IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 29 June 2024, accepted 14 August 2024, date of publication 16 August 2024, date of current version 28 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3445308

RESEARCH ARTICLE

Effective Credit Risk Prediction Using Ensemble Classifiers With Model Explanation

IDOWU ARULEBA AND YANXIA SUN, (Senior Member, IEEE)
Department of Electrical and Electronic Engineering Science, University of Johannesburg, Johannesburg 2006, South Africa

Corresponding author: Yanxia Sun (ysun@uj.ac.za)

This work was supported in part by South African National Research Foundation under Grant 141951, Grant 137951, and Grant AJCR230704126719120106.

ABSTRACT Credit risk prediction is a critical task in the financial industry, allowing lenders to assess the likelihood of a borrower defaulting on a loan. Traditional machine learning (ML) classifiers have been widely used for this purpose, and they often struggle with imbalanced data and lack interpretability, making it challenging for financial institutions to make informed decisions. This article explores the use of ensemble classifiers and Synthetic minority over-sampling Edited nearest neighbor (SMOTE-ENN) technique in credit risk prediction, aiming to improve the classification performance. The ensemble classifiers include Random Forest, adaptive boosting (AdaBoost), extreme gradient boosting (XGBoost), and light gradient boosting machine (LightGBM). The study addresses the class imbalance issue by leveraging ensemble classifiers and the SMOTE-ENN technique while employing Shapley additive exPlanations (SHAP) for model interpretability. The experimental results showed that the proposed approach resulted in improved classification performance. Specifically, on the German credit dataset, XGBoost outperformed the other models with a Recall of 0.930 and a Specificity of 0.846, while Random Forest obtained the best performance on the Australian dataset, achieving a Recall of 0.907 and Specificity of 0.922. Additionally, the integration of SHAP enhanced the models' transparency by providing valuable insights into the contribution of individual features, which is crucial for informed financial decision-making.

INDEX TERMS CART, credit risk, ensemble learning, XAI, machine learning, SHAP.

I. INTRODUCTION

Credit risk prediction is a critical task in the financial industry, as it allows lenders to assess the likelihood of a borrower defaulting on a loan. Machine learning (ML) techniques have become increasingly popular in credit risk prediction due to their ability to handle large amounts of data and complex relationships between variables [1], [2], [3]. However, traditional methods of credit risk prediction often rely on single classifiers, which may not capture the complexity of the data and lead to poor performance.

Imbalanced data poses a challenge in credit risk prediction, as there are typically far more instances of non-defaulting borrowers than defaulting borrowers in the dataset [4], [5]. This imbalance can lead to biased models that prioritise the majority class and overlook the minority class [6], [6]. In order to address the class imbalance, researchers have been exploring various techniques, including resampling, ensemble learning, and cost-sensitive learning methods, ensuring that both defaulting and non-defaulting borrowers are represented equally [7], [8], [9].

Additionally, the black-box nature of machine learning algorithms makes it difficult to understand how the model arrives at its predictions, which is a critical aspect in the financial industry where transparency and interpretability are crucial for decision-making [10]. Recently, explainable artificial intelligence (XAI) techniques have been developed to increase the transparency of machine learning algorithms [11]. XAI enables researchers to gain insights into how the model makes its predictions [12]. This allows for a better understanding of the factors contributing to credit risk, enabling financial institutions to make more informed decisions.

The associate editor coordinating the review of this manuscript and approving it for publication was Juan A. Lara.

The combination of efficient data resampling, ensemble learning, and XAI has the potential to enhance credit risk prediction. Therefore, in this study, we aim to explore the effectiveness of ensemble classifiers in predicting credit risk while providing model explanations. Specifically, the study aims to compare the performance of some popular ensemble classifiers, including random forest, AdaBoost, XGBoost, and LightGBM. These classifiers are also compared with traditional single classifiers, such as decision trees, logistic regression, and Multilayer perceptron (MLP). The study also aims to examine the interpretability of the models by generating SHAP values and plots to understand the factors influencing credit risk prediction. The main contributions of this study are:

- Investigating the impact of ensemble classifiers on credit risk prediction.
- Demonstrating the potential of SMOTE-ENN in improving credit risk prediction models.
- Providing insights into the key factors influencing credit risk through model interpretation.
- Discussing the implications of our findings for financial institutions and the future of credit risk prediction.

The rest of this paper is structured as follows: Section II reviews related credit risk literature. Section III presents the various datasets, ML algorithms, performance evaluation metrics used in the study, and the proposed credit risk framework. Section IV presents and discusses the experimental results, and Section IV concludes the study.

II. RELATED WORKS

Credit risk assessment is pivotal in the financial domain, with early works primarily leveraging logistic regression and decision tree algorithms for prediction tasks [13], [14], [15], [16]. These traditional methods, while foundational, exhibit limitations, particularly when dealing with imbalanced datasets, which is a common scenario in credit risk modelling [17], [18].

To address these challenges, recent studies have been using ensemble learning techniques that combine multiple classifiers to enhance the predictive accuracy [19]. Random Forest and AdaBoost, for instance, have been demonstrated to offer improved performance by aggregating diverse decision trees to reduce variance and bias [9], [20], [21], [22]. More advanced gradient boosting frameworks like XGBoost and LightGBM further build upon this, introducing more sophisticated optimization and regularization techniques to tackle overfitting and speed the training processes [4], [23], [24].

Meanwhile, resampling techniques, such as SMOTE and ENN, have been used to address the class imbalance problem. For example, Mahbobi et al. [25] incorporated SMOTE and artificial neural networks for credit risk prediction, obtaining an accuracy of 98.6%. Also, Gicic and Donko [26] used SMOTE with deep learning architectures, including long short-term memory (LSTM), stacked LSTM, Bidirectional LSTM (BiLSTM), and Stacked BiLSTM networks, with the stacked BiLSTM obtaining the highest accuracy of 87.19%.

The challenges noted in these studies typically include handling large volumes of data, integrating diverse types of data (e.g., structured and unstructured), and ensuring the fairness and unbiased nature of the predictions.** **Recent literature also highlights the struggle with computational efficiency and scalability when applying complex models to real-world data, pointing to an ongoing need for innovations in computational strategies and hardware optimization [27], [28], [29].

Furthermore, interpretability in machine learning has received significant attention, leading to the utilization of SHapley Additive exPlanations for explaining individual predictions in different applications [30], [31], [32]. SHAP values, based on cooperative game theory, provide a mathematically rigorous approach to quantify the contribution of each feature to a given prediction, thus offering transparency and insight into model decisions [33]. For instance, Wang et al. [34] employed SHAP values to interpret the prediction of ML models applied to student loan default prediction. The integration of the SHAP technique led to several insights, such as how the risk of student loan default is influenced by factors like college entrance examination scores, academic performance, and the number of scholarships received by the student.

Major findings in the field suggest that while advanced ML techniques can significantly improve predictive accuracy, they often sacrifice transparency. This trade-off calls for continuous efforts to balance complexity and interpretability, ensuring that models perform well and are understandable to stakeholders [28], [35]. Meanwhile, the reviewed research works and the challenges identified indicate the need for a multi-faceted approach to credit risk evaluation. Therefore, this study proposes an approach that combines the strengths of ensemble classifiers, resampling techniques, and interpretability methods to create robust, understandable models.

This study advances existing models by employing sophisticated ensemble classifiers, including random forest, AdaBoost, XGBoost, and LightGBM, which, when combined with the SMOTE-ENN technique, effectively addresses the issue of data imbalance that affects traditional single classifiers. Furthermore, using SHAP for model interpretability provides deeper insights into the decision-making processes, enhancing the transparency and usability of the predictive models in practical scenarios.

III. MATERIALS AND METHODS

A. DATASETS

In this study, two publicly available datasets are employed. The first is the German Credit Dataset [36], which is comprised of 1000 instances, each representing an individual credit applicant. The features include a range of attributes, shown in Table 1. Each attribute contributes to the final decision of classifying the credit risk of the applicant as either

TABLE 1. German credit dataset.

S/N	Feature	Description	Data Type
1	Status of existing checking account	Check account	Categorical
2	Duration	Duration of the credit in months	Numeric
3	Purpose	Purpose of the credit	Categorical
4	Credit amount	Credit amount requested by the applicant	Numeric
5	Savings account	Savings account	Categorical
6	Age	Age of the applicant in years	Numeric
7	Employment	Employment duration	Categorical
8	Sex	Sex	Categorical
9	Housing	Housing status (i.e., own/rent/free)	Categorical
10	Job	Applicant's Job category	Categorical
11	Credit Risk	Class variable (1=good and 2=bad)	Binary

good or bad. This dataset is widely used for benchmarking classification algorithms in credit risk assessment.

The second dataset is the Australian Credit Approval dataset [37], a publicly available dataset for credit card applications. It comprises instances representing individual applicants for credit, where the various attributes relate to the applicant's financial history. The dataset's primary objective is to predict whether an application should be approved or denied, making it an essential resource for developing and testing credit scoring models. The dataset is described in Table 2.

TABLE 2. Australian credit approval dataset.

S/N	Feature	Description	Data Type
1	A1	Sex	Nominal
2	A2	Age	Continuous
3	A3	Mean time at addresses	Continuous
4	A4	Home status	Nominal
5	A5	Current occupation	Nominal
6	A6	Current job status	Nominal
7	A7	Mean time with employers	Continuous
8	A8	Other investments	Nominal
9	A9	Bank account	Nominal
10	A10	Time with bank	Continuous
11	A11	Liability reference	Nominal
12	A12	Account reference	Nominal
13	A13	Monthly housing expense	Continuous
14	A14	Savings account balance	Continuous
15	Class	Target variable (Accept/Reject)	Nominal

B. CLASSIFIERS

1) RANDOM FOREST

Random Forest is a robust ensemble learning algorithm that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes of the individual decision trees [38], [39]. Random Forests aim to reduce the overfitting of individual decision trees by averaging their predictions, thereby improving the accuracy and robustness of the model [40].

The decision trees are created using different subsets of the dataset. For each tree, a bootstrap sample is used, i.e., a random selection with replacement from the training data. This process is known as bootstrap aggregating or bagging. During the construction of each tree, a random subset of features is selected at each split point, introducing feature randomness [41], [42]. This ensures that the trees in the forest are uncorrelated, enhancing the diversity of the ensemble and, consequently, its predictive performance. The typical choice for the number of features considered at each split is $\sqrt{m}$ for classification, where m is the total number of features.

Assuming $\{T_1, T_2, \ldots, T_n\}$ is the set of decision trees in the random forest, and $h(x, T_i)$ is the prediction of tree T_i for input x, the model prediction, $H(x)$, is the mode of the predictions made by the individual trees:

$$H(x) = \text{mode}\{h(x, T_1), h(x, T_2), \ldots, h(x, T_n)\}. \quad (1)$$

2) XGBoost

The XGBoost is an efficient and scalable implementation of gradient-boosted trees designed for speed and performance [10], [43]. It is known for its ability to handle sparse data and its use of a novel tree-learning algorithm. XGBoost improves on the traditional gradient boosting method by introducing a regularization term in the objective function, which helps to control over-fitting [44]. The core principle involves sequentially adding predictors that correct its predecessor, thus improving the model with each iteration. The objective function that XGBoost optimizes is given by:

$$L(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k), \quad (2)$$

where l is a differentiable convex loss function that measures the difference between the prediction $\hat{y}_i$ and the target y_i, and Ω penalizes the complexity of the model to avoid overfitting [44].

3) AdaBoost

AdaBoost focuses on converting a set of weak learners into strong learners by iteratively adjusting the weights of incorrectly classified instances. It adapts by giving more weight to difficult-to-classify instances and less to those already classified correctly in previous iterations [45], [46]. The final model is a weighted sum of the weak learners, calculated as follows:

$$F(x) = \sum_{t=1}^{T} \alpha_t h_t(x), \quad (3)$$

where $h_t(x)$ is the output of the weak learner, and α_t is its weight in the final prediction, with T being the total number of iterations.

4) LightGBM

LightGBM is a gradient boosting framework that uses tree-based learning algorithms designed for distributed and efficient training, particularly on large datasets [47]. It introduces two novel techniques: gradient-based One-Side Sampling (GOSS) to filter out the data instances to focus on those with larger gradients, and exclusive feature bundling (EFB) to reduce the number of features in sparse datasets [48]. LightGBM builds trees leaf-wise (best-first), as opposed to level-wise, allowing it to achieve lower loss compared to level-wise growth [49]. The LightGBM model is highly customizable and capable of handling categorical features natively.

5) CLASSIFICATION AND REGRESSION TREE

The Classification and Regression Trees (CART) algorithm is a tree-building technique that splits a dataset into subsets based on a decision rule inferred from the input variables. CART can be used for classification or regression tasks. The decision rule at each node is chosen based on the Gini impurity, aiming to maximize the homogeneity of the subsets [50], [51]. The Gini impurity is defined as follows:

$$\text{Gini}(D) = 1 - \sum_{i=1}^{m} p_i^2 \tag{4}$$

where D is the dataset at a node, m is the number of classes, p_i is the proportion of class i in the dataset D, N is the number of samples, y_i is the actual value, and $\hat{y}$ is the predicted value. The algorithm recursively splits the training set until a specified maximum depth is reached or no further gains can be made [38], [52], [53].

6) LOGISTIC REGRESSION

Logistic regression is a statistical method that is used to predict the probability of a binary outcome based on one or more predictor variables. The model is built on the logistic function, an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits [54], [55]. Mathematically, if P is the probability of the outcome, the logistic regression equation can be written as:

$$\log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \tag{5}$$

where:

- $\log\left(\frac{P}{1-P}\right)$ is the logit function.
- P is the probability of the presence of the characteristic of interest.
- $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ are the regression coefficients indicating the weight of each factor.
- $X_1, X_2, \ldots, X_k$ are the independent variables.

7) MULTILAYER PERCEPTRON

A Multilayer Perceptron is a class of feedforward artificial neural network (ANN) that consists of at least three layers of nodes: an input layer, a hidden layer, and an output layer [56]. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. MLP utilizes the backpropagation technique for training. It is represented by the following:

$$\hat{y} = \sigma(W_2\sigma(W_1 x + b_1) + b_2) \tag{6}$$

where x is the input vector, W_1 and W_2 are weights, b_1 and b_2 are bias terms, σ is the activation function, and $\hat{y}$ is the output. The model learns by adjusting the weights and biases to minimize the difference between the actual and predicted outputs, typically using gradient descent.

C. SMOTE-ENN

The Synthetic Minority Over-sampling Technique - Edited Nearest Neighbors is a hybrid approach for handling imbalanced datasets, which combines the over-sampling approach of SMOTE with the under-sampling technique of ENN [57]. SMOTE-ENN aims to balance dataset distribution by synthetically generating new instances of the minority class and removing the instances of both classes that are identified as noise or in the class overlapping areas. This method not only augments the minority class but also ensures that the classifier boundary is more general and not overfitted to the noisy data [41], [58].

Algorithm 1 SMOTE-ENN Algorithm

1: **Input:** Dataset D with minority class M and majority class N, Over-sampling rate S, Number of nearest neighbors k
2: **Output:** Balanced Dataset D'
3: Apply SMOTE to D with over-sampling rate S and k nearest neighbors to generate synthetic samples for minority class M
4: Combine original dataset D with synthetic samples to form an augmented dataset D_{aug}
5: Apply ENN to D_{aug} to remove samples that do not agree with majority of its k nearest neighbors
6: $D' \leftarrow$ The resulting dataset after ENN
7: **return** D'

Furthermore, SMOTE generates synthetic samples by operating in the feature space rather than the data space. For each minority class sample x_i, SMOTE selects k nearest neighbors from the minority class, chooses one neighbor x_{nn} at random, and generates a new sample x_{new} as follows:

$$x_{\text{new}} = x_i + \lambda \cdot (x_{\text{nn}} - x_i), \tag{7}$$

where λ is a random number between 0 and 1. Meanwhile, ENN removes instances of the dataset that are misclassified by their k nearest neighbors, making the class boundaries cleaner and less prone to overfitting. The SMOTE-ENN

technique is described in Algorithm 1. The combination of SMOTE and ENN effectively balances the class distribution while removing noisy and borderline instances, thus improving the performance of classification models on imbalanced datasets.

D. SHAPLEY ADDITIVE EXPLANATIONS

Interpretable machine learning seeks to explain the decision-making process of complex models, offering insights into the contribution of individual features to the output of the model [59], [60], [61]. SHAP is a game theory-based approach that assigns each feature an importance value for a particular prediction. It connects optimal credit allocations with local explanations using the classic Shapley values from cooperative game theory and their related extensions [62], [63]. The Shapley value is the average marginal contribution of a feature value across all possible coalitions. For a prediction instance i and a feature value v_j, the Shapley value ϕ_j is computed as:

$$\phi_j(v_j) = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \tag{8}$$

$$[f_{S\cup\{j\}}(v_S \cup v_j) - f_S(v_S)] \tag{9}$$

where F is the set of all features, S is a subset of features, f_S is the model function trained on the feature subset S, v_S is the set of feature values for S, and $|\cdot|$ denotes the cardinality of a set. SHAP values interpret the impact of having a certain value for a given feature compared to the prediction we would make in the absence of that feature [64]. This aligns with the intuition that features contributing positively towards the prediction will have larger SHAP values and vice versa for features detracting from the prediction outcome. SHAP not only provides a measure of feature importance but also offers local explanations that reveal the effect of each feature on an individual prediction. This granular insight is valuable, especially in critical applications such as finance, where understanding model predictions is essential for trust and actionable intelligence.

E. PERFORMANCE EVALUATION METRICS

Performance evaluation metrics are important for assessing the effectiveness of ML models. In this study, the following metrics are used: accuracy, precision, recall, specificity, and F-measure. These metrics are obtained from the confusion matrix, a table used to describe the performance of a classification model on a set of test data for which the true values are known. It contains information about actual and predicted classifications done by a classifier and helps to visualize the performance of an algorithm. The terms involved are:

- True Positives (TP): Correctly predicted positive observations
- True Negatives (TN): Correctly predicted negative observations
- False Positives (FP): Incorrectly predicted positive observations
- False Negatives (FN): Incorrectly predicted negative observations

$$\text{Confusion Matrix} = \begin{bmatrix} TN & FP \\ FN & TP \end{bmatrix}$$

1) ACCURACY

Accuracy measures the proportion of correctly predicted observations in the dataset. It is calculated as the number of correct predictions divided by the total number of predictions [65].

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

2) SENSITIVITY

Sensitivity, as called recall, measures the proportion of correctly predicted positive instances out of all actual positive instances in the dataset [66]. It is calculated as the number of true positive predictions divided by the sum of true positive and false negative predictions. They are defined as:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{11}$$

3) SPECIFICITY

Specificity quantifies the test's ability to correctly identify negative instances among all actual negatives in the dataset. Specifically, it measures the proportion of true negative outcomes to the total number of actual negative cases [67], [68]. It is given by the formula:

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{12}$$

A high specificity value indicates that the model effectively identifies negative cases and minimises false positives [69].

4) F-MEASURE

The F-measure is the harmonic mean of precision and recall, providing a balance between the two [70]. It is defined as:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{13}$$

Meanwhile, Precision measures the proportion of correctly predicted positive instances out of all instances that were predicted as positive [71]. It is represented mathematically as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{14}$$

F. PROPOSED FRAMEWORK FOR CREDIT RISK PREDICTION

The proposed framework for credit risk prediction is designed to leverage the strengths of ensemble classifiers and the SMOTE-ENN technique. This framework integrates multiple stages, from data resampling through model development and evaluation to interpretation, emphasizing the importance of

class balance and interpretability. The significance of this approach lies in its capacity to handle imbalanced datasets, a common challenge in credit risk modeling. SMOTE-ENN enhances the representation of minority classes, thus mitigating model bias towards the majority class and improving the predictive accuracy across diverse applicant profiles. Additionally, the use of ensemble methods helps to reduce variance and bias, further enhancing model performance. Algorithm 2 outlines the step-by-step process of the proposed framework:

Algorithm 2 Proposed Credit Risk Prediction Framework

Input: Training data $\mathcal{D}_{train}$ and testing data $\mathcal{D}_{test}$
Output: Credit risk predictions $\mathcal{P}$
Step 1: Data Preprocessing
Normalize and encode features in $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$.
Apply SMOTE-ENN to $\mathcal{D}_{train}$ for class balancing.
Step 2: Train Ensemble Classifiers
Train the following models on the balanced $\mathcal{D}_{train}$:
    1) Random Forest
    2) AdaBoost
    3) XGBoost
    4) LightGBM
Step 3: Model Validation
Validate each model using k-fold cross-validation.
Evaluate model performance based on accuracy, recall, specificity, and F-measure.
Step 4: Feature Importance Evaluation with SHAP
Identify the best-performing model based on validation results.
Compute SHAP values for the best-performing model to assess the impact of each feature.

IV. RESULTS AND DISCUSSION

In this section, we present the experimental results obtained by the various classifiers before and after the data resampling. Since the German and Australian datasets were used in the study, it is important to analyze the performance of the models on these datasets separately.

A. EXPERIMENTAL RESULTS USING THE GERMAN DATASET

The German Credit Dataset, employed in this study, comprises 1,000 instances, each described by 11 features representing various financial and personal attributes of loan applicants. The dataset, described in Table 1, is characterized by a binary class distribution with applicants labeled as 'good' or 'bad' credit risks. This skewed distribution, where approximately 70% of instances are labeled 'good', presents inherent challenges in model training due to class imbalance. The performance of the models on the German dataset before and after resampling is shown in Tables 3 and 4, respectively. Prior to resampling, it was observed that all classifiers struggled particularly with respect to specificity, as evidenced by the values not surpassing the 0.474 mark. This suggests a potential bias towards the majority class, a common issue in imbalanced datasets. Meanwhile, the XGBoost demonstrated superior performance with the highest accuracy (0.765) and F-measure (0.841); however, the specificity was poor, showing the limitation in effectively identifying true negatives.

TABLE 3. Classifiers performance on German dataset before resampling.

Classifier	Accuracy	Recall	Specificity	F-measure
AdaBoost	0.750	0.781	0.400	0.834
LightGBM	0.732	0.767	0.390	0.810
Random forest	0.750	0.890	0.430	0.833
XGBoost	0.765	0.789	0.440	0.841
CART	0.650	0.724	0.474	0.750
Logistic regression	0.710	0.743	0.252	0.786
MLP	0.705	0.775	0.280	0.727

TABLE 4. Classifiers performance on German dataset after resampling.

Classifier	Accuracy	Recall	Specificity	F-measure
AdaBoost	0.820	0.857	0.730	0.860
LightGBM	0.790	0.829	0.698	0.858
Random forest	0.861	0.911	0.825	0.874
XGBoost	0.897	0.930	0.846	0.907
CART	0.686	0.748	0.599	0.791
Logistic regression	0.814	0.798	0.630	0.805
MLP	0.820	0.829	0.801	0.820

After the data resampling using the SMOTE-ENN technique, a substantial improvement in classifier performance was evident, with notable improvements in the various classifiers. XGBoost again leads with an increase in accuracy to 0.897 and recall to 0.930, signifying an increase in the model's ability to correctly identify positive class instances. These enhancements reflect the resampling's efficacy in providing a more representative distribution of classes, which is crucial for improving the model's learning and generalization capabilities. This enhancement in recall is also observed in the random forest, which showed a notable increase in specificity to 0.825. The improvement in specificity across classifiers post-resampling demonstrates the effectiveness of the data resampling in mitigating class imbalance, which resulted in a more robust predictive model.

Furthermore, the classifiers also improved with regard to the F-measure after resampling, particularly with XGBoost achieving a score of 90.7%. This improvement in F-measure across the board indicates a more harmonious balance between precision and recall, which further demonstrates the positive impact of the SMOTE-ENN resampling on classifier performance. The enhanced balance between sensitivity and specificity after resampling suggests that models became more skilled at generalizing, reducing the bias towards the majority class, and improving the predictive performance for the minority class. These improvements are critical for credit risk assessment, as they ensure a fair and accurate evaluation of potential credit risks, potentially leading to more stable financial portfolios.

Meanwhile, since the XGBoost obtained the best performance, it will be worthwhile to understand how it reached its predictions and which features were most significant in the decision-making process. Therefore, Figure 1 provides a summary plot that shows the distribution of the SHAP values for each feature across all the data points in the test set. The colour coding (red for high and blue for low) indicates the value of the feature for each instance, with red dots indicating higher feature values and blue dots indicating lower values. The horizontal dispersion of the dots represents the distribution of the SHAP values for each feature, with a wider spread indicating higher variability in the feature's impact on the model's output [72], [73].

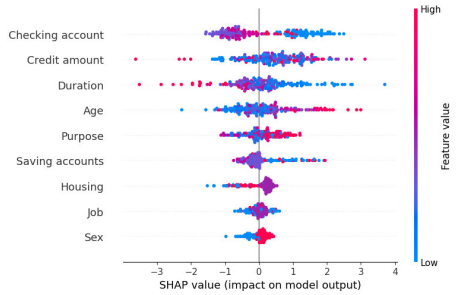


FIGURE 1. SHAP summary plot 2 for German credit dataset.

From Figure 1, the feature at the top, i.e., Checking account, is the most impactful feature. According to the model, high values of 'Checking account' generally lead to a higher prediction value, potentially indicating a 'good' credit risk. Credit amount also has a significant impact but in both positive and negative directions, suggesting a complex relationship that may depend on other factors or interactions between features. Duration and Age show a mix of positive and negative influences, indicating that their effect on the model's output is not uniform across all data points. Lastly, Purpose, Saving accounts, Housing, Job, and Sex seem to have a smaller impact on the model's output, as indicated by the closer clustering of SHAP values around zero.

B. EXPERIMENTAL RESULTS USING THE AUSTRALIAN DATASET

The Australian Credit Approval Dataset contains 690 instances and 15 features related to various personal and financial attributes of credit card applicants. The class distribution is nearly balanced, with approximately 55% of the instances classified as positive (credit approved) and 45% as negative (credit denied), providing a diverse scenario for evaluating model performance. The performance metrics for the various classifiers on the Australian dataset before and after resampling are shown in Tables 5 and 6, respectively. Before resampling, XGBoost outperformed other classifiers with the highest accuracy (0.895), recall (0.870), and F-measure (0.886), indicating a robust predictive performance that is balanced across the various classes. Notably, while AdaBoost and LightGBM trailed closely in performance, the CART algorithm demonstrated comparatively lower performance, with an accuracy of only 0.790 and an F-measure of 0.780.

TABLE 5. Classifiers performance on Australian dataset before resampling.

Classifier	Accuracy	Recall	Specificity	F-measure
AdaBoost	0.860	0.855	0.883	0.880
LightGBM	0.859	0.850	0.884	0.860
Random forest	0.840	0.831	0.850	0.827
XGBoost	0.895	0.870	0.899	0.886
CART	0.790	0.772	0.816	0.780
Logistic regression	0.836	0.848	0.842	0.820
MLP	0.806	0.845	0.779	0.780

TABLE 6. Classifiers performance on Australian dataset after resampling.

Classifier	Accuracy	Recall	Specificity	F-measure
AdaBoost	0.881	0.899	0.910	0.890
LightGBM	0.865	0.840	0.875	0.854
Random forest	0.916	0.907	0.922	0.910
XGBoost	0.910	0.898	0.914	0.903
CART	0.829	0.794	0.850	0.827
Logistic regression	0.860	0.866	0.880	0.871
MLP	0.849	0.880	0.832	0.830

After applying the SMOTE-ENN resampling to address the class imbalance, improvements across all classifiers were observed, particularly in recall and F-measure. The random forest showed the most noticeable enhancement, with accuracy increasing to 0.916 and F-measure reaching 0.910, surpassing the XGBoost. This suggests that the random forest benefited more from the resampling, potentially due to its inherent handling of feature selection and decision boundaries in a more balanced dataset.

The resampling process evidently mitigated the bias towards the majority class, as demonstrated by improved sensitivity values across all classifiers. The consistent increase in specificity after resampling also indicates that the classifiers maintained an improved true negative rate, confirming the notion that balancing the dataset can lead to models with better generalization and a fairer representation of both classes.

Furthermore, having obtained the best classification performance, the random forest model is explored further to understand the features contributing more to the predictions. Therefore, the SHAP technique is applied, and the summary plot is shown in Figure 2.

From Figure 2, it can be observed that feature A9 appears to be the most significant predictor, with high values leading to a higher output of the model, which could be interpreted as an increased likelihood of credit approval. Feature A15 shows a high impact on the model output with a mix of positive and negative contributions, suggesting a complex, non-linear relationship with the target variable. Features A11 and A8 also play essential roles but with a more moderate impact compared to A9. Lower-impact features, such as A6, A3, A2, A5, A4, and A7, still contribute to the model's predictions,

TABLE 7. Comparison with studies that used German dataset.

Reference	Method	Accuracy	Recall	Specificity	F-measure	Interpretability
[74]	MLP	0.730	-0	-	-	X
[75]	Optimized Random forest	0.856	0.828	-	0.852	X
[76]	Random forest with Feature selection	0.788	0.735	0.764	-	X
[76]	MLP with Feature selection	0.761	0.691	0.700	-	X
[13]	Gradient Boosted Decision Tree and SMOTE	0.824	0.838	-	0.834	X
[13]	Gradient Boosted Decision Tree and SMOTETomek	0.835	0.868	-	0.844	X
[77]	three-layer stacked LSTM	0.871	-	-	-	X
[78]	LSTM Ensemble with SMOTE-ENN	0.904	-	-	-	X
[79]	MLP and Grid Search	0.810	-	-	-	X
[79]	XGBoost and Grid Search	0.816	-	-	-	X
[80]	Gaussian Mixture Model	0.742	0.756	-	0.741	X
[81]	CART based Ensemble with Majority Voting	0.772	0.960	0.370	-	X
[82]	Rotation Forest	0.770	-	-	-	X
[82]	Random Subspace Decision Tree	0.761	-	-	-	X
[83]	Boosted Logistic Regression Ensemble	0.810	-	-	-	X
Our Approach	XGBoost and SMOTE-ENN	0.895	0.870	0.899	0.886	✓

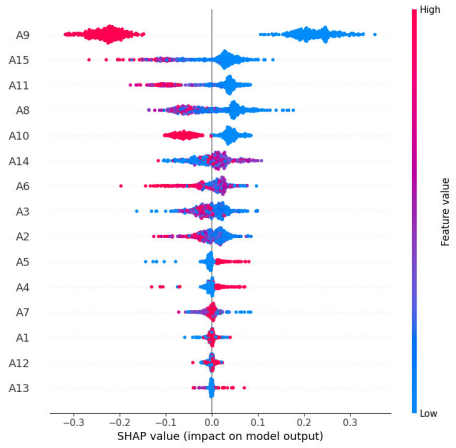


FIGURE 2. SHAP summary plot 2 for australian credit approval dataset.

but to a lesser extent. This lesser impact suggests that while they are relevant, these features do not singularly drive credit decisions but might still be considered in a more holistic assessment of credit risk.

Understanding these feature impacts through SHAP values enhances the transparency and fairness of the credit scoring process and provides actionable insights that can guide future data collection and feature engineering efforts. By focusing on high-impact features, financial institutions can streamline their data collection strategies to capture the most relevant information, improving the efficiency and accuracy of their credit risk assessments.

C. COMPARISON WITH RECENT STUDIES

This section presents a comparative analysis of our approach with other recent studies that utilized the German and Australian datasets for credit risk prediction. Tables 7 and 8 summarize the performances of various methodologies, providing a comprehensive view of how our approach stands in comparison to the state-of-the-art.

Firstly, for the German dataset, our methodology not only outperforms others in terms of overall accuracy and balance between recall and specificity but also offers the added benefit of interpretability, a crucial aspect often overlooked in many studies. For instance, Esenogho et al. [78] employed an LSTM Ensemble with SMOTE-ENN, achieving a high accuracy of 0.904 but lacking interpretability. Similarly, Alam et al. [13] achieved competitive results with Gradient Boosted Decision Trees; however, their models also lacked the interpretability provided by our use of SHAP values. This feature of our approach is critical, as it allows for better understanding and trust in model decisions, which is particularly important in financial applications.

Similarly, for the Australian dataset, our Random Forest and SMOTE-ENN approach outperformed most models, achieving an accuracy of 0.916, recall of 0.907, and specificity of 0.922 with an F-measure of 0.910. This performance is notable when compared to [81], which, while achieving a high recall of 0.953 with a CART-based Ensemble, did not manage to maintain as high specificity, showcasing our model's ability to maintain balance. Additionally, our model's interpretability, marked by the ✓ symbol, sets it apart from other studies.

D. DISCUSSION

This study's findings demonstrate the significant advancements in credit risk prediction models, particularly through the integration of ensemble learning techniques and advanced data preprocessing methods like SMOTE-ENN. The results highlight the robustness of these methods in enhancing classification performance and their ability to balance the trade-offs between recall and specificity effectively. The application of ensemble classifiers such as XGBoost and Random Forest, combined with SMOTE-ENN, resulted in notably high accuracy, recall, specificity, and F-measure across both German and Australian credit datasets. This indicates a robust capability to identify both defaulters and non-defaulters accurately, a critical requirement in credit risk assessments to minimize financial losses.

One of the standout features of the proposed approach in this study is its emphasis on model interpretability, enabled by the use of SHAP values. Financial institutions highly

TABLE 8. Comparison with studies that used the Australian dataset.

Reference	Method	Accuracy	Recall	Specificity	F-measure	Interpretability
[74]	MLP	0.869	-	-	-	X
[76]	Random forest with Feature selection	0.873	0.882	0.883	-	X
[76]	MLP with Feature selection	0.853	0.854	0.858	-	X
[79]	MLP and Grid Search	0.790	-	-	-	X
[79]	XGBoost and Grid Search	0.890	-	-	-	X
[80]	Gaussian Mixture Model	0.867	0.898	-	0.872	X
[81]	CART based Ensemble with Majority Voting	0.913	0.953	0.871	-	X
[82]	Rotation Forest	0.865	-	-	-	X
[82]	Random Subspace Decision Tree	-	0.869	-	-	X
[83]	Boosted Logistic Regression Ensemble	0.880	-	-	-	X
[84]	Decision Tree and SMOTE	0.838	-	-	0.836	X
[84]	Decision Tree and ADASYN	0.816	-	-	0.816	X
[85]	Convolutional Neural Network	0.880	-	-	-	✓
[86]	MLP-based Bagging Ensemble	0.865	-	-	-	X
[87]	Evolutionary Extreme Learning Machine	0.871	0.859	0.878	-	X
Our Approach	Random Forest and SMOTE-ENN	0.916	0.907	0.922	0.910	✓

value transparency in predictive models, as it aids in the justification of decision-making processes and compliance with regulatory requirements. The inclusion of SHAP values helps in identifying the features that most significantly impact the model's predictions and facilitates a deeper understanding of the underlying decision-making processes of the algorithms. This aspect is particularly important in gaining stakeholder trust and for the broader acceptance and deployment of these models in sensitive financial sectors.

When compared with other studies, our approach demonstrates superior performance and interpretability. Many recent methods achieve high performance but often at the cost of model transparency. Our methodology bridges this gap by maintaining high performance while also providing clear insights into the contributing factors of the predictions. Meanwhile, the practical implications of these findings are enormous. By improving the accuracy and transparency of credit risk assessments, financial institutions can make more informed decisions, potentially leading to lower default rates and better management of financial risk. Furthermore, the approach outlined in this study provides a scalable and adaptable framework that can be tailored to various types of credit products and diverse geographical markets.

V. CONCLUSION

This study presented a robust approach for credit risk prediction using ensemble learning algorithms, SMOTE-ENN resampling, and SHAP for model interpretation. The experimental results indicate the robustness of the proposed approach, with the XGBoost and random forest achieving the best performance on the German and Australian datasets, respectively. The proposed approach significantly outperformed other methods in recent literature, indicating its robustness. Meanwhile, understanding the decision-making process of ML models is critical, especially in financial applications like credit risk prediction. Therefore, this study examined the SHAP summary plot to interpret the contributions of different features to a predictive model's output.

The study contributes to the field of credit risk prediction by demonstrating how advanced ML techniques can be effectively applied to improve both the performance and transparency of predictive models. The integration of ensemble learning, SMOTE-ENN, and SHAP for interpretability has shown to be effective, offering a comprehensive approach that could be crucial for future developments in credit risk management. Meanwhile, despite the promising results, this study is not without limitations. The dependency on the quality and the representativeness of the data used can significantly influence the outcomes. Future research could explore the integration of more diverse data sources to further enhance the predictive power of the models. Additionally, testing the models across more varied datasets could help in establishing their robustness and generalizability.

REFERENCES

[1] S. Bhatore, L. Mohan, and Y. R. Reddy, "Machine learning techniques for credit risk evaluation: A systematic literature review," *J. Banking Financial Technol.*, vol. 4, pp. 111–138, Jun. 2020.

[2] W. Bao, N. Lianju, and K. Yue, "Integration of unsupervised and supervised machine learning algorithms for credit risk assessment," *Expert Syst. Appl.*, vol. 128, pp. 301–315, Aug. 2019.

[3] P. Addo, D. Guegan, and B. Hassani, "Credit risk analysis using machine and deep learning models," *Risks*, vol. 6, no. 2, p. 38, Apr. 2018.

[4] C. Rao, Y. Liu, and M. Goh, "Credit risk assessment mechanism of personal auto loan based on PSO-XGBoost model," *Complex Intell. Syst.*, vol. 9, no. 2, pp. 1391–1414, Apr. 2023.

[5] I. D. Mienye and N. Jere, "Deep learning for credit card fraud detection: A review of algorithms, challenges, and solutions," *IEEE Access*, vol. 12, pp. 96893–96910, 2024.

[6] S. Zian, S. A. Kareem, and K. D. Varathan, "An empirical evaluation of stacked ensembles with different meta-learners in imbalanced classification," *IEEE Access*, vol. 9, pp. 87434–87452, 2021.

[7] Y. Li and W. Chen, "A comparative performance assessment of ensemble learning for credit scoring," *Mathematics*, vol. 8, no. 10, p. 1756, Oct. 2020.

[8] G. Obaido, B. Ogbuokiri, C. W. Chukwu, F. J. Osaye, O. F. Egbelowo, M. I. Uzochukwu, I. D. Mienye, K. Aruleba, M. Primus, and O. Achilonu, "An improved ensemble method for predicting hyperchloremia in adults with diabetic ketoacidosis," *IEEE Access*, vol. 12, pp. 9536–9549, 2024.

[9] I. D. Mienye and Y. Sun, "A machine learning method with hybrid feature selection for improved credit card fraud detection," *Appl. Sci.*, vol. 13, no. 12, p. 7254, Jun. 2023.

[10] J. He, Y. Hao, and X. Wang, "An interpretable aid decision-making model for flag state control ship detention based on SMOTE and XGBoost," *J. Mar. Sci. Eng.*, vol. 9, no. 2, p. 156, Feb. 2021.

[11] J. Mary, C. Calauzenes, and N. El Karoui, "Fairness-aware learning for continuous attributes and treatments," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 4382–4391.

[12] G. Obaido, B. Ogbuokiri, T. G. Swart, N. Ayawei, S. M. Kasongo, K. Aruleba, I. D. Mienye, I. Aruleba, W. Chukwu, F. Osaye, O. F. Egbelowo, S. Simphiwe, and E. Esenogho, "An interpretable machine learning approach for hepatitis b diagnosis," *Appl. Sci.*, vol. 12, no. 21, p. 11127, Nov. 2022.

[13] T. M. Alam, K. Shaukat, I. A. Hameed, S. Luo, M. U. Sarwar, S. Shabbir, J. Li, and M. Khushi, "An investigation of credit card default prediction in the imbalanced datasets," *IEEE Access*, vol. 8, pp. 201173–201198, 2020.

[14] Y.-C. Chang, K.-H. Chang, H.-H. Chu, and L.-I. Tong, "Establishing decision tree-based short-term default credit risk assessment models," *Commun. Statist.-Theory Methods*, vol. 45, no. 23, pp. 6803–6815, Dec. 2016.

[15] J. Kruppa, A. Schwarz, G. Arminger, and A. Ziegler, "Consumer credit risk: Individual probability estimates using machine learning," *Expert Syst. Appl.*, vol. 40, no. 13, pp. 5125–5131, Oct. 2013.

[16] S. Mestiri and M. Hamdi, "Credit risk prediction: A comparative study between logistic regression and logistic regression with random effects," *Int. J. Manage. Sci. Eng. Manage.*, vol. 7, no. 3, pp. 200–204, Jan. 2012.

[17] S. Birla, K. Kohli, and A. Dutta, "Machine learning on imbalanced data in credit risk," in *Proc. IEEE 7th Annu. Inf. Technol., Electron. Mobile Commun. Conf. (IEMCON)*, Oct. 2016, pp. 1–6.

[18] L. Wang, "Imbalanced credit risk prediction based on SMOTE and multi-kernel FCM improved by particle swarm optimization," *Appl. Soft Comput.*, vol. 114, Jan. 2022, Art. no. 108153.

[19] I. Emmanuel, Y. Sun, and Z. Wang, "A machine learning-based credit risk prediction engine system using a stacked classifier and a filter-based feature selection method," *J. Big Data*, vol. 11, p. 23, Feb. 2024.

[20] L. Zhu, D. Qiu, D. Ergu, C. Ying, and K. Liu, "A study on predicting loan default based on the random forest algorithm," *Proc. Comput. Sci.*, vol. 162, pp. 503–513, Jan. 2019.

[21] L. Tang, F. Cai, and Y. Ouyang, "Applying a nonparametric random forest algorithm to assess the credit risk of the energy industry in China," *Technol. Forecasting Social Change*, vol. 144, pp. 563–572, Jul. 2019.

[22] F. Shen, X. Zhao, G. Kou, and F. E. Alsaadi, "A new deep learning ensemble credit risk evaluation model with an improved synthetic minority oversampling technique," *Appl. Soft Comput.*, vol. 98, Jan. 2021, Art. no. 106852.

[23] Y. Li, "Credit risk prediction based on machine learning methods," in *Proc. 14th Int. Conf. Comput. Sci. Educ. (ICCSE)*, Aug. 2019, pp. 1011–1013.

[24] W. Yang and L. Gao, "A study on RB-XGBoost algorithm-based e-commerce credit risk assessment model," *J. Sensors*, vol. 2021, no. 1, 2021, Art. no. 7066304.

[25] M. Mahbobi, S. Kimiagari, and M. Vasudevan, "Credit risk classification: An integrated predictive accuracy algorithm using artificial and deep neural networks," *Ann. Oper. Res.*, vol. 330, nos. 1–2, pp. 609–637, Nov. 2023.

[26] A. Gicic and D. Donko, "Proposal of a model for credit risk prediction based on deep learning methods and SMOTE techniques for imbalanced dataset," in *Proc. XXIX Int. Conf. Inf., Commun. Autom. Technol. (ICAT)*, Jun. 2023, pp. 1–6.

[27] M. R. Machado and S. Karray, "Assessing credit risk of commercial customers using hybrid machine learning algorithms," *Expert Syst. Appl.*, vol. 200, Aug. 2022, Art. no. 116889.

[28] F. M. Talaat, A. Aljadani, M. Badawy, and M. Elhosseini, "Toward interpretable credit scoring: Integrating explainable artificial intelligence with deep learning for credit card default prediction," *Neural Comput. Appl.*, vol. 36, no. 9, pp. 4847–4865, Mar. 2024.

[29] C. Rudin and Y. Shaposhnik, "Globally-consistent rule-based summary-explanations for machine learning models: Application to credit-risk evaluation," *J. Mach. Learn. Res.*, vol. 24, no. 16, pp. 1–44, 2023.

[30] N. Nordin, Z. Zainol, M. H. Mohd Noor, and L. F. Chan, "An explainable predictive model for suicide attempt risk using an ensemble learning and Shapley additive explanations (SHAP) approach," *Asian J. Psychiatry*, vol. 79, Jan. 2023, Art. no. 103316.

[31] Y. Gebreyesus, D. Dalton, S. Nixon, D. De Chiara, and M. Chinnici, "Machine learning for data center optimizations: Feature selection using Shapley additive exPlanation (SHAP)," *Future Internet*, vol. 15, no. 3, p. 88, Feb. 2023.

[32] P. S. Palar, L. R. Zuhal, and K. Shimoyama, "Enhancing the explainability of regression-based polynomial chaos expansion by Shapley additive explanations," *Rel. Eng. Syst. Saf.*, vol. 232, Apr. 2023, Art. no. 109045.

[33] L. Antwarg, R. M. Miller, B. Shapira, and L. Rokach, "Explaining anomalies detected by autoencoders using Shapley additive explanations," *Expert Syst. Appl.*, vol. 186, Dec. 2021, Art. no. 115736.

[34] Y. Wang, Y. Zhang, M. Liang, R. Yuan, J. Feng, and J. Wu, "National student loans default risk prediction: A heterogeneous ensemble learning approach and the SHAP method," *Comput. Educ., Artif. Intell.*, vol. 5, Jan. 2023, Art. no. 100166.

[35] M. A. M. Hassan, U. M. Mansur, R. Jha, F. H. Fahim, and T. Mahesh, "Interpretable machine learning models for credit risk assessment," in *Proc. 11th Int. Conf. Comput. Sustain. Global Develop. (INDIACom)*, Feb./Mar. 2024, pp. 361–365.

[36] H. Hofmann, "Statlog (German credit data)," UCI Mach. Learn. Repository, 1994, doi: 10.24432/C5NC77.

[37] R. Quinlan, "Statlog (Australian credit approval)," UCI Mach. Learn. Repository, doi: 10.24432/C59012.

[38] I. D. Mienye and N. Jere, "A survey of decision trees: Concepts, algorithms, and applications," *IEEE Access*, vol. 12, pp. 86716–86727, 2024.

[39] I. D. Mienye and N. Jere, "Optimized ensemble learning approach with explainable AI for improved heart disease prediction," *Information*, vol. 15, no. 7, p. 394, Jul. 2024.

[40] R. Y. Zou and M. Schonlau, "The random forest algorithm for statistical learning," *Stata J.*, vol. 20, no. 3, pp. 3–29, 2018.

[41] Z. Xu, D. Shen, T. Nie, and Y. Kou, "A hybrid sampling algorithm combining M-SMOTE and ENN based on random forest for medical imbalanced data," *J. Biomed. Informat.*, vol. 107, Jul. 2020, Art. no. 103465.

[42] W. Lin, Z. Wu, L. Lin, A. Wen, and J. Li, "An ensemble random forest algorithm for insurance big data analysis," *IEEE Access*, vol. 5, pp. 16568–16575, 2017.

[43] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 785–794.

[44] I. D. Mienye and Y. Sun, "A survey of ensemble learning: Concepts, algorithms, applications, and prospects," *IEEE Access*, vol. 10, pp. 99129–99149, 2022.

[45] I. D. Mienye and Y. Sun, "Effective feature selection for improved prediction of heart disease," in *Proc. Pan-Afr. Artif. Intell. Smart Syst. Conf.* Cham, Switzerland: Springer, 2021, pp. 94–107.

[46] C. Ying, M. Qi-Guang, L. Jia-Chen, and G. Lin, "Advance and prospects of AdaBoost algorithm," *Acta Automatica Sinica*, vol. 39, no. 6, pp. 745–758, 2013.

[47] C. Chen, Q. Zhang, Q. Ma, and B. Yu, "LightGBM-PPI: Predicting protein–protein interactions through LightGBM with multi-information fusion," *Chemometric Intell. Lab. Syst.*, vol. 191, pp. 54–64, Aug. 2019.

[48] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2017, pp. 1–9.

[49] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–9.

[50] E. Carrizosa, C. Molero-Río, and D. Romero Morales, "Mathematical optimization in classification and regression trees," *TOP*, vol. 29, no. 1, pp. 5–33, Apr. 2021.

[51] R. J. Marshall, "The use of classification and regression trees in clinical epidemiology," *J. Clin. Epidemiol.*, vol. 54, no. 6, pp. 603–609, Jun. 2001.

[52] R. K. Zimmerman, G. K. Balasubramani, M. P. Nowalk, H. Eng, L. Urbanski, M. L. Jackson, L. A. Jackson, H. Q. McLean, E. A. Belongia, A. S. Monto, R. E. Malosh, M. Gaglani, L. Clipper, B. Flannery, and S. R. Wisniewski, "Classification and regression tree (CART) analysis to predict influenza in primary care patients," *BMC Infectious Diseases*, vol. 16, p. 503, Sep. 2016.

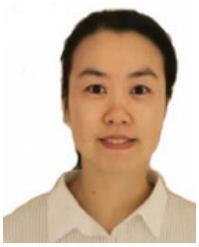
[53] B. Choubin, G. Zehtabian, A. Azareh, E. Rafiei-Sardooi, F. Sajedi-Hosseini, and Ã. KiÅŶi, "Precipitation forecasting using classification and regression trees (CART) model: A comparative study of different approaches," *Environ. Earth Sci.*, vol. 77, p. 314, Apr. 2018.

[54] J. C. Stoltzfus, "Logistic regression: A brief primer," *Academic Emergency Med.*, vol. 18, no. 10, pp. 1099–1104, Oct. 2011.

[55] P. C. Austin and S. van Buuren, "Logistic regression vs. predictive mean matching for imputing binary covariates," *Stat. Methods Med. Res.*, vol. 32, no. 11, pp. 2172–2183, Nov. 2023.

[56] I. D. Mienye, P. Kenneth Ainah, I. D. Emmanuel, and E. Esenogho, "Sparse noise minimization in image classification using genetic algorithm and DenseNet," in *Proc. Conf. Inf. Commun. Technol. Soc. (ICTAS)*, Mar. 2021, pp. 103–108.
[57] M. Lin, X. Zhu, T. Hua, X. Tang, G. Tu, and X. Chen, "Detection of ionospheric scintillation based on XGBoost model improved by SMOTE-ENN technique," *Remote Sens.*, vol. 13, no. 13, p. 2577, Jul. 2021.
[58] I. D. Mienye and Y. Sun, "A deep learning ensemble with data resampling for credit card fraud detection," *IEEE Access*, vol. 11, pp. 30628–30638, 2023.
[59] P. A. Moreno-Sanchez, "Features importance to improve interpretability of chronic kidney disease early diagnosis," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2020, pp. 3786–3792.
[60] M. Carletti, M. Terzi, and G. A. Susto, "Interpretable anomaly detection with DIFFI: Depth-based feature importance of isolation forest," *Eng. Appl. Artif. Intell.*, vol. 119, Mar. 2023, Art. no. 105730.
[61] A. Moncada-Torres, M. C. van Maaren, M. P. Hendriks, S. Siesling, and G. Geleijnse, "Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival," *Sci. Rep.*, vol. 11, p. 6968, Mar. 2021.
[62] Y. Nohara, K. Matsumoto, H. Soejima, and N. Nakashima, "Explanation of machine learning models using Shapley additive explanation and application for real data in hospital," *Comput. Methods Programs Biomed.*, vol. 214, Feb. 2022, Art. no. 106584.
[63] C. Yang, M. Chen, and Q. Yuan, "The application of XGBoost and SHAP to examining the factors in freight truck-related crashes: An exploratory analysis," *Accident Anal. Prevention*, vol. 158, Aug. 2021, Art. no. 106153.
[64] Y. Wu and Y. Zhou, "Hybrid machine learning model and Shapley additive explanations for compressive strength of sustainable concrete," *Construct. Building Mater.*, vol. 330, May 2022, Art. no. 127298.
[65] N. S. Morales and I. C. Fernández, "Land-cover classification using MaxEnt: Can we trust in model quality metrics for estimating classification accuracy?" *Entropy*, vol. 22, no. 3, p. 342, Mar. 2020.
[66] G. Obaido, B. Ogbuokiri, I. D. Mienye, and S. M. Kasongo, "A voting classifier for mortality prediction post-thoracic surgery," in *Proc. Int. Conf. Intell. Syst. Design Appl.* Cham, Switzerland: Springer, 2022, pp. 263–272.
[67] S. J. Kim, K. J. Cho, and S. Oh, "Development of machine learning models for diagnosis of glaucoma," *PLoS ONE*, vol. 12, no. 5, May 2017, Art. no. e0177726.
[68] J.-H. Wu, T. Y. A. Liu, W.-T. Hsu, J. H.-C. Ho, and C.-C. Lee, "Performance and limitation of machine learning algorithms for diabetic retinopathy screening: Meta-analysis," *J. Med. Internet Res.*, vol. 23, no. 7, Jul. 2021, Art. no. e23863.
[69] I. D. Mienye, G. Obaido, K. Aruleba, and O. A. Dada, "Enhanced prediction of chronic kidney disease using feature selection and boosted classifiers," in *Proc. Int. Conf. Intell. Syst. Design Appl.* Cham, Switzerland: Springer, 2021, pp. 527–537.
[70] R. Ghorbani and R. Ghousi, "Comparing different resampling methods in predicting students' performance using machine learning techniques," *IEEE Access*, vol. 8, pp. 67899–67911, 2020.
[71] G. Obaido, O. Achilonu, B. Ogbuokiri, C. S. Amadi, L. Habeebullahi, T. Ohalloran, C. W. Chukwu, E. D. Mienye, M. Aliyu, O. Fasawe, I. A. Modupe, E. J. Omietimi, and K. Aruleba, "An improved framework for detecting thyroid disease using filter-based feature selection and stacking ensemble," *IEEE Access*, vol. 12, pp. 89098–89112, 2024.
[72] S. Mangalathu, S.-H. Hwang, and J.-S. Jeon, "Failure mode and effects analysis of RC members based on machine-learning-based Shapley additive exPlanations (SHAP) approach," *Eng. Struct.*, vol. 219, Sep. 2020, Art. no. 110927.
[73] H. Errousso, E. A. Abdellaoui Alaoui, S. Benhadou, and H. Medromi, "Exploring how independent variables influence parking occupancy prediction: Toward a model results explanation with SHAP values," *Prog. Artif. Intell.*, vol. 11, no. 4, pp. 367–396, Dec. 2022.
[74] T. N. Pandey, A. K. Jagadev, S. K. Mohapatra, and S. Dehuri, "Credit risk analysis using machine learning classifiers," in *Proc. Int. Conf. Energy, Commun., Data Anal. Soft Comput. (ICECDS)*, Aug. 2017, pp. 1850–1854.
[75] Z. Hassani, M. Alambardar Meybodi, and V. Hajihashemi, "Credit risk assessment using learning algorithms for feature selection," *Fuzzy Inf. Eng.*, vol. 12, no. 4, pp. 529–544, Oct. 2020.
[76] D. Tripathi, D. R. Edla, A. Bablani, A. K. Shukla, and B. R. Reddy, "Experimental analysis of machine learning methods for credit score classification," *Prog. Artif. Intell.*, vol. 10, no. 3, pp. 217–243, Sep. 2021.
[77] A. Gicić, D. Donko, and A. Subasi, "Intelligent credit scoring using deep learning methods," *Concurrency Comput., Pract. Exper.*, vol. 35, Feb. 2023, Art. no. e7637.
[78] E. Esenogho, I. D. Mienye, T. G. Swart, K. Aruleba, and G. Obaido, "A neural network ensemble with feature engineering for improved credit card fraud detection," *IEEE Access*, vol. 10, pp. 16400–16407, 2022.
[79] O. Koc, O. Ugur, and A. S. Kestel, "The impact of feature selection and transformation on machine learning methods in determining the credit scoring," 2023, *arXiv:2303.05427*.
[80] X. Han, R. Cui, Y. Lan, Y. Kang, J. Deng, and N. Jia, "A Gaussian mixture model based combined resampling algorithm for classification of imbalanced credit data sets," *Int. J. Mach. Learn. Cybern.*, vol. 10, no. 12, pp. 3687–3699, Dec. 2019.
[81] D. Tripathi, R. Cheruku, and A. Bablani, "Relative performance evaluation of ensemble classification with feature reduction in credit scoring datasets," in *Advances in Machine Learning and Data Science* (Advances in Intelligent Systems and Computing). Singapore: Springer 2018, pp. 293–304.
[82] G. Wang, J. Ma, L. Huang, and K. Xu, "Two credit scoring models based on dual strategy ensemble trees," *Knowl.-Based Syst.*, vol. 26, pp. 61–68, Feb. 2012.
[83] A. Lawi, F. Aziz, and S. Syarif, "Ensemble GradientBoost for increasing classification accuracy of credit scoring," in *Proc. 4th Int. Conf. Comput. Appl. Inf. Process. Technol. (CAIPT)*, Aug. 2017, pp. 1–4.
[84] S. R. Lenka, S. K. Bisoy, R. Priyadarshini, and M. Sain, "Empirical analysis of ensemble learning for imbalanced credit scoring datasets: A systematic review," *Wireless Commun. Mobile Comput.*, vol. 2022, no. 1, Jun. 2022, Art. no. 6584352.
[85] X. Dastile and T. Celik, "Making deep learning-based predictions for credit scoring explainable," *IEEE Access*, vol. 9, pp. 50426–50440, 2021.
[86] D. Tripathi, A. K. Shukla, B. R. Reddy, G. S. Bopche, and D. Chandramohan, "Credit scoring models using ensemble learning and classification approaches: A comprehensive survey," *Wireless Pers. Commun.*, vol. 123, no. 1, pp. 785–812, Mar. 2022.
[87] D. Tripathi, D. R. Edla, V. Kuppili, and A. Bablani, "Evolutionary extreme learning machine with novel activation function for credit scoring," *Eng. Appl. Artif. Intell.*, vol. 96, Nov. 2020, Art. no. 103980.

IDOWU ARULEBA received the B.Sc. degree (Hons.) in computer science from Joseph Ayo Babalola University, in 2017, and the M.Sc. degree in computer science from the University of KwaZulu-Natal, in 2021. He is currently pursuing the Ph.D. degree with the Department of Electrical and Electronic Engineering Science, University of Johannesburg. His research interests include machine learning and deep learning for healthcare applications.



YANXIA SUN (Senior Member, IEEE) received the D.Tech. degree in electrical engineering from the Tshwane University of Technology, South Africa, and the Ph.D. degree in computer science from the University Paris-EST, France, in 2012. She is currently a Professor with the Department of Electrical and Electronic Engineering Science, University of Johannesburg, South Africa. Her research interests include renewable energy, evolutionary optimization, neural networks, non-linear dynamics, and control systems.