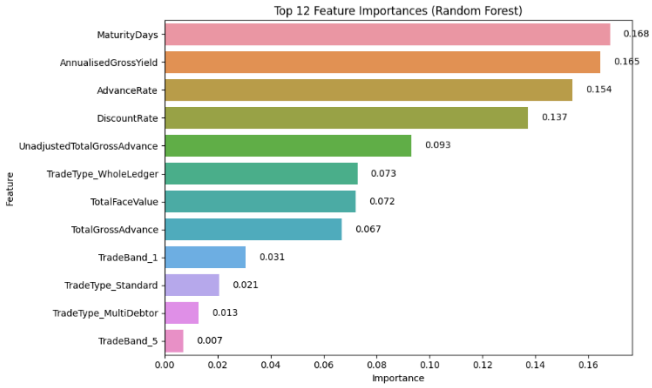

Fig. 6. Bar plot displaying the 12 most important predictor variables in the RF model for credit risk assessment.

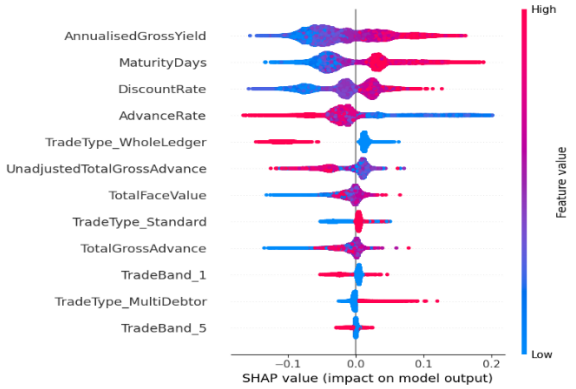

Fig. 7. SHAP global interpretability plot showing the directional effects of the 12 most important predictor variables in the RF model for credit risk assessment.

4.2.6. SHAP Analysis

To complement the feature importance analysis, SHAP values were computed using the Random Forest (RF) model to capture both the direction and magnitude of feature contributions to P2B credit risk (Figure 7). Leveraging the TreeSHAP algorithm, which is specifically optimized for tree-based models, the RF framework enables efficient and exact computation of Shapley values. This approach provides consistent, locally accurate, and computationally efficient feature attributions by exploiting the hierarchical structure of decision trees, thereby enhancing the interpretability and reliability of the model's predictions (Lundberg & Lee, 2018; Yang, 2021). The SHAP summary plot confirms that AnnualisedGrossYield, MaturityDays, DiscountRate, and AdvanceRate are the most influential features. Higher AnnualisedGrossYield and MaturityDays values are associated with increased risk, reflecting that loans priced with higher yields and longer maturities tend to involve riskier borrower profiles. Similarly, higher DiscountRate values elevate risk predictions, highlighting the burden of financing costs on borrower sustainability. In contrast, AdvanceRate exhibits a negative association with credit risk. Blue points on the right of the SHAP distribution show that higher advance rates reduce risk predictions, suggesting that such loans are selectively granted to more creditworthy borrowers. Finally, exposure-related variables (UnadjustedTotalGrossAdvance, TotalFaceValue, and TotalGrossAdvance) show mixed effects, while trade structure features (TradeType_WholeLedger, TradeType_Standard, TradeType_MultiDebtor, and TradeBands) contribute modestly, acting as secondary modifiers. Overall, the SHAP analysis validates the dominance of loan terms, yield, and financing costs as primary drivers of credit risk, while also revealing the protective effect of higher advance rates—a nuance not captured by feature importance alone

4.2.7. Partial Dependence Analysis

Partial Dependence Plots (PDPs) were generated for the four most important features, AnnualisedGrossYield, MaturityDays, DiscountRate, and AdvanceRate, to illustrate their marginal effects on predicted credit risk (Fig. 8). The plots show a nonlinear upward trend for AnnualisedGrossYield, with risk increasing steadily beyond yield levels of ~0.10. For MaturityDays, the effect is monotonic, with risk rising sharply after ~60 days and peaking for the longest maturities. DiscountRate displays sharp transitions: very low discount rates correspond to lower risk, but increases beyond ~0.006 lead to a substantial rise in predicted risk.

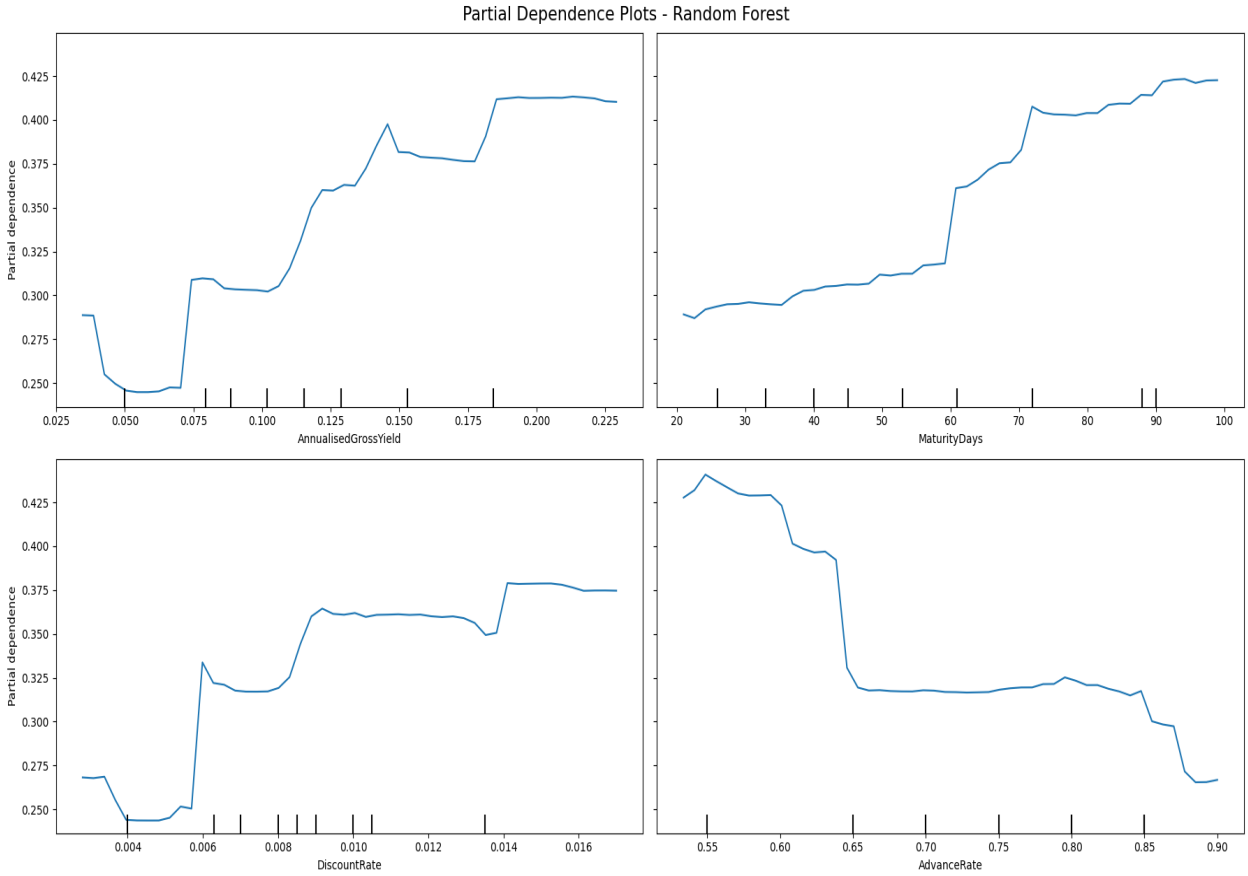

Fig. 8. Partial dependence plots for the four most important predictive variables.

By contrast, AdvanceRate demonstrates a negative relationship with credit risk. Predicted risk is elevated at lower advance rates (<0.65), but declines significantly as the advance rate increases, reinforcing the evidence from SHAP that higher advance rates are selectively associated with lower-risk borrowers. Overall, the PDPs confirm that loan yield, maturity, and financing costs drive higher risk, while higher advance rates mitigate it, offering important insights into how loan terms shape credit outcomes in P2B lending.

The interpretability results align closely with established financial theory and practical lending practices in credit markets. The positive association of yield, maturity, and discount rates with credit risk reflects standard risk–return trade-offs, where higher expected returns and longer repayment horizons typically signal elevated uncertainty and default likelihood. These findings are consistent with real-world credit pricing, where lenders compensate for risk through higher yields and fees. Conversely, the negative relationship between advance rate and credit risk suggests a selection effect: higher advance rates are generally offered only to borrowers deemed more creditworthy, supported by thorough due diligence and collateral assessment. This echoes evidence from trade finance and factoring markets, where financing terms are contingent on borrower reliability and portfolio quality. The relatively minor role of trade structure features indicates that while contractual form matters, it is secondary to core financial terms and exposure metrics in driving credit risk (Corrales et al., 2024; Long et al., 2022; Stiglitz & Weiss, 1981).

4.3. Discussion of Results

The empirical findings highlight the effectiveness of the proposed voting ensemble in predicting defaults in the P2B invoice trading context. Compared with the neural network benchmark reported by Corrales et al. (2024), which achieved an ROC-AUC of 0.717 and accuracy of 0.855 on the Kriya dataset, the ensemble model attained an ROC-AUC of 0.767 and accuracy of 0.907. These improvements are particularly notable given that the model relies exclusively on transaction-level features disclosed on the Kriya platform, whereas Corrales et al. incorporated macroeconomic indicators such as GDP and exports to improve explanatory power. This finding suggests that invoice-level attributes, such as advance rate, yield, and maturity, are the most influential predictors of firm creditworthiness in short-term financing arrangements.

Specifically, the SHAP analysis confirmed that loan-level characteristics exert strong influence on default probability. Advance rate, annualized gross yield, and maturity emerged as the most important predictors, consistent with the broader literature emphasizing the role of borrower- and loan-specific features. Serrano-Cinca et al. (2015) and Xu et al. (2021) similarly identified financial variables such as interest rate (analogous to yield), term length, and loan purpose as key

determinants of default. The positive relationship between annualized gross yield and default risk supports the adverse selection hypothesis, indicating that borrowers willing to accept higher borrowing costs are often riskier. Conversely, the negative association between advance rate and default aligns with the findings of Dorfleitner et al. (2017), who showed that platforms offering higher advances tend to apply stricter borrower screening, thereby reducing default risk. The positive effect of loan maturity corroborates earlier studies (Corrales et al., 2024)), which demonstrated that longer repayment horizons increase exposure to macroeconomic fluctuations, thereby elevating default risk.

Interestingly, SHAP also highlighted the role of trade type as a relevant predictor, with Whole Ledger trades associated with lower default risk. Corrales et al. (2024) also found trade type to be significant, particularly in standard and multi-debtor transactions, which exhibited a strong positive relationship with defaults. In the present analysis, Trade Band showed significance only for Bands 1 and 5, with mixed effects, whereas Corrales et al. reported sectoral variation in default risk, especially elevated in Trade Band 10. These differences may reflect methodological choices: Corrales et al. combined macroeconomic variables with invoice features using logistic regression, while the present approach isolates the predictive power of transaction-level data and applies RF with SHAP for feature importance and model interpretation.

From a methodological perspective, the ensemble's performance demonstrates the benefits of combining complementary learners in highly imbalanced settings. Prior studies have extensively explored regression-based models (Avgeri & Psillaki, 2023) and probit analysis (Nigmonov et al., 2022) ; however, they have often struggled with class imbalance. Ensemble methods, however, have demonstrated notable advantages: Aleksandrova & Parusheva (2021) showed that ensembles can reduce default ratios while improving portfolio allocation. The findings here are consistent with this evidence, as the proposed voting ensemble achieved statistically significant gains in ROC-AUC over its base learners ($p < 0.001$), indicating its robustness.

Nevertheless, the relatively low recall and F1 scores underline the persistent challenge of rare-event detection in P2B lending, consistent with prior work (Corrales et al., 2024). Corrales et al. also encountered limited recall despite applying sampling strategies. The ensemble improves upon their baseline, nearly doubling the F1 score, but absolute gains remain constrained by the structural imbalance of the dataset (1.7% defaults). This finding suggests avenues for further work, such as cost-sensitive learning and advanced resampling techniques, to enhance minority class detection without sacrificing specificity. Taken together, the findings illustrate the value of employing a heterogeneous ensemble model that leverages the strengths of diverse individual classifiers while mitigating their respective weaknesses. By integrating multiple learners, the ensemble is able to provide more robust and reliable predictive performance, even when relying on a relatively limited set of transaction-level variables. This underscores the practical utility of platform-generated data for credit risk assessment in P2B markets and further demonstrates how SHAP-based interpretability can offer meaningful insights into the relative contribution of invoice-level features to default prediction.

5. Conclusion

This study develops an explainable ensemble ML framework for credit risk assessment in P2B lending, addressing the challenges of information asymmetry, class imbalance, and limited transparency inherent in digital invoice trading platforms. By integrating RF, LGBM, and DL models within a heterogeneous soft voting ensemble, the proposed approach delivers superior predictive accuracy and interpretability compared with individual base learners as well as traditional statistical and state-of-the-art ML methods. The ensemble attains an ROC-AUC of 0.767, along with strong accuracy (0.907) and specificity (0.915), substantially improving on prior benchmarks while relying solely on transaction-level data. SHAP-based analysis further enhances model transparency, revealing that loan maturity, yield, discount rate, and advance rate are the dominant drivers of credit risk.

From an accounting analytics perspective, these findings demonstrate how explainable ensemble learning can provide decision-relevant insights for investors, platform managers, and regulators. The model supports more consistent and transparent credit evaluations, contributing to risk-informed investment strategies and promoting financial inclusion among small and medium-sized enterprises. Its explainability facilitates compliance with emerging regulatory expectations for algorithmic transparency in FinTech credit scoring. However, some limitations should be acknowledged. The model was developed and evaluated using data from a single UK-based P2B lending platform, Kriya. Future research should explore cross-platform validation to assess model robustness, integrate macro-financial indicators and textual disclosures to capture broader risk factors. Future research could also explore cost-sensitive learning and advanced resampling techniques to address the issue of extreme class imbalance.

Funding

The author extends her appreciation to Prince Sattam bin Abdulaziz University, Saudi Arabia for funding this research work through the project number (PSAU/2025/02/32836).

Conflict of Interest

The author has no relevant financial or non-financial interests to disclose.

Data Availability

All data sources are provided in the paper along with the relevant references.

References

Aleksandrova, Y., & Parusheva, S. (2021). Optimizing financial results for credit risk prediction in peer to peer lending platforms using machine learning. *2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS)*, 369–374.

Ali, G. M. (2026a). Digital Asset Analytics for DeFi Protocol Valuation: An Explainable Optuna-Tuned Super Learner Ensemble Framework. *Journal of Risk and Financial Management*, *19*(1), 63.

Ali, G. M. (2026b). Enhancing project financial performance prediction: An explainable machine learning framework integrating frontier efficiency and super learner. *Journal of Project Management (Canada)*, *11*(1), 151–168. https://doi.org/10.5267/j.jpm.2025.10.003

Ali, G. M. (2026c). Estimating project cost of equity using explainable ensemble learning: An empirical assessment of annual report readability. *Journal of Project Management (Canada)*, *11*(2), 541–560. https://doi.org/10.5267/J.JPM.2025.12.006

Ali, G. M. (2026d). Firm valuation using accounting-based capital structure and cash holdings: An explainable machine learning approach. *International Journal of Data and Network Science*, *10*(2), 577–596. https://doi.org/10.5267/j.ijdns.2026.2.001

Ali, G. M., Gomaa, I. I., & Darwish, S. M. (2022). An Intelligent Model for Success Prediction of Initial Coin Offerings. *IEEE Access*, *10*, 58589–58602. https://doi.org/10.1109/ACCESS.2022.3178369

Ali, G., & Yahia, Z. (2025). Disclosure Determinants of Blockchain Crowdfunding Performance for Sustainable Smart City Financing: An Explainable Optuna-Optimized Machine Learning Approach. *Journal of Posthumanism*, *5*(7), 802–826. https://doi.org/10.63332/joph.v5i7.2843

Altman, D. G., & Bland, J. M. (1994). Diagnostic tests. 1: Sensitivity and specificity. *BMJ : British Medical Journal*, *308*(6943), 1552. https://doi.org/10.1136/BMJ.308.6943.1552

Appiah, T., & Agblewornu, V. V. (2025). The interplay of perceived benefit, perceived risk, and trust in Fintech adoption: Insights from Sub-Saharan Africa. *Heliyon*, *11*(2), e41992. https://doi.org/10.1016/J.HELIYON.2025.E41992

Avgeri, E., & Psillaki, M. (2023). Factors determining default in P2P lending. *Journal of Economic Studies*, *51*(4), 823–840.

Avgeri, E., Psillaki, M., & Zervoudi, E. (2023). Peer-to-Peer Lending as a Determinant of Federal Housing Administration-Insured Mortgages to Meet Sustainable Development Goals. *Sustainability 2023, Vol. 15, Page 13618*, *15*(18), 13618. https://doi.org/10.3390/SU151813618

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, *6*(1), 20–29. https://doi.org/10.1145/1007730.1007735

Bogaert, M., & Delaere, L. (2023). Ensemble Methods in Customer Churn Prediction: A Comparative Analysis of the State-of-the-Art. *Mathematics 2023, Vol. 11, Page 1137*, *11*(5), 1137. https://doi.org/10.3390/MATH11051137

Byanjankar, A. (2017). Predicting credit risk in Peer-to-Peer lending with survival analysis. *2017 IEEE Symposium Series on Computational Intelligence, SSCI 2017 - Proceedings*, *2018-January*, 1–8. https://doi.org/10.1109/SSCI.2017.8280927

Carvalho, M., Pinho, A. J., & Brás, S. (2025). Resampling approaches to handle class imbalance: a review from a data perspective. *Journal of Big Data*, *12*(1), 1–58. https://doi.org/10.1186/S40537-025-01119-4/FIGURES/4

Chawla, N. V, Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chen, S. F., Charkaborty, G., Li, L. H., & Lin, C. T. (2019). Credit Risk Assessment Using Regression Model on P2P Lending. *International Journal of Applied Science and Engineering*, *16*(2), 149–157. https://doi.org/10.6703/IJASE.201909_16(2).149

Cho, P., Chang, W., & Song, J. W. (2019). Application of Instance-Based Entropy Fuzzy Support Vector Machine in Peer-To-Peer Lending Investment Decision. *IEEE Access*, *7*, 16925–16939. https://doi.org/10.1109/ACCESS.2019.2896474

Chulawate, N., & Kiattisin, S. (2023). Success factors influencing peer-to-peer lending to support financial innovation. *Sustainability*, *15*(5), 4028.

Corrales, C. M., González, L. A. O., & Santomil, P. D. (2024). Estimation of default and pricing for invoice trading (P2B) on crowdlending platforms. *Financial Innovation*, *10*(1), 109.

Donner, A., & Klar, N. (1996). The statistical analysis of kappa statistics in multiple samples. *Journal of Clinical Epidemiology*, *49*(9), 1053–1058. https://doi.org/10.1016/0895-4356(96)00057-1

Dorfleitner, G., Rad, J., & Weber, M. (2017). Pricing in the online invoice trading market: First empirical evidence. *Economics Letters*, *161*, 56–61.

Gao, X., Xie, D., Zhang, Y., Wang, Z., Chen, C., He, C., Yin, H., & Zhang, W. (2025). *A Comprehensive Survey on Imbalanced Data Learning*. https://doi.org/10.1007/s11704-025-50274-7

Gomes Mantovani, R., Horváth, T., Rossi, A. L. D., Cerri, R., Barbon Junior, S., Vanschoren, J., & Carvalho, A. C. de. (2024). Better trees: an empirical study on hyperparameter tuning of classification decision tree induction algorithms: R. Gomes Mantovani et al. *Data Mining and Knowledge Discovery*, *38*(3), 1364–1416.

Guijarro, F., Long Tran, K., Anh Le, H., Hien Nguyen, T., Trung Nguyen, D., Thai Binh Ward, N., & Chi Minh City, H. (2022). Explainable Machine Learning for Financial Distress Prediction: Evidence from Vietnam. *Data 2022, Vol. 7, Page 160*, *7*(11), 160. https://doi.org/10.3390/DATA7110160

Guo, Y., Zhou, W., Luo, C., Liu, C., & Xiong, H. (2016). Instance-based credit risk assessment for investment decisions in P2P lending. *European Journal of Operational Research*, *249*(2), 417–426.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Huang, H. (2024). Technology-Driven Financial Risk Management: Exploring the Benefits of Machine Learning for Non-Profit Organizations. *Systems 2024, Vol. 12, Page 416*, *12*(10), 416. https://doi.org/10.3390/SYSTEMS12100416

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, Number 1). Springer.

Kgoroeadira, R., Burke, A., Di Pietro, F., & van Stel, A. (2023). Determinants of firms' default on unsecured loans in the P2P crowdfunding market. *Journal of International Financial Markets, Institutions and Money*, *89*, 101882.

Kuang, J., Ji, X., Cheng, P., & Kallinterakis, V. B. (2023). Media News and Social Media Information in the Chinese Peer-to-Peer Lending Market. *Systems 2023, Vol. 11, Page 133*, *11*(3), 133. https://doi.org/10.3390/SYSTEMS11030133

Kunapuli, G. (2023). *Ensemble Methods for Machine Learning*. Simon and Schuster.

Kwak, S. K., & Kim, J. H. (2017). Statistical data preparation: management of missing values and outliers. *Korean Journal of Anesthesiology*, *70*(4), 407. https://doi.org/10.4097/KJAE.2017.70.4.407

Lin, X., Li, X., & Zheng, Z. (2017). Evaluating borrower's default risk in peer-to-peer lending: evidence from a lending platform in China. *Applied Economics*, *49*(35), 3538–3545.

Liu, Y., Baals, L. J., Osterrieder, J., & Hadji-Misheva, B. (2024). Leveraging network topology for credit risk assessment in P2P lending: A comparative study under the lens of machine learning. *Expert Systems with Applications*, *252*, 124100.

Long, J., Jiang, C., Dimitrov, S., & Wang, Z. (2022). Clues from networks: quantifying relational risk for credit risk evaluation of SMEs. *Financial Innovation*, *8*(1), 1–41. https://doi.org/10.1186/S40854-022-00390-1/TABLES/38

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Lundberg, S. M., & Lee, S.-I. (2018). *Consistent feature attribution for tree ensembles*. http://arxiv.org/abs/1706.06060

Machado, M. R., Karray, S., & De Sousa, I. T. (2019). LightGBM: An effective decision tree gradient boosting method to predict customer loyalty in the finance industry. *14th International Conference on Computer Science and Education, ICCSE 2019*, 1111–1116. https://doi.org/10.1109/ICCSE.2019.8845529

Maiolini, R., Cappa, F., Franco, S., & Quaratino, G. R. (2024). The impact of sustainable development goals in lending-based prosocial crowdfunding: A topic modeling analysis on the kiva platform. *International Review of Financial Analysis*, *95*, 103441.

Mienye, I. D., & Sun, Y. (2022). A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects. *IEEE Access*, *10*, 99129–99149. https://doi.org/10.1109/ACCESS.2022.3207287

Mienye, I. D., & Swart, T. G. (2024). A Comprehensive Review of Deep Learning: Architectures, Recent Advances, and Applications. *Information 2024, Vol. 15, Page 755*, *15*(12), 755. https://doi.org/10.3390/INFO15120755

Mohammed, A., & Kora, R. (2023). A comprehensive review on ensemble deep learning: Opportunities and challenges. *Journal of King Saud University - Computer and Information Sciences*, *35*(2), 757–774. https://doi.org/10.1016/J.JKSUCI.2023.01.014

Nigmonov, A., Shams, S., & Alam, K. (2022). Macroeconomic determinants of loan defaults: Evidence from the U.S. peer-to-peer lending market. *Research in International Business and Finance*, *59*, 101516. https://doi.org/10.1016/J.RIBAF.2021.101516

Nwafor, C. N., Nwafor, O., & Brahma, S. (2024). Enhancing transparency and fairness in automated credit decisions: an explainable novel hybrid machine learning approach. *Scientific Reports 2024 14:1*, *14*(1), 1–20. https://doi.org/10.1038/s41598-024-75026-8

Polena, M., & Regner, T. (2018). Determinants of borrowers' default in P2P lending under consideration of the loan risk class. *Games*, *9*(4), 82.

Porwik, P., Doroz, R., & Wrobel, K. (2019). An ensemble learning approach to lip-based biometric verification, with a dynamic selection of classifiers. *Expert Systems with Applications*, *115*, 673–683. https://doi.org/10.1016/J.ESWA.2018.08.037

Probst, P., Boulesteix, A.-L., & Bischl, B. (2019). Tunability: Importance of hyperparameters of machine learning algorithms. *Journal of Machine Learning Research*, *20*(53), 1–32.

Sánchez-Medina, A. J., Blázquez-Santana, F., Cerviño-Cortínez, D. L., & Pellejero, M. (2025). Ensemble Methods for Bankruptcy Resolution Prediction: A New Approach. *Computational Economics*, 1–36. https://doi.org/10.1007/S10614-024-10709-Y/TABLES/3

Sarıkaya, A., Günel Kılıç, B., & Demirci, M. (2022). GRU-GBM: A combined intrusion detection model using LightGBM and gated recurrent unit. *Expert Systems*, *39*(9), e13067. https://doi.org/10.1111/EXSY.13067

Sayari, S., Mgadmi, N., Dhaou, I. Ben, Almehdar, M., Chishty, S. K., & Rabeh, A. (2025). Advancing Sustainable Development Through Digital Transformation and Fintech Innovation. *Sustainability 2025, Vol. 17, Page 4924*, *17*(11), 4924. https://doi.org/10.3390/SU17114924

Serrano-Cinca, C., Gutiérrez-Nieto, B., & López-Palacios, L. (2015). Determinants of default in P2P lending. *PloS One*, *10*(10), e0139427.

Shan, W., & Gao, B. (2025). Stacked Ensemble Model with Enhanced TabNet for SME Supply Chain Financial Risk Prediction. *Systems 2025, Vol. 13, Page 892*, *13*(10), 892. https://doi.org/10.3390/SYSTEMS13100892

Stiglitz, J. E., & Weiss, A. (1981). Credit Rationing in Markets with Rationing Credit Information Imperfect. *The American Economic Review*, *71*(3), 393–410. https://doi.org/10.2307/1802787

Teichmann, F. M. J., Boticiu, S. R., & Sergi, B. S. (2024). Compliance concerns in sustainable finance: an analysis of peer-to-peer (P2P) lending platforms and sustainability. *Journal of Financial Crime*, *31*(4), 993–1001.

Thabet, H. H., Darwish, S. M., & Ali, G. M. (2024). Measuring the efficiency of banks using high-performance ensemble technique. *Neural Computing and Applications*, 1–19. https://doi.org/10.1007/S00521-024-09929-Y/TABLES/5

Thai, H. T. (2022). Machine learning for structural engineering: A state-of-the-art review. *Structures*, *38*, 448–491. https://doi.org/10.1016/J.ISTRUC.2022.02.003

Trimulato, T., Nafis, M. C., & Amalia, E. (2022). The Role Sharia Fintech Support Sustanaible Development Goals Program (SDGs). *Jurnal Ilmiah Ekonomi Islam*, *8*(1), 251–259.

Wang, Z., Jiang, C., Ding, Y., Lyu, X., & Liu, Y. (2018). A novel behavioral scoring model for estimating probability of default over time in peer-to-peer lending. *Electronic Commerce Research and Applications*, *27*, 74–82.

Weng, Z., & Luo, P. (2021). The Impact of Guarantees on Peer-to-Peer Lending Platform: Evolutionary Game Analysis and Empirical Evidence from China. *Journal of Theoretical and Applied Electronic Commerce Research 2021, Vol. 16, Pages 2708-2731*, *16*(7), 2708–2731. https://doi.org/10.3390/JTAER16070149

Xiao, J., Li, S., Tian, Y., Huang, J., Jiang, X., & Wang, S. (2025). Example dependent cost sensitive learning based selective deep ensemble model for customer credit scoring. *Scientific Reports 2025 15:1*, *15*(1), 1–24. https://doi.org/10.1038/s41598-025-89880-7

Xu, B., Su, Z., & Celler, J. (2021). Evaluating default risk and loan performance in UK peer-to-peer lending: evidence from funding circle. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, *25*(5), 530–538.

Yang, J. (2021). Fast treeshap: Accelerating shap value computation for trees. *ArXiv Preprint ArXiv:2109.09847*.

Yang, Y., Wu, Y., Wang, P., & Jiali, X. (2021). Stock Price Prediction Based on XGBoost and LightGBM. *E3S Web of Conferences*, *275*. https://doi.org/10.1051/E3SCONF/202127501040

Yuan, Z., Wang, Z., & Xu, H. (2018). Credit risk assessment of peer-to-peer lending borrower utilizing BP neural network. *Advances in Internetworking, Data & Web Technologies: The 5th International Conference on Emerging Internetworking, Data & Web Technologies (EIDWT-2017)*, 22–33.

Zhu, S., Wu, H., Ngai, E. W. T., Ren, J., He, D., Ma, T., & Li, Y. (2024). A Financial Fraud Prediction Framework Based on Stacking Ensemble Learning. *Systems, 12*(12), 588. https://doi.org/10.3390/SYSTEMS12120588

© 2026 by the authors; licensee Growing Science, Canada. This is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC, BY) license (http://creativecommons.org/licenses/by/4.0/).