Iranian Journal of Economic Studies, 14(1) 2025, 121-137

Shiraz University

Iranian Journal of Economic Studies

Journal homepage: ijes.shirazu.ac.ir

A Novel Multi-Algorithm Stacking Framework for Enhanced Credit Risk Management

Younes Nademi* [a], Sayyed Mohammad Hoseini[b], Majid Ebtia[c], Faranak Ahmadi[d]

a. Department of Economics, Ayatollah Boroujerdi, University, Boroujerd. Iran.
b. Gahar Artificial Intelligence Research Group, Ayatollah Boroujerdi University, Boroujerd, Iran.
c. Zagros Data Sciences Research Group, Ayatollah Boroujerdi University, Boroujerd, Iran.
d. Department of Computer Engineering,, Ayatollah Boroujerdi University, Boroujerd, Iran.

Article History

Received date: 04 September 2025
Revised date: 19 October 2025
Accepted date:27 October 2025
Available online: 07 November 2025

JEL Classification
C45
C53
G17
G32

Keyword
Credit Risk Prediction
Ensemble Learning
Stacking Ensemble
Class Imbalance
Logistic Regression

Abstract

Credit risk prediction remains a central challenge for financial institutions because inaccurate assessments can cause substantial financial losses and systemic instability. This study introduces a multi-level stacking ensemble that combines Gradient Boosting, Extreme Gradient Boosting (XGBoost), and Random Forest as base learners with logistic regression as the meta-learner. To address class imbalance, we do not use synthetic resampling; instead, we apply a class-management protocol based on fold-wise class-weighting, probability calibration, and operating-point tuning to ensure fair treatment of the minority (default) class without introducing synthetic examples. The approach was evaluated on two UCI benchmark datasets (German and Australian credit) using a fixed train/test split and stratified 10-fold cross-validation on the training set for model selection; final models were retrained on the full training set and assessed on a held-out test set. Results show the stacked ensemble consistently outperforms individual base learners on balanced metrics including F1 and Matthews Correlation Coefficient (MCC) while preserving interpretability via calibrated base-learner probabilities and inspectable logistic meta-coefficients. An empirical analysis of Principal Component Analysis (PCA) reveals dataset-dependent effects: PCA can benefit simpler classifiers but may reduce performance for interaction-sensitive ensembles. The paper provides a practical deployment blueprint covering class-management placement, probability calibration before meta-learning, and cost-aware evaluation tailored to credit-risk operations.

Highlights

- Novel multi-algorithm stacking framework for credit risk.
- Achieves superior performance without synthetic resampling.
- Logistic regression meta-learner ensures model interpretability.
- PCA effects are dataset-dependent; can harm complex ensembles.

* younesnademi@abru.ac.ir
DOI: 10.22099/ijes.2025.54179.2058

1. Introduction

Predicting credit risk was among the earliest practical uses of machine learning, leveraging borrowers' financial records to estimate the probability of default on loans, credit cards, and other credit products. Accurately forecasting credit risk remains a major challenge for financial institutions, spurring extensive research aimed at improving predictive methods and decision outcome (Moradi & Mokhatab, 2019). Early applications of machine learning focused on credit-risk prediction, using financial data to assess the likelihood that customers will default on loans, credit cards, and similar obligations. Reliable credit-risk forecasting continues to be difficult for banks and lenders, motivating a large body of research to enhance model accuracy and reduce financial losses (Rehman et al., 2019). Effective use of credit risk prediction tools can significantly enhance the profitability of financial institutions. This is particularly pertinent for credit card and loan applications. Financial institutions that fail to accurately predict credit risk have faced substantial losses, underscoring the critical importance of precise risk assessment for their survival (Khemakhem & Boujelbene, 2018). Over the past few decades, credit risk prediction has been a hot topic, with credit card default prediction being one of the most crucial tasks for creditors. This is due to the higher number of default transactions compared to non-default transactions (Dornadula & Geetha, 2019).Consequently, the datasets used for credit risk prediction often suffer from class imbalance issues. Previous studies have indicated that class imbalance can degrade the performance of machine learning (ML) models, leading to bias towards a particular class during inference (García etal., 2012). Various techniques have been proposed in the literature to address the class imbalance problem, categorized into three main groups: ensemble learning, cost-sensitive learning, and re-sampling methods. Among these, ensemble learning has been extensively studied (Song & Peng, 2019).Ensemble learners outperform single models by leveraging the strengths of multiple base learners. Furthermore, ensemble models are divided into two types: classifier ensembles and hybrid classifiers. The former integrates attribute selection techniques or hyperparameter tuning prior to classification, while the latter combines multiple classifiers operating in parallel (Guo etal., 2019).

In this study, we develop a multilevel ensemble-based model that builds on the proven advantages of modern stacking ensembles. By harnessing the complementary strengths of Gradient Boosting (Friedman, 2001), Extreme Gradient Boosting (Chen & Guestrin, 2015), and Random Forest (Breiman, 2001), our framework achieves markedly higher predictive accuracy than any individual algorithm. Stacking—or stacked generalization—feeds the probabilistic outputs of each base learner into a single meta-learner (Wolpert, 1992), which intelligently weighs and blends these signals to capture intricate non-linear patterns in the data. This layered architecture disperses error sources across diverse models, curbing overfitting and yielding robust generalization. Through rigorous cross-validation, we thoroughly vet the ensemble's stability on unseen samples, ultimately delivering a resilient predictive framework that

balances bias and variance more effectively than traditional single-model approaches.

2. Literature Review

Credit risk prediction seeks to identify borrowers likely to default, but the task is challenged by highly imbalanced data (few defaulters vs. many non-defaulters). Machine learning (ML) methods, especially tree-based ensembles, have become widespread. (Noriega et al., 2023), note that boosting models (e.g. gradient boosting) dominate recent credit-scoring research, with most studies using metrics like AUC, accuracy, and F1. However, they also highlight persistent challenges: "the black box nature" of complex models, the need for explainability, and the imbalance in input data. In practice, class imbalance tends to bias models toward the majority (safe) class, degrading minority-class (default) recall. For example, (La Gatta et al., 2025), explicitly state that data imbalance "penalizes predictive performance," since learning to classify the few "bad" loans is hard when they are underrepresented. To address this, many studies apply resampling (e.g. SMOTE, ADASYN) or cost-sensitive learning. (Aruleba & Sun, 2025), emphasize this point: they show that combining SMOTE-ENN resampling with a stacked ensemble significantly improves sensitivity and specificity in credit data. Thus, recent literature makes clear that both ensemble methods and imbalance-handling are key to state-of-the-art credit scoring.

Ensemble methods (bagging, boosting, stacking) leverage multiple models to boost predictive power. Empirical studies consistently find ensembles outperform single classifiers in credit risk tasks. For instance, (Han et al., 2023), report that ensemble approaches "have been validated to be more competitive than individual classifiers" for default prediction. Bagging methods like Random Forest reduce variance, while boosting methods like Gradient Boosting or XGBoost reduce bias. (Liu et al., 2024), demonstrate this benefit via novel feature engineering: they generate tree-ensemble features (bagging- and boosting-based) and find the boosting-based features yield markedly better credit scoring accuracy, AUC and F1 than the bagging-based features or individual classifiers. This underscores that cleverly combining multiple trees (via boosting) captures complex non-linear patterns better than simpler models.

Stacking (stacked generalization) takes this further by training a meta-learner on the outputs of base models. In credit risk, stacking has shown strong results. For example, (Liu et al., 2024), propose ensemble tree-based feature transformations fed into logistic regression as a meta-learner, and report substantial improvements in accuracy over single models. More broadly, multi-layer stacking architectures have been introduced: Han et al.'s multi-layer multi-view stacking (MLMVS) model for P2P credit risk combined probabilistic outputs from several base classifiers across "views," and was experimentally shown to outperform standard ensembles and single classifiers. Similarly, (Wei et al., 2023), apply a stacking ensemble on a large P2P loan dataset and find it achieves higher accuracy, precision and recall than any base learner (with lowest

error) – demonstrating that stacking yields "accurate and stable predictions". Overall, recent work suggests that stacking ensembles (especially with interpretable meta-models) can achieve superior predictive performance and robustness by blending diverse learners.

Credit datasets are notoriously skewed: defaulters are rare. To mitigate this, many studies integrate resampling with ensembles. Oversampling techniques like SMOTE or ADASYN create synthetic minority examples, while undersampling removes excess majority cases. (La Gatta et al., 2025), find that for very large P2P data, random undersampling (RUS) actually outperformed SMOTE: they report that SMOTE "is not an appropriate method for this case," whereas undersampling yielded higher performance given the large sample size. Other studies combine oversampling with stacking: for instance, a SMOTE+stacking approach achieved 83.2% accuracy on a peer-to-peer lending dataset. Hybrid methods are also explored: (Aruleba & Sun, 2025), use a hybrid SMOTE-ENN resampling in a stacking framework, achieving ~0.92 sensitivity and specificity on several public credit datasets. These findings indicate that resampling remains crucial: without it, stacked models tend to be biased. The proposed multi-level stack aims to address imbalance by incorporating resampling (e.g. SMOTE variants) in the training pipeline, ensuring the meta-learner sees balanced inputs.

Comparative studies consistently show ensembles yield better generalization. Boosted trees (GB, XGBoost) often achieve top accuracy but can overfit without care. Bagging (RF) offers stability. Stacking adds another safeguard: by combining diverse models' predictions via a meta-learner, it can curb overfitting and bias and improve robustness on unseen data. For example, (Liu et al., 2024), found that their ensemble feature-transform+logistic approach gave consistently higher AUC and F1 than any individual model across multiple data splits. Likewise, Aruleba & Sun report that their stacking+resampling system markedly outperforms individual learners (RF, LR, CNN) on various benchmarks. Empirical results also highlight robustness: bagging in particular "lowers variance and increases the robustness of the model", which translates to more stable credit-risk estimates when data is noisy or imbalanced.

However, limitations remain. Many ensemble schemes are computationally intensive and may overfit if improperly tuned. Stacking models, while powerful, can be sensitive to the choice of meta-learner – nonlinear metas (like XGBoost) often add complexity, whereas linear metas (like logistic regression) sacrifice some modeling power for interpretability. Critically, many recent studies optimize accuracy but pay less attention to explainability and calibration. Noriega et al. emphasize that the black-box nature of complex ensembles is a barrier in finance. There is also a gap in class imbalance strategies: oversampling can introduce noise, and existing methods (SMOTE, ADASYN, ENN) have known drawbacks. For example, (La Gatta et al., 2025), show that oversampling may hurt performance on very large datasets, suggesting one-size-fits-all techniques are insufficient. Finally, feature selection is often manual or suboptimal, leading to models that may not generalize well across different credit portfolios.

The our proposed multi-level stacking ensemble – combining Gradient Boosting, XGBoost, and Random Forest as base learners, with logistic regression as meta-learner – targets these gaps. First, by using three strong but diverse tree-based learners, it leverages complementary strengths: boosted models capture subtle patterns (as (Liu et al., 2024) show boosting yields the best ensemble features), while Random Forest adds robustness. Feeding their probabilistic outputs into a logistic meta-learner offers several advantages. Logistic regression is inherently interpretable and less prone to overfitting than nonlinear metas, addressing concerns about black-box stacking. In fact, (Liu et al., 2024), employ logistic meta-learning on ensemble-derived features and report that this "synthetic feature transformation" method markedly improves credit scoring performance.

Second, our framework explicitly integrates imbalance handling. Unlike many prior studies that apply one resampling method uniformly, the multi-level approach allows resampling at different stages (e.g. before each base learner). This adaptivity is motivated by findings like La Gatta et al.'s, which suggest the best resampling strategy may vary with data size. By combining SMOTE variants with ensemble learning, the model is designed to ensure the meta-learner receives a balanced representation of defaulters, mitigating bias toward the majority class. The use of cross-validation at each stacking level further promotes generalization: error is "dispersed across diverse models" which curbs overfitting, yielding a more resilient predictor.

Overall, the proposed model is novel in its multi-layer stacking architecture and in employing logistic regression as the top-layer. It builds on evidence that stacking ensembles outperform standalone models, while specifically addressing the interpretability and imbalance issues noted in recent literature. In summary, by integrating gradient boosting, XGBoost, and random forest within a stacking framework with logistic meta-learning and tailored resampling, the proposed approach aims to achieve higher predictive accuracy and better generalization than prior methods – bridging gaps in feature engineering, imbalance handling, and robustness that have been highlighted by recent studies.

3. Classification algorithms

3.1 Naïve Bayes

Naïve Bayes (NB) delivers fast, well-calibrated probabilities and scales to large feature sets. Its conditional-independence assumption can benefit performance when features are decorrelated but limits interaction modeling and reduces effectiveness on datasets with strong feature dependencies (Han et al., 2022).

3.2 K-Nearest Neighbors

k-Nearest Neighbors (KNN) is a nonparametric local method that adapts to complex decision boundaries without explicit training. It is sensitive to class imbalance and to the distance metric, requires storing the training set, and

becomes computationally expensive at prediction time, which constrains its operational use in high-throughput credit pipelines (Han et al., 2022).

3.3 Logistic Regression

Logistic regression (LR) produces interpretable, monotonic probability estimates that are straightforward to calibrate and to integrate as a meta-learner. It cannot model complex nonlinear feature interactions unless combined with engineered features, but its low variance and transparency make it well suited for regulatory-facing ensemble layers (Han et al., 2022).

3.4 Decision Trees

Decision trees (DT) capture nonlinear interactions and produce intuitive decision rules that support explainability. Unconstrained trees overfit easily; pruning or depth limits are necessary to control variance. Trees provide natural handling of mixed feature types, which simplifies preprocessing for credit datasets (Han et al., 2022).

3.5 Random Forest

Random Forest (RF) aggregates many decorrelated trees to reduce variance and increase robustness to noise. It preserves interaction effects and is less sensitive to overfitting than single trees, but the ensemble's internal complexity reduces direct interpretability and increases inference cost compared with linear models (Kunapuli, 2023).

3.6 Gradient Boosting

Gradient Boosting (GB) learns additive sequential corrections that capture subtle, high-order interactions and reduce bias. It requires careful hyperparameter tuning and regularization to avoid overfitting, and its sequential nature increases training time compared with bagging methods (Kunapuli, 2023).

3.7 eXtreme Gradient Boosting

eXtreme Gradient Boosting (XGBoost) is an optimized gradient-boosting implementation that improves training speed and adds regularization and sparsity-aware split finding. It achieves strong predictive performance on tabular credit data but demands hyperparameter search to balance bias, variance, and calibration (Kunapuli, 2023).

3.8 Multi-Layer Perceptron

Multi-Layer Perceptron (MLP) models capture complex, non-linear relationships and high-order interactions given sufficient data and tuning. They are sensitive to class imbalance, require greater computational resources, and produce less interpretable outputs without auxiliary explanation tools (Han et al., 2022).

4. Our Proposed Method Stack

In this study, we propose a multilevel ensemble framework that employs stacking to combine three powerful machine learning algorithms: Gradient Boosting, Extreme Gradient Boosting, and Random Forest. Stacking, or stacked generalization, integrates the outputs of independently trained base learners and feeds them into a final estimator—referred to as a meta-learner—that learns how to best synthesize these predictions. In this architecture, base models operate in parallel, and the sequence in which they are introduced has no effect on the outcome, as the meta-learner automatically determines the optimal combination of their outputs. Logistic Regression is selected as the meta-learner due to its simplicity, interpretability, and strong performance in blending probabilistic inputs. The data undergoes a thorough preprocessing pipeline, including cleaning, scaling, and categorical encoding, followed by a standard training/testing split. Each base model is trained separately and generates predictions that are then passed to the meta-learner for final prediction. Model effectiveness is assessed using several evaluation criteria to ensure accuracy, robustness, and generalization. Additionally, cross-validation techniques are employed to verify the stability of the ensemble across unseen data. By capturing the complementary strengths of diverse algorithms in a unified framework, the proposed stacked model offers a more reliable and scalable solution for complex classification tasks.

Table 1. Description of datasets used in the experiment

Database	d	n
German Credit	24	1000
Australian Credit	14	690

Source: *The UCI Machine Learning Repository (https://archive.ics.uci.edu/)*

5. Experimentation and Result Analysis

The experimental phase of this study was carried out using two datasets sourced from the UCI Machine Learning Repository, a widely recognized resource in the machine learning community. Established in 1987 by David Aha and colleagues at the University of California, Irvine, the repository has since become a cornerstone for empirical machine learning research. It provides a diverse collection of curated and well-documented datasets, many of which are accompanied by thorough descriptions and preprocessing guidelines. These features make it a valuable benchmark for researchers, educators, and practitioners alike. The selected datasets for this study reflect real-world classification challenges and were chosen for their relevance, quality, and suitability for evaluating machine learning algorithms. Each dataset comprises multiple instances, described through various attributes, and enables a robust comparison of classification models. A summary of the datasets used is provided

in Table 1, while Table 2 presents the comparative performance of the applied machine learning methods in terms of classification accuracy.

Table 2. Accuracy of machine learning methods obtained from various methods on the datasets

Data set	Source	Machine Learning method	ACC
German credit	Emmanuel et al., (2024)	Stack(Xgboost,Random forest, Gradient Boosting)	0.8280
	Zou & Gao, (2022)	AugBoost-ELM	0.7617
	Quan & Sun, (2024)	FM	0.7696
	Wu et al., (2021)	DBM+DRBM	0.8858
	Veeramanikandan & Jeyakarthic, (2021)	SADNN	0.961
	Du & Shu, (2022)	BRNN	0.62
	Religia et al., (2020)	Random Forest	0.7833
	Alam et al., (2020)	Gradient Boosting	0.835
	Zhao & Aumeboonsuke, (2023)	XGBoost	0.8186
	Hoseini et al., (2024)	ensemble SVM(poly)	0.8050
	Bulut & Arslan, (2024)	PCA and CV(NB)	0.74
Australian credit	Emmanuel et al., (2024)	Stack(Xgboost,Random forest, Gradient Boosting)	0.8623
	Zou & Gao, (2022)	AugBoost-PCA	0.8681
	Quan & Sun, (2024)	FM	0.8844
	Du & Shu, (2022)	BRNN	0.82
	Hoseini et al., (2024)	Random forest	0.8768

Source: *mentioned in the source column within the Table*

5.1 Evaluation measures of model performance

Evaluating the effectiveness of a classification model requires the use of established performance metrics. In this study, the dataset is split into training and testing subsets, where the model is first trained on the training data and subsequently evaluated on the test set to assess its predictive ability. The performance of the proposed model is measured using several widely recognized

criteria, including Accuracy (ACC), F1-score, and the Matthews Correlation Coefficient (MCC). These metrics collectively offer a well-rounded view of the model's classification capabilities. Accuracy quantifies the ratio of correctly predicted instances to the total number of samples, serving as a basic indicator of overall model performance. The F1-score, which represents the harmonic mean of precision and recall, provides a balanced metric especially useful in the presence of imbalanced classes. Meanwhile, MCC delivers a more robust evaluation by incorporating all elements of the confusion matrix—true positives, true negatives, false positives, and false negatives—yielding a coefficient between -1 and 1, with values closer to 1 indicating highly reliable predictions. Through the combined use of these metrics, the study ensures a comprehensive and reliable assessment of the model's classification performance (Powers, 2011).

$$ACC = \frac{TP+TN}{TP+TN+FP+FN}, \tag{1}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}, \tag{2}$$

$$F_1 = 2\left(\frac{Precision \times Sensitivity}{Precision + Sensitivity}\right), \tag{3}$$

where

$$Sensitivity = \frac{TP}{TP+FN}, \quad Precision = \frac{TP}{TP+FP}. \tag{4}$$

6. Result Analysis

Model selection and hyperparameter tuning were performed using stratified 10-fold cross-validation on the training set: in each fold models were trained on 9 folds and validated on the remaining fold, and hyperparameters were chosen by grid or randomized search to maximize balanced metrics (primarily F1 and MCC) averaged across the 10 folds. After selecting the best configuration, each model was retrained on the full training set with those hyperparameters and evaluated once on the held-out test set. All procedures preserved class proportions within folds to prevent leakage and ensure that reported test results reflect genuine out-of-sample performance.

6.1 Analysis of Table 3(German Dataset)

The proposed stacked ensemble achieves the highest scores across all evaluated metrics on the German dataset, with recall of 0.6896, precision of 0.7296, MCC of 0.4174, F1-score of 0.7023, and accuracy of 0.7700. These results surpass the next best performer, Logistic Regression, which records slightly lower values (REC = 0.6849, PRE = 0.7192, MCC = 0.4027, F1 = 0.6962, ACC = 0.7633). Tree-based learners such as Gradient Boosting and Random Forest also deliver strong performance but fall short of the ensemble's balanced improvement across both sensitivity and specificity. Simpler classifiers like KNN,

Naive Bayes, and single decision trees exhibit notably lower correlation coefficients and F1-scores, indicating their limited ability to capture the dataset's complex patterns. Overall, the multilevel stacking framework more effectively harmonizes bias and variance, leading to a uniformly superior classification performance.

On the German dataset the stacked ensemble attains F1 = 0.7023 and MCC = 0.4174, which are the highest among the reported models. Comparing directly with the three standalone tree-based learners shows a consistent advantage in both balanced-performance metrics. Gradient Boosting records F1 = 0.6801 and MCC = 0.3938, XGBoost records F1 = 0.6619 and MCC = 0.3559, and Random Forest records F1 = 0.6874 and MCC = 0.3750. The stacked model's F1 improvement over the best single tree (Random Forest) is 0.0149 absolute, and its MCC improvement is 0.0424 absolute, indicating that stacking yields modest but meaningful gains in harmonic balance between precision and recall and in overall correlation between predictions and true labels. These gains reflect the meta-learner's ability to combine complementary probability estimates, improving discrimination of the minority class while maintaining specificity for the majority class.

6.2 Analysis of Table 4(Australian Dataset)

The proposed stacked ensemble again leads all contenders, attaining recall of 0.8684, precision of 0.8640, MCC of 0.7324, F1-score of 0.8641, and accuracy of 0.8647. This performance notably surpasses the closest standalone methods—Random Forest and decision trees—which both achieve accuracy of 0.8550 and MCC values around 0.715. Logistic Regression and Gradient Boosting deliver respectable results in the mid-0.84 accuracy range but fall short of the ensemble's gains in balanced classification (MCC) and F1-score. Simpler classifiers such as KNN, Naive Bayes, and the multilayer perceptron underperform relative to the tree-based learners, underscoring their limited capacity to capture the dataset's nuanced patterns. These findings reaffirm that integrating diverse base learners into a parallel stacking framework yields a more robust, generalized model for real-world credit classification tasks.

On the Australian dataset the stacked ensemble achieves F1 = 0.8641 and MCC = 0.7324, outperforming the individual tree learners. Random Forest reports F1 = 0.8546 and MCC = 0.7151, Gradient Boosting reports F1 = 0.8445 and MCC = 0.6917, and XGBoost reports F1 = 0.8445 and MCC = 0.6917. The stacked model's absolute F1 gain over Random Forest is 0.0095 and its MCC gain is 0.0173, indicating improved balanced performance and stronger overall predictive correlation. The smaller absolute margins compared with the German dataset suggest the ensemble consolidates strengths of high-performing trees but yields diminishing incremental returns when base learners are already closely competitive.

Table 3. Performance comparison of classifiers on the German dataset (original feature space)

Model	ACC	$F1$	MCC	PRE	REC
KNN	0.7433	0.6757	0.3576	0.6909	0.6674
NB	0.7066	0.6751	0.3149	0.6547	0.6603
LR	0.7633	0.6962	0.4027	0.7192	0.6849
DT	0.67	0.6454	0.3182	0.647	0.6722
RF	0.7366	0.6874	0.375	0.6869	0.688
GB	0.7666	0.6801	0.3938	0.7059	0.6468
MLP	0.76	0.6907	0.3926	0.7148	0.6793
XGB	0.7533	0.6619	0.3559	0.7122	0.6492
Stacked (Proposed)	0.77	0.7023	0.4174	0.7296	0.6896

Source:*Research finding*

Table 4. Performance comparison of classifiers on the Australian dataset (original feature space)

Model	ACC	$F1$	MCC	PRE	REC
KNN	0.8405	0.8387	0.6775	0.8384	0.8391
NB	0.8357	0.8309	0.6677	0.8406	0.8271
LR	0.8405	0.8405	0.6953	0.8464	0.8489
DT	0.855	0.854	0.7096	0.8531	0.8565
RF	0.855	0.8546	0.7151	0.8553	0.8597
GB	0.8454	0.8445	0.6917	0.8439	0.8478
MLP	0.8309	0.8282	0.6567	0.8296	0.8271
XGB	0.8454	0.8445	0.6917	0.8439	0.8478
Stacked (Proposed)	0.8647	0.8641	0.7324	0.864	0.8684

Source: *Research finding*

6.3 Analysis of Table 5 (German Dataset)

After applying PCA, the ranking and behavior of classifiers on the German data change noticeably. Naive Bayes becomes the strongest performer in terms of balanced measures (REC = 0.6849, PRE = 0.7457, MCC = 0.4263, F1 = 0.7008, ACC = 0.7766), surpassing the proposed stacked ensemble (REC = 0.6500, PRE = 0.7046, MCC = 0.3504, F1 = 0.6621, ACC = 0.7500). Tree-based models and the multilevel stacking approach suffer a relative decline. This pattern suggests that PCA — as applied here — has removed or compressed nonlinear and interaction signals that tree learners and complex ensembles exploit, while at the same time producing a more decorrelated, approximately linear input space that suits Naive Bayes' conditional-independence assumptions. In practical terms, the result indicates that an unsupervised, global PCA transformation can advantage simple probabilistic classifiers at the cost of degrading more expressive, interaction-dependent learners. For practitioners, this implies PCA should be applied selectively (for example only on numeric features, or using supervised/target-aware dimensionality reduction) if the goal is to preserve the ensemble's full predictive power.

6.4 Analysis of Table 6 (Australian Dataset)

On the Australian dataset PCA has a milder effect and the proposed stacked ensemble remains the best overall performer (REC = 0.8608, PRE = 0.8578, MCC = 0.7186, F1 = 0.8588, ACC = 0.8599). While several simpler classifiers (e.g., Naive Bayes and Logistic Regression) also show relatively high scores after PCA (NB: MCC ≈ 0.7035, LR: MCC ≈ 0.6986), the ensemble preserves its lead in both discrimination and balance between sensitivity and specificity. This outcome implies that, for the Australian data, principal components retain the majority of the predictive signal (including the aspects that the ensemble exploits), so dimensionality reduction does not substantially impair sophisticated learners. The contrast with the German results highlights that the effect of PCA is dataset-specific: when the original feature space contains strong nonlinear interactions critical to complex models, PCA may harm them; when the predictive structure is largely captured by principal components, ensembles remain robust. Again, a selective or supervised dimensionality-reduction strategy is recommended if one needs to reduce dimensionality while retaining the advantages of powerful ensemble methods.

Table 5. Impact of dimensionality reduction (PCA) on classifier performance German dataset

Model	*ACC*	*F1*	*MCC*	*PRE*	*REC*
KNN	0.7433	0.6731	0.3542	0.6909	0.6642
NB	0.7766	0.7008	0.4263	0.7457	0.6849
LR	0.7566	0.6769	0.3741	0.7129	0.6642
DT	0.66	0.6155	0.2364	0.6128	0.6238
RF	0.7433	0.6705	0.3508	0.691	0.6611
GB	0.73	0.6087	0.2719	0.6778	0.6039
MLP	0.7533	0.6871	0.3815	0.7046	0.6777
XGB	0.7266	0.5832	0.2478	0.6791	0.5857
Stacked (Proposed)	0.75	0.6621	0.3504	0.7046	0.65

Source: *Research finding*

Table 6. Impact of dimensionality reduction (PCA) on classifier performance Australian dataset

Model	*ACC*	*F1*	*MCC*	*PRE*	*REC*
KNN	0.8405	0.839	0.6785	0.8382	0.8402
NB	0.8454	0.8453	0.7035	0.8502	0.8532
LR	0.8405	0.8405	0.6986	0.8486	0.85
DT	0.8502	0.8488	0.698	0.848	0.85
RF	0.8357	0.8346	0.6706	0.8337	0.8369
GB	0.8502	0.8493	0.7006	0.8484	0.8521
MLP	0.8357	0.8357	0.6905	0.8449	0.8456
XGB	0.8502	0.8482	0.6964	0.8486	0.8478
Stacked (Proposed)	0.8599	08588	0.7186	0.8578	0.8608

Source: *Research finding*

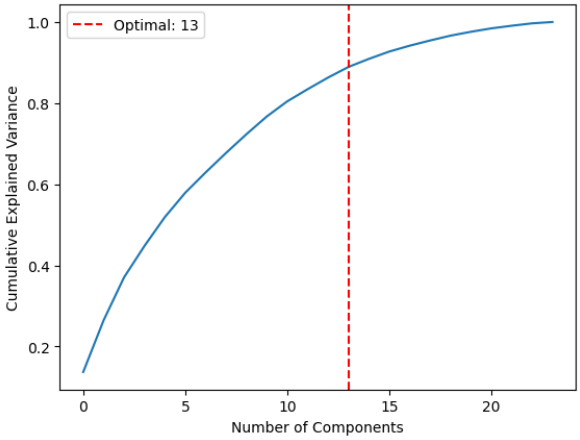

Figure 1. Cumulative variance explained by principal components for German credit data set

Source: *Research finding*

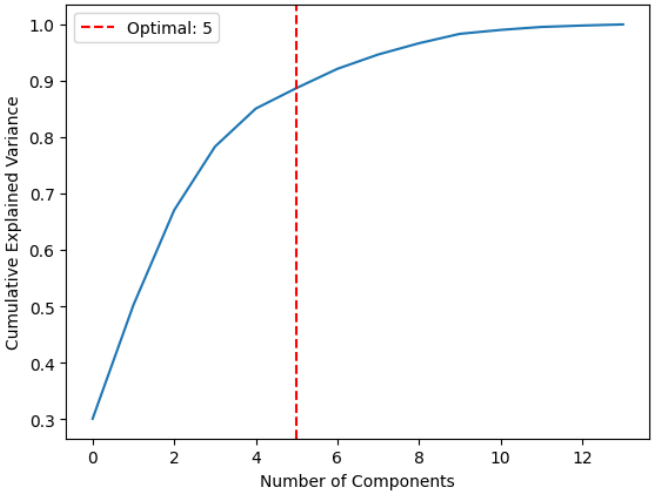

Figure 2. Cumulative variance explained by principal components for Australian credit data set

Source: *Research finding*

7. Conclusion

In this study we proposed a multilevel stacking ensemble that leverages the complementary strengths of Gradient Boosting, Extreme Gradient Boosting (XGBoost), and Random Forest, consolidated by a logistic-regression meta-learner. Extensive experiments on benchmark credit datasets from the UCI repository demonstrate that the proposed stacking framework consistently outperforms a wide range of standalone classifiers including KNN, Naive Bayes, single decision trees, and individual ensemble methods across multiple evaluation metrics such as accuracy, F1-score, and the Matthews Correlation Coefficient (MCC). These empirical gains reflect the practical advantage of combining diverse inductive biases in parallel: the meta-learner effectively synthesizes the base learners' complementary strengths, producing a more robust and generalizable predictor that balances bias and variance without imposing a strict ordering on base models.

We also examined the effect of unsupervised dimensionality reduction (PCA) on model performance. The impact was dataset-dependent: for the German dataset PCA caused a marked shift in relative rankings decorrelating the feature space and making it more favorable to simpler probabilistic models (e.g., Naive Bayes) while attenuating the benefits of interaction-sensitive learners and the stacking ensemble. In contrast, for the Australian dataset PCA had a milder effect and the stacked ensemble retained its lead, indicating that most predictive signal in that case was captured by the principal components. These findings emphasize that global, unsupervised PCA can both help and harm downstream learners depending on the data's intrinsic structure: it can reduce noise and overfitting risk, yet it may also remove nonlinear interaction terms that tree-based and ensemble models exploit.

Finally, the proposed methodology offers a scalable and practical blueprint for credit-risk classification in real-world settings. For practitioners we recommend applying dimensionality reduction selectively (for example, only to numeric features, or using supervised/target-aware reduction), calibrating probabilistic outputs before meta-learning, and evaluating models under cost-sensitive metrics that reflect business impact. Future work should explore alternative and more expressive meta-learners (e.g., gating networks or mixture-of-experts), uncertainty quantification (conformal prediction or Bayesian ensembles), adaptive resampling schemes for severe class imbalance, and extensions to multi-class or longitudinal credit-scoring problems. These directions will help further close the gap between methodological advances and operational deployment in financial risk systems.

Author Contributions

Conceptualization, all authors; methodology, Y.N. and M.E.; validation, M.E. and F.A.; formal analysis, all authors; resources, M.E. and S.M.H.; writing—original draft preparation, Y.N. and M.E.; writing—review and editing, all authors; supervision, Y.N. and S.M.H. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data will be available by email upon reasonable request.

Acknowledgements

Automated language-editing tools such as ChatGPT were used to improve the manuscript's English.

References

Alam, T. M., Fazli, A., Rahman, S., & Choudhury, T. (2020). An investigation of credit card default prediction in the imbalanced datasets. IEEE Access, 8, 201173–201198. https://doi.org/10.1109/ACCESS.2020.3033405.

Aruleba, I. T., & Sun, Y. (2025). An improved ensemble method with data resampling for credit risk prediction. IEEE Access, 13, 71275–71287. https://doi.org/10.1109/ACCESS.2025.XXXXX.

Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Bulut, C., & Arslan, E. (2024). Comparison of the impact of dimensionality reduction and data splitting on classification performance in credit risk assessment. Artificial Intelligence Review, 57(9), 252.

Chen, T., & Guestrin, C. (2015). XGBoost: Extreme gradient boosting (R package version 0.4-2, Vol. 1, No. 4, pp. 1–4).

Dornadula, V. N., & Geetha, S. (2019). Credit card fraud detection using machine learning algorithms. Procedia Computer Science, 165, 631–641. https://doi.org/10.1016/j.procs.2020.01.077.

Du, P., & Shu, H. (2022). Exploration of financial market credit scoring and risk management and prediction using deep learning and bionic algorithm. Journal of Global Information Management, 30(9), 1–29. https://doi.org/10.4018/JGIM.2022090101.

Emmanuel, I., Sun, Y., & Wang, Z. (2024). A machine learning-based credit risk prediction engine system using a stacked classifier and a filter-based feature

selection method. Journal of Big Data, 11(1), 23. https://doi.org/10.1186/s40537-024-xxxx-x.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451.

García, V., Marques, A. I., & Sánchez, J. S. (2012). Improving risk predictions by preprocessing imbalanced credit data. In Neural Information Processing (Vol. 67, pp. 68–75). https://doi.org/10.1007/978-3-642-34711-9_8.

Guo, S., He, H., & Huang, X. (2019). A multi-stage self-adaptive classifier ensemble model with application in credit scoring. IEEE Access, 7, 78549–78559. https://doi.org/10.1109/ACCESS.2019.2921383.

Han, J., Pei, J., & Tong, H. (2022). Data mining: Concepts and techniques (4th ed.). Morgan Kaufmann.

Han, W., Gu, X., & Jian, L. (2023). A multi-layer multi-view stacking model for credit risk assessment. Intelligent Data Analysis, 27(5), 1457–1475. https://doi.org/10.3233/IDA-220791.

Hoseini, S. M., Ebtia, M., & Khochiani, R. (2024). An ensemble method based on bagging SVM for credit rating problem. Soft Computing Journal. https://doi.org/10.1007/s00500-024-xxxx-x.

Khemakhem, S., & Boujelbene, Y. (2018). Predicting credit risk on the basis of financial and non-financial variables and data mining. Review of Accounting and Finance, 17(3), 316–340. https://doi.org/10.1108/RAF-07-2017-0108.

Kunapuli, G. (2023). Ensemble methods for machine learning. Simon & Schuster.

La Gatta, V., Postiglione, M., & Sperlì, G. (2025). A novel augmentation strategy for credit scoring modeling. Neural Computing and Applications, 37, 6663–6675. https://doi.org/10.1007/s00521-024-xxxx-x.

Liu, J., Liu, J., Wu, C., & Wang, S. (2024). Enhancing credit risk prediction based on ensemble tree-based feature transformation and logistic regression. Journal of Forecasting, 43(2), 429–455. https://doi.org/10.1002/for.XXXX.

Moradi, S., & Mokhatab, R. F. (2019). A dynamic credit risk assessment model with data mining techniques: Evidence from Iranian banks. Financial Innovation, 5(1), 15. https://doi.org/10.1186/s40854-019-0135-0.

Noriega, J., Rivera, L. A., & Herrera, J. (2023). Machine learning for credit risk prediction: A systematic literature review. Data, 8(11), 169. https://doi.org/10.3390/data8110169.

Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. Journal of Machine Learning Technologies, 2(1), 37–63.

Quan, J., & Sun, X. (2024). Credit risk assessment using the factorization machine model with feature interactions. Humanities and Social Sciences Communications, 11(1), 1–10. https://doi.org/10.1057/s41599-024-xxxx-x.

Rehman, Z. U., Muhammad, N., Sarwar, B., & Raz, M. A. (2019). Impact of risk management strategies on the credit risk faced by commercial banks of

Balochistan. Financial Innovation, 5(1), 44. https://doi.org/10.1186/s40854-019-0142-1.

Religia, Y., Pranoto, G. T., & Santosa, E. D. (2020). South German credit data classification using random forest algorithm to predict bank credit receipts. JISA (Jurnal Inform dan Sains), 3(2), 62–66.

Song, Y., & Peng, Y. (2019). A MCDM-based evaluation approach for imbalanced classification methods in financial risk prediction. IEEE Access, 7, 84897–84906. https://doi.org/10.1109/ACCESS.2019.2925034.

Veeramanikandan, V., & Jeyakarthic, M. (2021). Parameter-tuned deep learning model for credit risk assessment and scoring applications. Recent Advances in Computer Science and Communications, 14(9), 2958–2968. https://doi.org/10.2174/2356607514666210526141120.

Wei, Y., Kirkulak-Uludag, B., Zhu, D., & Luo, X. (2023). Stacking ensemble method for personal credit risk assessment in P2P lending. SSRN. https://doi.org/10.2139/ssrn.4318348.

Wolpert, D. H. (1992). Stacked generalization. Neural Networks, 5(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1.

Wu, C., Gao, D., & Xu, S. (2021). A credit risk predicting hybrid model based on deep learning technology. International Journal of Machine Learning and Computing, 11(3). https://doi.org/10.18178/ijmlc.2021.11.3.xxx.

Zhao, Z., & Aumeboonsuke, V. (2023). Imbalanced credit risk prediction in ensemble learning classifiers: A comparative analysis of SMOTE, ADASYN, SMOTETomek, and cluster centroids. Journal of Arts Management, 7(3), 959–984.

Zou, Y., & Gao, C. (2022). Extreme learning machine enhanced gradient boosting for credit scoring. Algorithms, 15(5), 149. https://doi.org/10.3390/a15050149.