

APPLIED RESEARCH

A Deep Learning Ensemble With Data Resampling for Credit Card Fraud Detection

IBOMOIYE DOMOR MIENYE^{ID}, (Member, IEEE), AND YANXIA SUN^{ID}, (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 the South African National Research Foundation under Grant 120106 and Grant 132797, and in part by the South African National Research Foundation Incentive under Grant 132159.

ABSTRACT Credit cards play an essential role in today's digital economy, and their usage has recently grown tremendously, accompanied by a corresponding increase in credit card fraud. Machine learning (ML) algorithms have been utilized for credit card fraud detection. However, the dynamic shopping patterns of credit card holders and the class imbalance problem have made it difficult for ML classifiers to achieve optimal performance. In order to solve this problem, this paper proposes a robust deep-learning approach that consists of long short-term memory (LSTM) and gated recurrent unit (GRU) neural networks as base learners in a stacking ensemble framework, with a multilayer perceptron (MLP) as the meta-learner. Meanwhile, the hybrid synthetic minority oversampling technique and edited nearest neighbor (SMOTE-ENN) method is employed to balance the class distribution in the dataset. The experimental results showed that combining the proposed deep learning ensemble with the SMOTE-ENN method achieved a sensitivity and specificity of 1.000 and 0.997, respectively, which is superior to other widely used ML classifiers and methods in the literature.

INDEX TERMS Credit card, deep learning, ensemble learning, fraud detection, machine learning, neural network.

I. INTRODUCTION

Information technology advancements have significantly impacted the financial sector, leading to the broad adoption of electronic commerce (e-commerce) platforms. Also, the recent outbreak of the novel coronavirus (COVID-19) pandemic has further shown the need for a more digital world and further expanded the e-commerce industry [1], [2]. One of the major issues associated with modern e-commerce is the high cases of credit card fraud [3]. Also, in the last decade, there has been an increase in credit card fraud, which is a huge burden on financial institutions [4]. The increased credit card fraud rate is associated with the expansion of e-commerce and increased online transactions. Therefore, credit card fraud detection (CCFD) is crucial for financial companies to avoid losses.

Artificial intelligence (AI) and machine learning applications in the financial sector can produce excellent results for companies, such as improved efficiency, reduced opera-

tional cost, and enhanced customer satisfaction [5]. Several ML-based systems have been developed to detect credit card fraud. For example, Malik et al. [6] studied the use of hybrid models in CCFD. The hybrid models were achieved by combining a variety of ML algorithms, including extreme gradient boosting (XGBoost), random forest, adaptive boosting (AdaBoost), and light gradient boosting machine (LGBM). The experimental results indicated that the hybrid model based on AdaBoost and LGBM obtained the best classification performance. In a similar research work, Alfaiz and Fati [7] conducted a performance evaluation of ML classifiers and data resampling techniques for detecting credit card fraud. The classifiers used in the study include LGBM, XGBoost, random forest, categorical boosting (CatBoost), logistic regression, and naïve Bayes. The results indicated that the CatBoost classifier integrated with a k-nearest neighbor-based undersampling technique performed better than the other methods.

Meanwhile, building robust machine learning-based CCFD models has remained a challenge for some reasons. Firstly, conventional classifiers make predictions based on the

The associate editor coordinating the review of this manuscript and approving it for publication was Massimo Cafaro^{ID}.

transaction details only, such as amount, transaction country, and transaction type, ignoring the sequence of transactions that defines the clients' shopping behaviour, which is useful in identifying appropriate fraud patterns [8], [9]. Secondly, credit card fraud datasets are highly imbalanced since genuine transactions significantly outnumber fraudulent transactions [10]. Imbalance classification is a predictive modelling problem where there is an uneven distribution of samples across the classes [11]. The class that makes up a large proportion of the dataset is called the majority class, while the class with a smaller proportion is called the minority class. Imbalance classification is a challenge because most ML algorithms were designed with the assumption of an even class distribution. Therefore, using imbalanced data such as the credit card dataset results in models with poor classification performance, especially for the minority class, i.e., fraudulent transactions. Furthermore, correctly identifying the minority class samples is of utmost importance in imbalance classification problems [12].

Deep learning (DL) and ensemble learning have recently dominated the ML field [13], [14], [15], [16], achieving excellent prediction performances in complex problems, and they could be applied to solve the challenges in credit card fraud detection. Deep learning, a subset of machine learning, is mainly a neural network with multiple layers [17]. Deep learning models using recurrent neural networks (RNN) have been employed for different sequential modelling-based ML tasks [18], [19], [20]. For example, Shen et al. [21] noted that algorithms that utilize sequential modelling, such as RNNs, usually perform better than conventional ML models. Meanwhile, simple RNN-based models are prone to the vanishing gradient problem, a situation where the RNN is unable to propagate relevant gradient information from the model's output end back to the layers near the input end [22]. However, LSTM and GRU-based RNNs were proposed to solve the vanishing gradient problem and have shown good performances in different sequence classification tasks [8], [23], [24].

Meanwhile, ensemble learning involves training multiple base classifiers and combining their outputs to obtain better performance than the single base classifiers. Some ML algorithms build models with low accuracy, high bias, or high variance [25]. Ensemble learning-based classifiers tend to outperform single classifiers [26]. However, DL-based ensemble models have rarely been employed for CCFD. Therefore, to fill this research gap and solve the challenges facing CCFD, this paper considers the CCFD as a sequence classification problem and uses the LSTM and GRU networks as base learners in a stacking ensemble model. Also, most ensemble learning-based CCFD studies employed voting-based methods. Hence, to contribute to existing literature, this study uses the MLP neural network as the meta-learner in the stacking ensemble. The proposed method uses the advantages of sequential modelling and ensemble learning to achieve enhanced CCFD.

Furthermore, this study uses the SMOTE-ENN method for solving the imbalance class problem. Hybrid resampling techniques such as SMOTE-ENN have shown enhanced effectiveness in dealing with imbalanced data compared to oversampling and undersampling techniques [27], [28], [29]. In SMOTE-ENN, the SMOTE step adds minority class samples to the dataset. Meanwhile, SMOTE obtains synthetic samples via linear interpolation between majority class samples in the neighbourhood and often generates noisy samples [30]. Therefore, it is essential to eliminate such noisy samples, and in the hybrid SMOTE-ENN, the ENN removes noisy samples. Meanwhile, the main contributions of this work include the following:

- An improved credit card fraud detection approach is proposed using the LSTM and GRU as base learners and MLP as a meta-learner in a stacking ensemble model.
- Improving the credit card fraud detection rate by combining the robustness of ensemble learning and deep learning.
- Overcoming the class imbalance problem in credit card datasets using the SMOTE-ENN method.
- Constructing the credit card fraud detection model using recurrent neural networks (i.e., LSTM and GRU), capable of learning client's spending behaviour and transaction sequences.
- A comparative study with recently proposed credit card fraud detection methods is conducted

The rationale behind the proposed approach is that considering the CCFD as a sequential modelling problem could better detect slight differences in genuine clients' spending behaviour. The main objective of the research is to build a well-performing CCFD model using the SMOTE-ENN data resampling technique and deep learning-based stacking ensemble. The following classifiers would be implemented and used as the baseline for performance comparison: LSTM, GRU, MLP, AdaBoost, and random forest.

The remainder of the paper is organized as follows: Section II presents a literature review, while Section III presents the dataset and the various algorithms used in the study. Section IV introduces the proposed CCFD approach. Section V presents and discusses the experimental results, and the paper is concluded in Section VI.

II. RELATED WORKS

Several machine learning methods have been proposed for credit card fraud detection [31], [32]. Specifically, supervised learning algorithms have shown to be highly effective in detecting credit card fraud, where labelled datasets containing previous transaction records are utilized to build machine learning models that can detect new fraudulent transactions. These supervised learning algorithms include logistic regression [33], support vector machines (SVM) [34], decision trees [35], adaptive boosting (AdaBoost) [36], random forest [37], and artificial neural networks (ANN) [38], [39], [40].

Meanwhile, numerous researchers have proposed different methods to improve the performance of machine learning algorithms for credit card fraud detection. For example, Taha and Malebary [41] proposed an optimized light gradient boosting machine (LightGBM) to detect fraud in credit card transactions. The method involves optimizing the hyperparameters of the LightGBM using a Bayesian-based algorithm. The authors used two credit card datasets in the study. They performed a comparative analysis with other classifiers such as SVM, decision tree, random forest, naïve Bayes, and categorical boosting (CatBoost). The optimized LightGBM performed better than the benchmarked classifiers, having obtained an accuracy of 98.40%, a precision of 97.34%, and an area under the receiver operating characteristic curve (AUC) of 92.88%.

In another research, Ileberi et al. [42] proposed a credit card fraud detection model based on feature selection using a genetic algorithm (GA). The selected features were then used to train different ML models using naïve Bayes, logistic regression, decision tree, ANN, and random forest algorithms. The study showed that the GA-based feature selection enhanced the performance of the various ML classifiers, and the random forest achieved the highest accuracy of 99.98%.

Furthermore, Salekshahrezaee et al. [43] aimed to develop robust CCFD models by studying the impact of feature extraction and data resampling on the following machine learning classifiers: CatBoost, random forest, extreme gradient boosting (XGBoost), and LightGBM. The feature extraction was achieved using principal component analysis (PCA) and convolutional autoencoder (CAE), while the data resampling methods include SMOTE, random undersampling (RUS), and SMOTE Tomek techniques. The experimental results indicated a significant increase in the performance of the classifiers when the RUS and CAE techniques were used for resampling and feature extraction, respectively.

Meanwhile, the class imbalance in most credit card datasets has made fraud detection difficult [44], [45]. In imbalanced datasets, the number of transactions labelled as legitimate (majority class) is significantly higher than those labelled as fraud (minority class). Therefore, conventional machine learning algorithms tend to underperform, especially in identifying fraud cases [9], [46]. Also, these conventional ML algorithms build CCFD models using individual transaction details, such as transaction location and amount, without considering the sequential information associated with the credit card clients [8].

Hence, deep learning techniques such as RNNs that consider the client's shopping behaviour and transaction sequences can be vital in detecting important fraud patterns [8], [47], [48]. For example, Benchaji et al. [8] proposed a credit card fraud detection method using an LSTM network to achieve sequential modelling and ensure improved fraud detection. The LSTM network was coupled with a uniform manifold approximation and projection (UMAP) technique to identify the most relevant attributes and an attention mechanism to improve the performance of the LSTM. The experi-

mental results demonstrated that the proposed approach was robust in detecting credit card fraud.

However, deep learning-based ensemble models have rarely been employed for credit card fraud detection, even though combining deep learning-based techniques such as LSTM in an ensemble model could result in more robust models. Therefore, this paper aims to use LSTM and GRU networks to achieve sequential modelling and enhance the identification of fraudulent transactions. Meanwhile, to ensure the classification task is significantly improved, the LSTM and GRU networks would be employed as base learners in a stacking ensemble model, with an MLP classifier as the meta-learner. Additionally, the SMOTE-ENN technique would be employed to overcome the class imbalance problem and ensure effective machine learning.

III. MATERIALS AND METHODS

This section describes the credit card dataset used in the study and provides a detailed explanation of the various algorithms and methods used in formulating the proposed credit card fraud detection method.

A. CREDIT CARD DATASET

This study uses a credit card dataset containing transactions performed by European cardholders in September 2013, which is publicly available [49]. It contains 283,807 transactions, among which 492 transactions are labelled as fraudulent. The dataset is highly imbalanced, with only 0.172% labelled as fraudulent transactions. Most of the features were transformed to numerical variables using principal component analysis (PCA) because of confidentiality issues, and the names of the features were anonymized as V1, V2, V3, . . . , and V28, excluding the "Time" and "Amount" features. The "Class" feature is the target variable, and it has values 1 and 0, representing fraud and non-fraud transactions, respectively.

B. MULTILAYER PERCEPTRON

Multilayer perceptron is a type of feedforward neural network comprising three layers, including the input layer, hidden layer, and output layer. It is a powerful neural network with applications in several domains [50], [51], [52]. In the MLP network, data flows from the input to the output layer. The hidden layer, placed between the input and output layers, is the core of the MLP, which processes the input information and transfers it to the output layer [53]. The neurons are the processing elements in the MLP, and the neurons in each layer are connected to every neuron in the next layer. The input layer feeds the network with the input variables, and subsequent layers receive their inputs from the output of the previous layers [54]. The MLP network is usually trained using the backpropagation algorithm [55], enabling the network to update its weights to minimize the output error. The mean squared error (MSE) is the commonly used error

function, and it is represented as:

$$E = \frac{1}{2} \sum_{i=1}^n \|p_i - t_i\|^2 \quad (1)$$

where n is the number of data points, and p_i and t_i are the predicted output and target output for sample i , respectively. Meanwhile, this layer-to-layer transfer of information is achieved using activation functions, such as the sigmoid function $\sigma(k) = \frac{1}{(1+e^{(-k)})}$, where e is Euler's number [53], [56].

C. LONG SHORT-TERM MEMORY

The LSTM network is a modified form of a recurrent neural network. Unlike conventional neural networks like MLP, RNNs are not limited to a unidirectional data flow [57]. They can loop through several layers and temporarily memorize information that can be used later. Meanwhile, the simple RNN is susceptible to the vanishing gradient problem, and the LSTM and GRU were developed to solve the problem [58]. The LSTM can learn long-term dependencies, making it suitable for classifying sequential data, such as credit card data. LSTM networks consist of a memory cell c_t , with an input gate i_t , a forget gate f_t , and an output gate o_t . The three gates control how the data is processed and used [9]. The following mathematical formulations represent the flow of information within the LSTM layers:

$$i_t = \sigma(V_i x_t + W_i h_{t-1} + b_i) \quad (2)$$

$$f_t = \sigma(V_f x_t + W_f h_{t-1} + b_f) \quad (3)$$

$$\tilde{c}_t = \tanh(V_c x_t + W_c h_{t-1} + b_c) \quad (4)$$

$$c_t = f_t \otimes c_{(t-1)} + i_t \otimes \tilde{c}_t \quad (5)$$

$$o_t = \sigma(V_o x_t + W_o h_{t-1} + b_o) \quad (6)$$

$$h_t = o_t \otimes \tanh(c_t) \quad (7)$$

where V_* , W_* , and b_* are learnable parameters, h_* is the hidden state, where $*$ is used in place of f , i , o , or c to represent the given gates and memory cell. Meanwhile, σ and \tanh are the sigmoid and tanh activation functions and \otimes is the element-wise product [59].

D. GATED RECURRENT UNIT

The gated recurrent unit, developed by Cho et al. [60], is a gating mechanism in recurrent neural networks similar to the LSTM network but without an output gate. The GRU replaces the three gates in LSTM with two, i.e., the update gate z_t and reset gate r_t . The update gate and reset gate control information that flows into memory and information that flows out of memory, respectively [61]. These gates are basically vectors that determine the information passed on to the output and can be trained to keep past information or discard unnecessary information that does not contribute to the prediction. The gates in GRU are given sigmoid activations, thereby ensuring their values are in the range (0,1), which could be "fraud" and "non-fraud" classifications. Furthermore, in the GRU, the hidden state h_t and cell state c_t blend into one, i.e., $h_t = c_t$.

The GRU update equations include the following:

$$r_t = \sigma(V_r x_t + W_r h_{(t-1)} + b_r) \quad (8)$$

$$z_t = \sigma(V_z x_t + W_z h_{(t-1)} + b_z) \quad (9)$$

$$\tilde{c}_t = \tanh(V_c x_t + W_c (r_t \otimes h_{(t-1)}) + b_c) \quad (10)$$

$$c_t = (1 - z_t) \otimes c_{(t-1)} + z_t \otimes \tilde{c}_t \quad (11)$$

$$h_t = c_t \quad (12)$$

where V_r , W_r , V_z , and W_z represent the weight matrices, b_r and b_z denotes the bias vectors [62].

E. ENSEMBLE LEARNING

Ensemble learning is a machine learning approach that combines multiple algorithms to achieve better classification performance than the individual base models [63] [64]. ML models usually have shortcomings, such as high bias, high variance, and low accuracy, and are not exempted from making errors [25], [65]. Therefore, rather than relying on one classifier, ensemble learning methods harness the strengths of two or more classifiers and often obtain higher accuracy than the individual base classifiers [66], [67], [68].

Ensemble learning methods can be broadly grouped into bagging, boosting, and stacking [69]. Stacking, which is the focus of this paper, involves using different base algorithms to build models, termed level-0 models, and a different algorithm called a meta-learner (or level-1 classifier) is trained to combine the predictions of the base models [70]. The trained level-0 models are tested with out-of-sample instances, and the predicted class labels, combined with the actual labels, make up the dependent and independent variables in the new dataset employed for training the meta-classifier [71].

Unlike bagging and boosting, which uses combination rules such as majority voting and weighted majority voting, the stacking ensemble model uses another ML algorithm (i.e., meta-learner) to aggregate the predictions from the level-0 models [72]. In the literature, stacking-based ensembles have been applied in diverse fields [73], [74], [75].

IV. PROPOSED CCFD METHOD

A. DATA RESAMPLING

Only 0.172% of the samples in the European credit card dataset are labelled as fraudulent. Therefore, the dataset is highly imbalanced, which usually leads to models with poor generalization ability. Both oversampling and undersampling methods have been widely utilized to solve the class imbalance challenge in different ML applications [76], [77], [78]. However, oversampling methods create balanced training sets by duplicating samples in the minority class, which could result in overfitting [79], while undersampling methods obtain balanced datasets by discarding selected majority class instances. Hence, undersampling could remove useful examples that might be crucial in building efficient ML models, and it is also inefficient in highly imbalanced datasets like the European credit card dataset.

Therefore, the SMOTE-ENN method is adopted to balance the credit card dataset in this study. The SMOTE-ENN,

summarized in Algorithm 1, is a hybrid method, having both undersampling and oversampling aspects using ENN and SMOTE, respectively [80]. The SMOTE technique oversamples the minority class instances while the ENN deletes overlapping samples. The ENN's neighborhood cleaning approach is utilized to discard samples that vary from two in the three nearest neighbors [81].

Algorithm 1 SMOTE-ENN Resampling Method

Input: training data $S = \{(x_1, y_1), \dots, (x_2, y_2), \dots, (x_m, y_m)\}$

Procedure:

Step 1: Oversampling

- 1) Select an instance x_i randomly from the minority class
- 2) Find the k nearest neighbors of x_i and let S_k represent the samples
- 3) Generate a synthetic data point p by randomly picking one of the samples in S_k called z , then connect p and z to obtain a line segment in the feature space
- 4) Assign the minority class label to p .
- 5) Generate consecutive synthetic instances as a convex combination of p and z .

Step 2: Undersampling

- 1) Select a random instance $x_r \in S$
- 2) Find the k nearest neighbors of x_r , where $k = 3$
- 3) Delete x_r if it has more neighbors from the other class.
- 4) Repeat 6 – 8 for the whole training data.

Output: A balanced dataset for effective CCFD.

B. DEEP LEARNING ENSEMBLE

The proposed CCFD method combines LSTM, GRU, and MLP neural networks to obtain a robust stacking-based ensemble model. The stacking framework comprises two layers, i.e., level-0 and level-1. At level-0, the base classifiers are trained and then tested with out-of-sample instances; the resulting predictions and their actual labels comprise the independent and dependent variables in the new dataset used to train the meta-classifier [71]. In this study, the LSTM and GRU networks are the level-0 learners, while the MLP is the level-1 learner. The main reason behind selecting the LSTM and GRU as the base learners include their robustness in modelling sequential data and their high prediction performance in such tasks. Also, their difference would ensure diversity in the ensemble, which is essential because different base models are likely to make different types of errors. A flowchart of the proposed methodology is shown in Figure 1.

From Figure 1, the proposed approach can be divided into three steps. The first step involves training the base models using the LSTM and GRU networks. Meanwhile, the 10-fold cross-validation (CV) approach is used to develop the ensemble, ensuring there is no data leakage. In the second step, the two base models generate out-of-fold predictions,

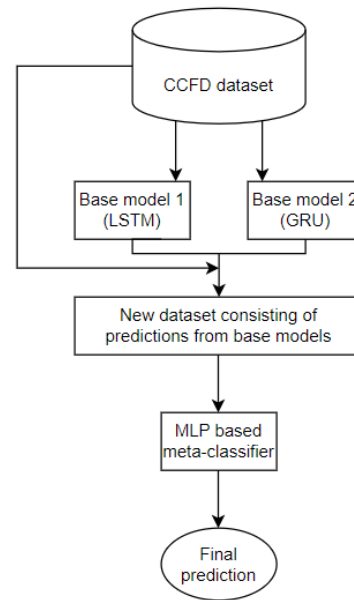


FIGURE 1. Structure of the proposed ensemble.

which are transformed, and, together with the actual labels, form a new dataset. Specifically, the predicted target labels are used as attributes in the new dataset, whereas the original class labels make up the response variable.

Furthermore, the base learners are trained individually on the training data using the same CV indices, while the meta-classifier is trained with the out-of-fold predictions and the actual labels associated with these instances; since they were not used in training the level-0 models. Therefore the output of the base learners is 1 and 0, representing fraud and non-fraud transactions. For instance, assuming each sample in a credit card dataset S is x_i, y_i , a new sample \hat{x}_i, y_i is created, where

$$\hat{x}_i = h_1(x_i), h_2(x_2), \dots, h_T(x_i) \tag{13}$$

where \hat{x}_i is the out-of-sample prediction and $\{h_1, h_2, \dots, h_T\}$ are the base models. The third step involves using the generated dataset to train the MLP-based meta-classifier, which is employed to combine the base models, i.e., assuming x is a test instance, the final ensemble prediction for x is $\hat{h}(h_1(x), h_2(x), \dots, h_T(x))$, where \hat{h} represents the meta-model [82]. The proposed approach is outlined in Algorithm 2.

V. RESULTS AND DISCUSSION

This study develops a deep learning ensemble with data resampling for improved credit card fraud detection. The experimental results have been split into two, i.e., the classifiers' performance before and after data resampling. Meanwhile, the proposed stacking ensemble was achieved using the LSTM and GRU neural networks as base learners and an MLP neural network as the meta-learner. It was implemented together with five other classifiers for performance

Algorithm 2 Proposed DL Ensemble

Input: Credit card dataset $S = (x_1, y_1), \dots, (x_2, y_2), \dots, (x_m, y_m)$
 The selected base learners, i.e., LSTM and GRU
 The meta-learner L , i.e., MLP

Procedure:
 Step 1: Training the level-0 models
for $t = 1, \dots, T$:
 Train a base classifier h_t using S
end for
 Step 2: Generating the new dataset S_N
for $i = 1, \dots, m$:
 Generate a new example \hat{x}_i, y_i using (13)
end for
 Step 3: Fit the MLP-based meta-model \hat{h} (or level-1 model) using S_N
return $H(x) = \hat{h}(h_1(x), h_2(x), \dots, h_T(x))$

Output: Stacking-based DL ensemble classifier H .

comparison. The classifiers include AdaBoost [83], random forest [84], MLP [85], LSTM [86], and GRU [60]. All the models were developed using scikit-learn [87], a well-known ML library in Python, and the hardware used in building the models is a 16 GB RAM windows computer having the following specification: Intel(R) Core(TM) i5-102100U CPU @ 1.60 GHz 2.10 GHz. Meanwhile, the 10-fold cross-validation technique is utilized to evaluate all the models' performance.

Furthermore, the following performance metrics were used to evaluate the models: sensitivity (SEN), specificity (SPE), receiver operating characteristic (ROC) curve, and area under the ROC curve (AUC). Sensitivity indicates the ability of the classifier to predict fraudulent transactions as fraud. A classifier with a high sensitivity score is usually preferred in CCFD [88]. Specificity indicates the ability of the classifier to predict legitimate transactions [89]. These performance metrics are represented mathematically as follows:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (14)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (15)$$

where true positive (TP) and false positive (FP) represent the number of instances correctly predicted and wrongly predicted as fraud; in contrast, true negative (TN) and false negative (FN) denotes the number of transactions correctly and wrongly predicted as non-fraud. Meanwhile, the ROC curve is a plot showing the ability of a classifier to distinguish between the fraud and non-fraud classes. It is obtained by plotting the true positive rate against the false positive rate at different threshold settings [90]. The AUC is a summary of the ROC curve, having a value of 0 to 1, where 0 indicates that all the classifiers' predictions are wrong, and 1 indicates a perfect classifier.

TABLE 1. Experimental results without SMOTE-ENN data resampling.

Classifier	Sensitivity	Specificity	AUC
AdaBoost	0.775	0.930	0.810
Random forest	0.802	0.931	0.820
MLP	0.729	0.910	0.770
LSTM	0.771	0.920	0.810
GRU	0.734	0.915	0.790
Proposed DL Ensemble	0.905	0.966	0.920

A. PERFORMANCE OF THE CLASSIFIERS WITHOUT DATA RESAMPLING

Table 1 shows the prediction performance of the proposed deep learning ensemble and the selected classifiers trained with the original credit card dataset without resampling. The proposed method achieved higher specificity, sensitivity, and AUC scores than the individual classifiers that make up the ensemble (i.e., LSTM, GRU, and MLP). Also, the proposed ensemble performed better than the well-known AdaBoost and random forest classifiers. Specifically, the proposed ensemble achieved the highest sensitivity value of 90.5%, followed by the random forest (80.2%) and AdaBoost (77.5%). A similar trend is observed for the specificity and AUC values, where the proposed ensemble achieved the best performance.

Meanwhile, Table 1 shows that the sensitivity values are lower than the specificity values, which implies the classifiers can correctly identify the majority class samples or non-fraud transactions and fail to correctly predict the minority class samples, i.e. fraud transactions. Hence, it is necessary to solve the class imbalance problem in order to achieve optimal classification performance.

B. PERFORMANCE OF THE CLASSIFIERS AFTER SMOTE-ENN RESAMPLING

Table 2 shows the performance of the classifier after the SMOTE-ENN-based resampling. The results suggest that the performance of the classifiers is high compared to Table 1. Results in Table 2 indicate improved sensitivity and specificity values, with the proposed deep learning ensemble obtaining the best performance. Specifically, the proposed method obtained a sensitivity of 100% and a specificity of 99.7%. The AdaBoost obtained the second highest sensitivity of 96.6%, followed by the GRU (95.2%) and random forest (94.5%). Regarding the specificity value, the AdaBoost achieved 98.4%, followed by the LSTM (97.1%) and random forest (96.1%).

Figure 2 shows the ROC curves of the various models. The proposed ensemble model's ROC curve is at the top-left, indicating it has the best classification performance compared to the benchmarked models. The DL ensemble achieved an AUC of 1.000, which is higher than the other methods. The high AUC indicates the DL ensemble performs better in distinguishing between fraud and non-fraud transactions.

Furthermore, Fig 3 and Fig 4 provides a comparison of the sensitivity and specificity score achieved by the classifiers

TABLE 2. Performance of the classifiers after resampling.

Classifier	Sensitivity	Specificity	AUC
AdaBoost	0.966	0.984	0.980
Random forest	0.945	0.961	0.940
MLP	0.928	0.952	0.930
LSTM	0.940	0.971	0.940
GRU	0.952	0.960	0.950
Proposed DL Ensemble	1.000	0.997	1.000

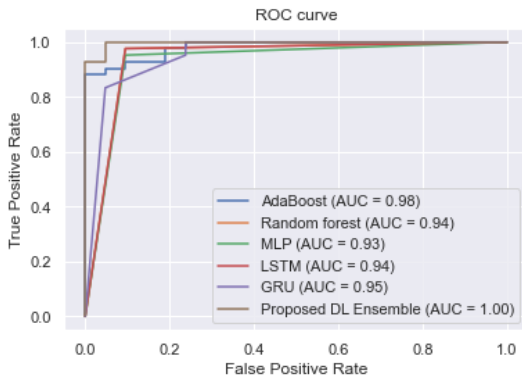


FIGURE 2. ROC curve of the models.

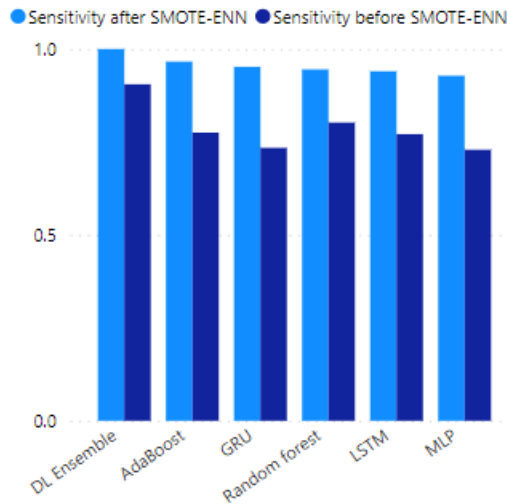


FIGURE 3. Sensitivity before and after data resampling.

before and after the resampling step. The figures demonstrate the impact and importance of the data resampling, as it significantly influenced the learning and prediction performance of the classifiers. Using the balanced dataset ensured the sensitivity or true positive rate is greatly improved, which implies an improved detection of the samples labelled as fraud.

C. COMPARISON WITH STATE-OF-THE-ART METHODS

This section compares the proposed method with well-performing methods in recent scholarly articles, including a weighted extreme learning machine (Weighted ELM) [91],

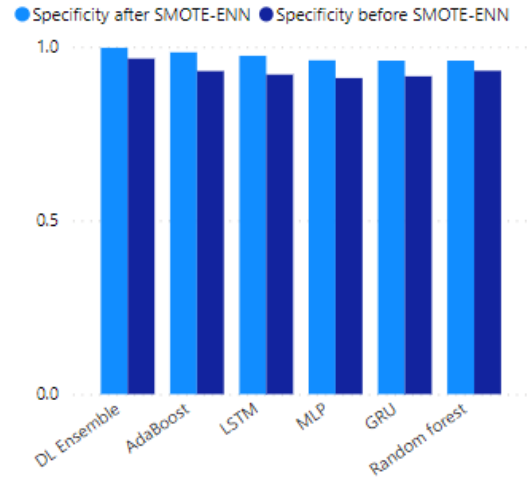


FIGURE 4. Specificity before and after data resampling.

TABLE 3. Performance comparison with other scholarly works.

Reference	Algorithm	SEN	SPE	AUC
Zhu et al. [71]	Weighted ELM	0.982	-	0.978
Taha et al. [72]	Optimized LightGBM	-	-	0.928
Alkhatib et al. [73]	DNN	0.955	-	0.990
Makki et al. [74]	Cost sensitive SVM	0.650	-	0.620
Yotsawat et al. [75]	Neural network ensemble	-	0.936	0.980
Ileberi et al. [76]	GA-RF	0.778	-	0.950
Kalid et al. [77]	C4.5-Naive Bayes	0.872	1.000	-
Feng et al. [78]	Markov Chain-based ensemble	-	-	0.660
Mrozek et al. [79]	RF-SMOTE	0.829	-	0.910
Xie et al. [80]	XGBoost – SMOTE	0.988	-	0.970
Esenogho et al. [9]	LSTM ensemble – SMOTE-ENN	0.996	0.998	0.990
Dang et al. [4]	DNN-ADASYN	0.829	0.998	0.970
Dang et al. [4]	DNN-SMOTE	0.857	0.997	0.952
Ahmad et al. [100]	ANN-RUS	0.818	1.000	0.909
Ahmad et al. [100]	ANN-SBS	0.929	0.960	0.943
Proposed method	Stacking-based DL Ensemble	1.000	0.997	1.000

an optimized LGBM [41], a deep neural network (DNN) based classifier [92], a cost-sensitive SVM [93], a neural network ensemble [94], a random forest-based genetic algorithm wrapper method (GA-RF) [42], a method that sequentially combines the C4.5 and naïve Bayes classifiers [95], a dynamic weighted ensemble technique using Markov Chain [96], a model developed using random forest algorithm and SMOTE based resampling (RF-SMOTE) [97], an XGBoost model with SMOTE based resampling [98], an LSTM ensemble with SMOTE-ENN [9], a comparison of SMOTE and ADASYN based resampling with a DNN classifier [4], and an ANN model with random undersampling (RUS) and similarity-based selection (SBS) resampling techniques [99].

The stacking-based DL ensemble obtained optimal performance in comparison with other well-performing methods in Table 3, reflecting the proposed method’s robustness. Meanwhile, it would be beneficial to observe how the proposed approach would perform using a different dataset. Therefore, the Taiwan credit card dataset [100] is employed to evaluate the proposed method and the selected baseline classifiers. The

TABLE 4. Experimental results before SMOTE-ENN resampling using the Taiwan dataset.

Classifier	Sensitivity	Specificity	AUC
AdaBoost	0.670	0.885	0.690
Random forest	0.705	0.893	0.710
MLP	0.600	0.785	0.630
LSTM	0.631	0.859	0.680
GRU	0.615	0.794	0.630
Proposed DL Ensemble	0.725	0.890	0.740

TABLE 5. Experimental results after SMOTE-ENN resampling using the Taiwan dataset.

Classifier	Sensitivity	Specificity	AUC
AdaBoost	0.860	0.910	0.880
Random forest	0.884	0.925	0.900
MLP	0.817	0.894	0.830
LSTM	0.840	0.902	0.850
GRU	0.774	0.860	0.790
Proposed DL Ensemble	0.930	0.961	0.940

Taiwan dataset is imbalanced, containing 30 000 instances, with 23 364 and 6 636 samples labelled as good and bad clients, respectively. Table 4 and Table 5 show the performance of the classifiers before and after the application of data resampling, and it can be seen that the proposed ensemble obtained superior performance.

D. DISCUSSIONS

Using both datasets, the proposed DL ensemble model performed optimally compared to the other models. Notably, the proposed ensemble outperformed the individual classifiers (GRU, LSTM, and MLP) in all the instances, indicating its robustness. For example, using the European Credit Card dataset, the proposed ensemble obtained sensitivity, specificity, and AUC of 1.000, 0.997, and 1.000, respectively, compared to the performance obtained by GRU (SEN = 0.952, SPE = 0.960, and AUC = 0.950), LSTM (SEN = 0.940, SPE = 0.971, and AUC = 0.940), and MLP (SEN = 0.928, SPE = 0.952, and AUC = 0.930).

Furthermore, the classification performance of the classifiers was enhanced after resampling the data. Using the European credit card dataset, the sensitivity, specificity, and AUC increased by 10.5%, 3.2%, and 8.7%, respectively. Similar performance increases were observed when the models were trained using the Taiwan credit card dataset. Also, the proposed method achieved excellent performance compared to other credit card fraud detection methods in the literature, including those that used resampling techniques, such as oversampling-based methods (SMOTE and ADASYN) and RUS undersampling methods.

From the above, it is fair to conclude that combining the hybrid SMOTE-ENN-based resampling and the stacking-based deep learning ensemble resulted in excellent classification performance. Additionally, the stacking-based deep learning ensemble proposed in this study effectively

learned from the credit card datasets, predicted both fraud and legitimate transactions efficiently, and could be deployed for real-time fraud detection.

VI. CONCLUSION

The growth in e-commerce and credit card usage has led to a significant rise in credit card fraud, affecting financial institutions and customers. Credit card fraud detection is an essential and challenging task. This study contributes to the existing literature by proposing a new approach using a deep learning-based stacking ensemble with data resampling to detect credit card fraud effectively. The stacking ensemble uses LSTM and GRU neural networks as base learners and an MLP as the meta-learner. Meanwhile, the data resampling was achieved using the hybrid SMOTE-ENN method. The proposed method achieved sensitivity, specificity, and AUC values of 1.000, 0.997, and 1.000, respectively, outperforming the baseline classifiers, including AdaBoost, random forest, MLP, LSTM, and GRU. In terms of performance comparison with other scholarly works, the proposed method presents an excellent performance. Future works would aim to introduce diversity in the base models using classifiers with different training methods, such as combining LSTM with random forest, logistic regression, or SVM. Also, future research works could consider feature importance and risk factor analysis.

REFERENCES

- [1] M. Habibpour, H. Gharoun, M. Mehdipour, A. Tajally, H. Asgharnezhad, A. Shamsi, A. Khosravi, M. Shafie-Khah, S. Nahavandi, and J. P. S. Catalao, "Uncertainty-aware credit card fraud detection using deep learning," 2021, *arXiv:2107.13508*.
- [2] A. Cherif, A. Badhib, H. Ammar, S. Alshehri, M. Kalkatawi, and A. Imine, "Credit card fraud detection in the era of disruptive technologies: A systematic review," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 35, no. 1, pp. 145–174, Jan. 2023, doi: 10.1016/j.jksuci.2022.11.008.
- [3] G. Zhang, Z. Li, J. Huang, J. Wu, C. Zhou, and J. Yang, "eFraudCom: An e-commerce fraud detection system via competitive graph neural networks," *ACM Trans. Inf. Syst.*, vol. 40, no. 3, pp. 1–27, Mar. 2022, doi: 10.1145/3474379.
- [4] T. K. Dang, T. C. Tran, L. M. Tuan, and M. V. Tiep, "Machine learning based on resampling approaches and deep reinforcement learning for credit card fraud detection systems," *Appl. Sci.*, vol. 11, no. 21, p. 10004, Oct. 2021, doi: 10.3390/app112110004.
- [5] J. Chaquet-Ulldemolins, F.-J. Gimeno-Blanes, S. Moral-Rubio, S. Muñoz-Romero, and J.-L. Rojo-Álvarez, "On the black-box challenge for fraud detection using machine learning (I): Linear models and informative feature selection," *Appl. Sci.*, vol. 12, no. 7, p. 3328, Mar. 2022, doi: 10.3390/app12073328.
- [6] E. F. Malik, K. W. Khaw, B. Belaton, W. P. Wong, and X. Chew, "Credit card fraud detection using a new hybrid machine learning architecture," *Mathematics*, vol. 10, no. 9, p. 1480, Apr. 2022, doi: 10.3390/math10091480.
- [7] N. S. Alfaiz and S. M. Fati, "Enhanced credit card fraud detection model using machine learning," *Electronics*, vol. 11, no. 4, p. 662, Feb. 2022, doi: 10.3390/electronics11040662.
- [8] I. Benchaji, S. Douzi, B. El Ouahidi, and J. Jaafari, "Enhanced credit card fraud detection based on attention mechanism and LSTM deep model," *J. Big Data*, vol. 8, no. 1, p. 151, Dec. 2021, doi: 10.1186/s40537-021-00541-8.
- [9] E. Ezenogho, 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, doi: 10.1109/ACCESS.2022.3148298.

- [10] E. Btoush, X. Zhou, R. Gururajan, K. Chan, and X. Tao, "A survey on credit card fraud detection techniques in banking industry for cyber security," in *Proc. 8th Int. Conf. Behav. Social Comput. (BESC)*, Oct. 2021, pp. 1–7, doi: [10.1109/BESC53957.2021.9635559](https://doi.org/10.1109/BESC53957.2021.9635559).
- [11] I. D. Mienye and Y. Sun, "Performance analysis of cost-sensitive learning methods with application to imbalanced medical data," *Inform. Med. Unlocked*, vol. 25, Jan. 2021, Art. no. 100690, doi: [10.1016/j.imu.2021.100690](https://doi.org/10.1016/j.imu.2021.100690).
- [12] S. A. Ebiaredoh-Mienye, T. G. Swart, E. Ezenogho, and I. D. Mienye, "A machine learning method with filter-based feature selection for improved prediction of chronic kidney disease," *Bioengineering*, vol. 9, no. 8, p. 350, Jul. 2022, doi: [10.3390/bioengineering9080350](https://doi.org/10.3390/bioengineering9080350).
- [13] C. Ho, Z. Zhao, X. F. Chen, J. Sauer, S. A. Saraf, R. Jialdasani, K. Taghipour, A. Sathe, L.-Y. Khor, K.-H. Lim, and W.-Q. Leow, "A promising deep learning-assistive algorithm for histopathological screening of colorectal cancer," *Sci. Rep.*, vol. 12, no. 1, pp. 1–9, Feb. 2022, doi: [10.1038/s41598-022-06264-x](https://doi.org/10.1038/s41598-022-06264-x).
- [14] P. Goel, R. Jain, A. Nayyar, S. Singhal, and M. Srivastava, "Sarcasm detection using deep learning and ensemble learning," *Multimedia Tools Appl.*, vol. 81, no. 30, pp. 43229–43252, Dec. 2022, doi: [10.1007/s11042-022-12930-z](https://doi.org/10.1007/s11042-022-12930-z).
- [15] I. D. Mienye, P. Kenneth Aina, I. D. Emmanuel, and E. Ezenogho, "Sparse noise minimization in image classification using genetic algorithm and DenseNet," in *Proc. Conf. Inf. Commun. Technol. Soc. (ICTAS)*, Mar. 2021, pp. 103–108, doi: [10.1109/ICTAS50802.2021.9395014](https://doi.org/10.1109/ICTAS50802.2021.9395014).
- [16] R. T. Aruleba, T. A. Adekiya, N. Ayawei, G. Obaido, K. Aruleba, I. D. Mienye, I. Aruleba, and B. Ogbuokiri, "COVID-19 diagnosis: A review of rapid antigen, RT-PCR and artificial intelligence methods," *Bioengineering*, vol. 9, no. 4, p. 153, Apr. 2022, doi: [10.3390/bioengineering9040153](https://doi.org/10.3390/bioengineering9040153).
- [17] G. Nguyen, S. Dlugolinsky, M. Bobák, V. Tran, L. L. García, I. Heredia, P. Malík, and L. Hluchý, "Machine learning and deep learning frameworks and libraries for large-scale data mining: A survey," *Artif. Intell. Rev.*, vol. 52, no. 1, pp. 77–124, Jun. 2019. [Online]. Available: <http://link.springer.com/10.1007/s10462-018-09679-z>
- [18] S. O. Alhumoud and A. A. Al Wazrah, "Arabic sentiment analysis using recurrent neural networks: A review," *Artif. Intell. Rev.*, vol. 55, no. 1, pp. 707–748, Jan. 2022, doi: [10.1007/s10462-021-09989-9](https://doi.org/10.1007/s10462-021-09989-9).
- [19] Z. Zhong, Y. Gao, Y. Zheng, B. Zheng, and I. Sato, "Real-world video deblurring: A benchmark dataset and an efficient recurrent neural network," *Int. J. Comput. Vis.*, vol. 131, no. 1, pp. 284–301, Jan. 2023, doi: [10.1007/s11263-022-01705-6](https://doi.org/10.1007/s11263-022-01705-6).
- [20] J. Van Gompel, D. Spina, and C. Develder, "Satellite based fault diagnosis of photovoltaic systems using recurrent neural networks," *Appl. Energy*, vol. 305, Jan. 2022, Art. no. 117874, doi: [10.1016/j.apenergy.2021.117874](https://doi.org/10.1016/j.apenergy.2021.117874).
- [21] 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, doi: [10.1016/j.asoc.2020.106852](https://doi.org/10.1016/j.asoc.2020.106852).
- [22] A. Tsantekidis, N. Passalis, and A. Tefas, "Chapter 5—Recurrent neural networks," in *Deep Learning for Robot Perception and Cognition*, A. Iosifidis and A. Tefas, Eds. New York, NY, USA: Academic, 2022, pp. 101–115, doi: [10.1016/B978-0-32-385787-1.00010-5](https://doi.org/10.1016/B978-0-32-385787-1.00010-5).
- [23] Y. Xie, G. Liu, C. Yan, C. Jiang, M. Zhou, and M. Li, "Learning transactional behavioral representations for credit card fraud detection," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Oct. 5, 2022, doi: [10.1109/TNNLS.2022.3208967](https://doi.org/10.1109/TNNLS.2022.3208967).
- [24] Y.-C. Wei, Y.-X. Lai, and M.-E. Wu, "An evaluation of deep learning models for chargeback fraud detection in online games," *Cluster Comput.*, vol. 26, pp. 927–943, Jul. 2022, doi: [10.1007/s10586-022-03674-4](https://doi.org/10.1007/s10586-022-03674-4).
- [25] S. Mishra, K. Shaw, D. Mishra, S. Patil, K. Kotecha, S. Kumar, and S. Bajaj, "Improving the accuracy of ensemble machine learning classification models using a novel bit-fusion algorithm for healthcare AI systems," *Frontiers Public Health*, vol. 10, May 2022, Art. no. 858282. [Online]. Available: <https://www.frontiersin.org/articles/10.3389/fpubh.2022.858282>
- [26] I. D. Mienye, G. Obaido, K. Aruleba, and O. A. Dada, "Enhanced prediction of chronic kidney disease using feature selection and boosted classifiers," in *Intelligent Systems Design and Applications*. Cham, Switzerland: Springer, 2022, pp. 527–537, doi: [10.1007/978-3-030-96308-8_49](https://doi.org/10.1007/978-3-030-96308-8_49).
- [27] N. L. Fitriyani, M. Syafrudin, G. Alfian, C.-K. Yang, J. Rhee, and S. M. Ulyah, "Chronic disease prediction model using integration of DBSCAN, SMOTE-ENN, and random forest," in *Proc. ASU Int. Conf. Emerg. Technol. Sustainability Intell. Syst. (ICETSIS)*, Jun. 2022, pp. 289–294, doi: [10.1109/ICETSIS55481.2022.9888806](https://doi.org/10.1109/ICETSIS55481.2022.9888806).
- [28] J. Yang and J. Guan, "A heart disease prediction model based on feature optimization and smote-Xgboost algorithm," *Information*, vol. 13, no. 10, p. 475, Oct. 2022, doi: [10.3390/info13100475](https://doi.org/10.3390/info13100475).
- [29] D. S. Sisodia, N. K. Reddy, and S. Bhandari, "Performance evaluation of class balancing techniques for credit card fraud detection," in *Proc. IEEE Int. Conf. Power, Control, Signals Instrum. Eng. (ICPCSI)*, Sep. 2017, pp. 2747–2752, doi: [10.1109/ICPCSI.2017.8392219](https://doi.org/10.1109/ICPCSI.2017.8392219).
- [30] H. Guan, Y. Zhang, M. Xian, H. D. Cheng, and X. Tang, "SMOTE-WENN: Solving class imbalance and small sample problems by oversampling and distance scaling," *Int. J. Speech Technol.*, vol. 51, no. 3, pp. 1394–1409, Mar. 2021, doi: [10.1007/s10489-020-01852-8](https://doi.org/10.1007/s10489-020-01852-8).
- [31] L. Ni, J. Li, H. Xu, X. Wang, and J. Zhang, "Fraud feature boosting mechanism and spiral oversampling balancing technique for credit card fraud detection," *IEEE Trans. Computat. Social Syst.*, early access, Feb. 13, 2023, doi: [10.1109/TCSS.2023.3242149](https://doi.org/10.1109/TCSS.2023.3242149).
- [32] H. Fanai and H. Abbasimehr, "A novel combined approach based on deep autoencoder and deep classifiers for credit card fraud detection," *Exp. Syst. Appl.*, vol. 217, May 2023, Art. no. 119562, doi: [10.1016/j.eswa.2023.119562](https://doi.org/10.1016/j.eswa.2023.119562).
- [33] F. Itoo, Meenakshi, and S. Singh, "Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection," *Int. J. Inf. Technol.*, vol. 13, no. 4, pp. 1503–1511, Aug. 2021, doi: [10.1007/s41870-020-00430-y](https://doi.org/10.1007/s41870-020-00430-y).
- [34] S. K. Saddam Hussain, E. Sai Charan Reddy, K. G. Akshay, and T. Akanksha, "Fraud detection in credit card transactions using SVM and random forest algorithms," in *Proc. 5th Int. Conf. I-SMAC (IoT Social, Mobile, Analytics Cloud) (I-SMAC)*, Nov. 2021, pp. 1013–1017, doi: [10.1109/I-SMAC52330.2021.9640631](https://doi.org/10.1109/I-SMAC52330.2021.9640631).
- [35] I. D. Mienye, Y. Sun, Z. Wang, "Prediction performance of improved decision tree-based algorithms: A review," *Proc. Manuf.*, vol. 35, pp. 698–703, Jan. 2019, doi: [10.1016/j.promfg.2019.06.011](https://doi.org/10.1016/j.promfg.2019.06.011).
- [36] K. Randhawa, C. K. Loo, M. Seera, C. P. Lim, and A. K. Nandi, "Credit card fraud detection using AdaBoost and majority voting," *IEEE Access*, vol. 6, pp. 14277–14284, 2018, doi: [10.1109/ACCESS.2018.2806420](https://doi.org/10.1109/ACCESS.2018.2806420).
- [37] T.-H. Lin and J.-R. Jiang, "Credit card fraud detection with autoencoder and probabilistic random forest," *Mathematics*, vol. 9, no. 21, p. 2683, Oct. 2021, doi: [10.3390/math9212683](https://doi.org/10.3390/math9212683).
- [38] A. Rb and S. K. Kr, "Credit card fraud detection using artificial neural network," *Global Transitions Proc.*, vol. 2, no. 1, pp. 35–41, Jun. 2021, doi: [10.1016/j.gltp.2021.01.006](https://doi.org/10.1016/j.gltp.2021.01.006).
- [39] S. C. Dubey, K. S. Mundhe, and A. A. Kadam, "Credit card fraud detection using artificial neural network and BackPropagation," in *Proc. 4th Int. Conf. Intell. Comput. Control Syst. (ICICCS)*, May 2020, pp. 268–273, doi: [10.1109/ICICCS48265.2020.9120957](https://doi.org/10.1109/ICICCS48265.2020.9120957).
- [40] O. N. Akande, S. Misra, H. B. Akande, J. Oluranti, and R. Damasevicius, "A supervised approach to credit card fraud detection using an artificial neural network," in *Applied Informatics*. Cham, Switzerland: Springer, 2021, pp. 13–25, doi: [10.1007/978-3-030-89654-6_2](https://doi.org/10.1007/978-3-030-89654-6_2).
- [41] A. A. Taha and S. J. Malebary, "An intelligent approach to credit card fraud detection using an optimized light gradient boosting machine," *IEEE Access*, vol. 8, pp. 25579–25587, 2020, doi: [10.1109/ACCESS.2020.2971354](https://doi.org/10.1109/ACCESS.2020.2971354).
- [42] E. Ileberi, Y. Sun, and Z. Wang, "A machine learning based credit card fraud detection using the GA algorithm for feature selection," *J. Big Data*, vol. 9, no. 1, p. 24, Feb. 2022, doi: [10.1186/s40537-022-00573-8](https://doi.org/10.1186/s40537-022-00573-8).
- [43] Z. Salekshahrezaee, J. L. Leevy, and T. M. Khoshgoftaar, "The effect of feature extraction and data sampling on credit card fraud detection," *J. Big Data*, vol. 10, no. 1, p. 6, Jan. 2023, doi: [10.1186/s40537-023-00684-w](https://doi.org/10.1186/s40537-023-00684-w).
- [44] A. Singh, R. K. Ranjan, and A. Tiwari, "Credit card fraud detection under extreme imbalanced data: A comparative study of data-level algorithms," *J. Experim. Theor. Artif. Intell.*, vol. 34, no. 4, pp. 571–598, Jul. 2022, doi: [10.1080/0952813X.2021.1907795](https://doi.org/10.1080/0952813X.2021.1907795).
- [45] Z. Li, M. Huang, G. Liu, and C. Jiang, "A hybrid method with dynamic weighted entropy for handling the problem of class imbalance with overlap in credit card fraud detection," *Exp. Syst. Appl.*, vol. 175, Aug. 2021, Art. no. 114750, doi: [10.1016/j.eswa.2021.114750](https://doi.org/10.1016/j.eswa.2021.114750).

- [46] S. H. Khan, M. Hayat, M. Bennamoun, F. A. Sohel, and R. Togneri, "Cost-sensitive learning of deep feature representations from imbalanced data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 8, pp. 3573–3587, Aug. 2018, doi: [10.1109/TNNLS.2017.2732482](https://doi.org/10.1109/TNNLS.2017.2732482).
- [47] J. Jurgovsky, M. Granitzer, K. Ziegler, S. Calabretto, P.-E. Portier, L. He-Guelton, and O. Caelen, "Sequence classification for credit-card fraud detection," *Exp. Syst. Appl.*, vol. 100, pp. 234–245, Jun. 2018, doi: [10.1016/j.eswa.2018.01.037](https://doi.org/10.1016/j.eswa.2018.01.037).
- [48] I. D. Mienye, Y. Sun, and Z. Wang, "Improved predictive sparse decomposition method with DenseNet for prediction of lung cancer," *Int. J. Comput.*, vol. 14, no. 4, pp. 533–541, Dec. 2020.
- [49] *Credit Card Fraud Detection*. Accessed: Oct. 26, 2021. [Online]. Available: <https://kaggle.com/mlg-ulb/creditcardfraud>
- [50] K. Aruleba, G. Obaido, B. Ogbuokiri, and A. O. Fadaka, "Applications of computational methods in biomedical breast cancer imaging diagnostics: A review," *J. Imag.*, vol. 6, no. 10, p. 105, 2020.
- [51] Z. Zhang, D. Zhou, J. Zhang, Y. Xu, G. Lin, B. Jin, Y. Liang, Y. Geng, and S. Zhang, "Multilayer perceptron-based prediction of stroke mimics in prehospital triage," *Sci. Rep.*, vol. 12, no. 1, Oct. 2022, Art. no. 17994, doi: [10.1038/s41598-022-22919-1](https://doi.org/10.1038/s41598-022-22919-1).
- [52] T. Bikku, "Multi-layered deep learning perceptron approach for health risk prediction," *J. Big Data*, vol. 7, no. 1, p. 50, Jul. 2020, doi: [10.1186/s40537-020-00316-7](https://doi.org/10.1186/s40537-020-00316-7).
- [53] Y. Moodi, M. Ghasemi, and S. R. Mousavi, "Estimating the compressive strength of rectangular fiber reinforced polymer-confined columns using multilayer perceptron, radial basis function, and support vector regression methods," *J. Reinforced Plastics Composites*, vol. 41, nos. 3–4, pp. 130–146, Feb. 2022, doi: [10.1177/07316844211050168](https://doi.org/10.1177/07316844211050168).
- [54] S. Agahian and T. Akan, "Battle royale optimizer for training multi-layer perceptron," *Evolving Syst.*, vol. 13, no. 4, pp. 563–575, Aug. 2022, doi: [10.1007/s12530-021-09401-5](https://doi.org/10.1007/s12530-021-09401-5).
- [55] I. D. Mienye and Y. Sun, "Improved heart disease prediction using particle swarm optimization based stacked sparse autoencoder," *Electronics*, vol. 10, no. 19, p. 2347, Jan. 2021, doi: [10.3390/electronics10192347](https://doi.org/10.3390/electronics10192347).
- [56] S. M. Kasongo and Y. Sun, "Performance analysis of intrusion detection systems using a feature selection method on the UNSW-NB15 dataset," *J. Big Data*, vol. 7, no. 1, p. 105, Nov. 2020, doi: [10.1186/s40537-020-00379-6](https://doi.org/10.1186/s40537-020-00379-6).
- [57] S. M. Kasongo, "A deep learning technique for intrusion detection system using a recurrent neural networks based framework," *Comput. Commun.*, vol. 199, pp. 113–125, Feb. 2023, doi: [10.1016/j.comcom.2022.12.010](https://doi.org/10.1016/j.comcom.2022.12.010).
- [58] M. Ragab, K. Eljaaly, N. A. Alhakamy, H. A. Alhadrami, A. A. Bahaddad, S. M. Abo-Dahab, and E. M. Khalil, "Deep ensemble model for COVID-19 diagnosis and classification using chest CT images," *Biology*, vol. 11, no. 1, p. 43, Dec. 2021, doi: [10.3390/biology11010043](https://doi.org/10.3390/biology11010043).
- [59] J. Yang, J. Qu, Q. Mi, and Q. Li, "A CNN-LSTM model for tailings dam risk prediction," *IEEE Access*, vol. 8, pp. 206491–206502, 2020, doi: [10.1109/ACCESS.2020.3037935](https://doi.org/10.1109/ACCESS.2020.3037935).
- [60] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder–decoder approaches," in *Proc. SSST 8th Workshop Syntax, Semantics Struct. Stat. Transl.*, Doha, Qatar, 2014, pp. 103–111, doi: [10.3115/v1/W14-4012](https://doi.org/10.3115/v1/W14-4012).
- [61] H. Xie, M. Randall, and K.-W. Chau, "Green roof hydrological modelling with GRU and LSTM networks," *Water Resour. Manage.*, vol. 36, no. 3, pp. 1107–1122, Feb. 2022, doi: [10.1007/s11269-022-03076-6](https://doi.org/10.1007/s11269-022-03076-6).
- [62] S. Gao, Y. Huang, S. Zhang, J. Han, G. Wang, M. Zhang, and Q. Lin, "Short-term runoff prediction with GRU and LSTM networks without requiring time step optimization during sample generation," *J. Hydrol.*, vol. 589, Oct. 2020, Art. no. 125188, doi: [10.1016/j.jhydrol.2020.125188](https://doi.org/10.1016/j.jhydrol.2020.125188).
- [63] J. Zhou, Z. Jiang, F.-L. Chung, and S. Wang, "Formulating ensemble learning of SVMs into a single SVM formulation by negative agreement learning," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 51, no. 10, pp. 6015–6028, Oct. 2021, doi: [10.1109/TSMC.2019.2958647](https://doi.org/10.1109/TSMC.2019.2958647).
- [64] I. D. Mienye and Y. Sun, "Effective feature selection for improved prediction of heart disease," in *Pan-African Artificial Intelligence and Smart Systems*. Cham, Switzerland: Springer, 2022, pp. 94–107, doi: [10.1007/978-3-030-93314-2_6](https://doi.org/10.1007/978-3-030-93314-2_6).
- [65] Y. Sun, Z. Li, X. Li, and J. Zhang, "Classifier selection and ensemble model for multi-class imbalance learning in education grants prediction," *Appl. Artif. Intell.*, vol. 35, no. 4, pp. 290–303, Mar. 2021, doi: [10.1080/08839514.2021.1877481](https://doi.org/10.1080/08839514.2021.1877481).
- [66] S. Doroudi, "The bias-variance tradeoff: How data science can inform educational debates," *AERA Open*, vol. 6, no. 4, pp. 1–18, Oct. 2020, doi: [10.1177/2332858420977208](https://doi.org/10.1177/2332858420977208).
- [67] S. Alelyani, "Stable bagging feature selection on medical data," *J. Big Data*, vol. 8, no. 1, p. 11, Jan. 2021, doi: [10.1186/s40537-020-00385-8](https://doi.org/10.1186/s40537-020-00385-8).
- [68] G. Brown, "Ensemble learning," in *Encyclopedia of Machine Learning*, C. Sammut and G. I. Webb, Eds. Boston, MA, USA: Springer, 2010, pp. 312–320, doi: [10.1007/978-0-387-30164-8_252](https://doi.org/10.1007/978-0-387-30164-8_252).
- [69] I. D. Mienye, Y. Sun, and Z. Wang, "An improved ensemble learning approach for the prediction of heart disease risk," *Inform. Med. Unlocked*, vol. 20, Jan. 2020, Art. no. 100402, doi: [10.1016/j.imu.2020.100402](https://doi.org/10.1016/j.imu.2020.100402).
- [70] R. Polikar, "Ensemble learning," in *Ensemble Machine Learning: Methods and Applications*, C. Zhang and Y. Ma, Eds. Boston, MA, USA: Springer, 2012, pp. 1–34, doi: [10.1007/978-1-4419-9326-7_1](https://doi.org/10.1007/978-1-4419-9326-7_1).
- [71] M. Liang, T. Chang, B. An, X. Duan, L. Du, X. Wang, and J. Miao, "A stacking ensemble learning framework for genomic prediction," *Frontiers Genet.*, vol. 12, May 2022, Art. no. 600040. [Online]. Available: <https://www.frontiersin.org/article/10.3389/fgene.2021.600040>
- [72] I. D. Mienye and Y. Sun, "A survey of ensemble learning: Concepts, algorithms, applications, and prospects," *IEEE Access*, vol. 10, pp. 99129–99149, 2022, doi: [10.1109/ACCESS.2022.3207287](https://doi.org/10.1109/ACCESS.2022.3207287).
- [73] Y. Xia, K. Chen, and Y. Yang, "Multi-label classification with weighted classifier selection and stacked ensemble," *Inf. Sci.*, vol. 557, pp. 421–442, May 2021, doi: [10.1016/j.ins.2020.06.017](https://doi.org/10.1016/j.ins.2020.06.017).
- [74] P. Saikia and R. D. Baruah, "Investigating stacked ensemble model for oil reservoir characterisation," in *Proc. IEEE Int. Conf. Syst., Man Cybern. (SMC)*, Oct. 2019, pp. 13–20, doi: [10.1109/SMC.2019.8914488](https://doi.org/10.1109/SMC.2019.8914488).
- [75] F. Anifowose, J. Labadin, and A. Abdurraheem, "Improving the prediction of petroleum reservoir characterization with a stacked generalization ensemble model of support vector machines," *Appl. Soft Comput. J.*, vol. 26, pp. 483–496, Jan. 2015, doi: [10.1016/j.asoc.2014.10.017](https://doi.org/10.1016/j.asoc.2014.10.017).
- [76] S. F. Abdoh, M. A. Rizka, and F. A. Maghraby, "Cervical cancer diagnosis using random forest classifier with SMOTE and feature reduction techniques," *IEEE Access*, vol. 6, pp. 59475–59485, 2018, doi: [10.1109/ACCESS.2018.2874063](https://doi.org/10.1109/ACCESS.2018.2874063).
- [77] A. Ishaq, S. Sadiq, M. Umer, S. Ullah, S. Mirjalili, V. Rupapara, and M. Nappi, "Improving the prediction of heart failure Patients' survival using SMOTE and effective data mining techniques," *IEEE Access*, vol. 9, pp. 39707–39716, 2021, doi: [10.1109/ACCESS.2021.3064084](https://doi.org/10.1109/ACCESS.2021.3064084).
- [78] K. Aruleba, N. Jere, and O. Matarirano, "An evaluation of technology adoption during remote teaching and learning at tertiary institution by gender," *IEEE Trans. Computat. Social Syst.*, early access, Apr. 13, 2022, doi: [10.1109/TCSS.2022.3163912](https://doi.org/10.1109/TCSS.2022.3163912).
- [79] Asniar, N. U. Maulidevi, and K. Surendro, "SMOTE-LOF for noise identification in imbalanced data classification," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, no. 6, pp. 3413–3423, Jun. 2022, doi: [10.1016/j.jksuci.2021.01.014](https://doi.org/10.1016/j.jksuci.2021.01.014).
- [80] I. D. Mienye and Y. Sun, "Heart disease prediction using enhanced machine learning techniques," in *Intelligent Systems and Machine Learning for Industry*, Boca Raton, FL, USA: CRC Press, 2022, pp. 93–114.
- [81] T. Le, M. T. Vo, B. Vo, M. Y. Lee, and S. W. Baik, "A hybrid approach using oversampling technique and cost-sensitive learning for bankruptcy prediction," *Complexity*, vol. 2019, pp. 1–12, Aug. 2019, doi: [10.1155/2019/8460934](https://doi.org/10.1155/2019/8460934).
- [82] Y. Li, J. Gao, Q. Li, and W. Fan, "Ensemble Learning," in *Data Classification*. Boca Raton, FL, USA: Chapman & Hall/CRC, 2014.
- [83] R. E. Schapire, "A brief introduction to boosting," in *Proc. IJCAI*, vol. 99, 1999, pp. 1401–1406.
- [84] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: [10.1023/A:1010933404324](https://doi.org/10.1023/A:1010933404324).
- [85] H. Taud and J. F. Mas, "Multilayer perceptron (MLP)," in *Geomatic Approaches for Modeling Land Change Scenarios*, M. T. C. Olmedo, M. Paegelow, J.-F. Mas, and F. Escobar, Eds. Cham, Switzerland: Springer, 2018, pp. 451–455, doi: [10.1007/978-3-319-60801-3_27](https://doi.org/10.1007/978-3-319-60801-3_27).
- [86] Y. Yu, X. Si, C. Hu, and Z. Jianxun, "A review of recurrent neural networks: LSTM cells and network architectures," *Neural Comput.*, vol. 31, no. 7, pp. 1235–1270, 2019, doi: [10.1162/neco_a_01199](https://doi.org/10.1162/neco_a_01199).
- [87] F. Pedregosa, G. Varoquaux, A. Gramfort, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. 10, pp. 2825–2830, 2012.

- [88] R. Trevethan, "Sensitivity, specificity, and predictive values: Foundations, pliabilitys, and pitfalls in research and practice," *Frontiers Public Health*, vol. 5, p. 307, Jun. 2022. [Online]. Available: <https://www.frontiersin.org/article/10.3389/fpubh.2017.00307>
- [89] 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, doi: [10.3390/app122111127](https://doi.org/10.3390/app122111127).
- [90] W. Li, Z. Fang, and Y. Wang, "Stacking ensemble of deep learning methods for landslide susceptibility mapping in the three Gorges reservoir area, China," *Stochastic Environ. Res. Risk Assessment*, vol. 36, no. 8, pp. 2207–2228, Aug. 2022, doi: [10.1007/s00477-021-02032-x](https://doi.org/10.1007/s00477-021-02032-x).
- [91] H. Zhu, G. Liu, M. Zhou, Y. Xie, A. Abusorrah, and Q. Kang, "Optimizing weighted extreme learning machines for imbalanced classification and application to credit card fraud detection," *Neurocomputing*, vol. 407, pp. 50–62, Sep. 2020, doi: [10.1016/j.neucom.2020.04.078](https://doi.org/10.1016/j.neucom.2020.04.078).
- [92] K. I. Alkhatib, A. I. Al-Aiad, M. H. Almahmoud, and O. N. Elayan, "Credit card fraud detection based on deep neural network approach," in *Proc. 12th Int. Conf. Inf. Commun. Syst. (ICICS)*, May 2021, pp. 153–156, doi: [10.1109/ICICS52457.2021.9464555](https://doi.org/10.1109/ICICS52457.2021.9464555).
- [93] S. Makki, Z. Assaghir, Y. Taher, R. Haque, M. S. Hacid, and H. Zeineddine, "An experimental study with imbalanced classification approaches for credit card fraud detection," *IEEE Access*, vol. 7, pp. 93010–93022, 2019, doi: [10.1109/ACCESS.2019.2927266](https://doi.org/10.1109/ACCESS.2019.2927266).
- [94] W. Yotsawat, P. Wattuya, and A. Srivihok, "A novel method for credit scoring based on cost-sensitive neural network ensemble," *IEEE Access*, vol. 9, pp. 78521–78537, 2021, doi: [10.1109/ACCESS.2021.3083490](https://doi.org/10.1109/ACCESS.2021.3083490).
- [95] S. N. Kalid, K.-H. Ng, G.-K. Tong, and K.-C. Khor, "A multiple classifiers system for anomaly detection in credit card data with unbalanced and overlapped classes," *IEEE Access*, vol. 8, pp. 28210–28221, 2020, doi: [10.1109/ACCESS.2020.2972009](https://doi.org/10.1109/ACCESS.2020.2972009).
- [96] X. Feng, Z. Xiao, B. Zhong, Y. Dong, and J. Qiu, "Dynamic weighted ensemble classification for credit scoring using Markov chain," *Int. J. Speech Technol.*, vol. 49, no. 2, pp. 555–568, Feb. 2019, doi: [10.1007/s10489-018-1253-8](https://doi.org/10.1007/s10489-018-1253-8).
- [97] P. Mrozek, J. Panneerselvam, and O. Bagdasar, "Efficient resampling for fraud detection during anonymised credit card transactions with unbalanced datasets," in *Proc. IEEE/ACM 13th Int. Conf. Utility Cloud Comput. (UCC)*, Dec. 2020, pp. 426–433, doi: [10.1109/UCC48980.2020.00067](https://doi.org/10.1109/UCC48980.2020.00067).
- [98] Y. Xie, A. Li, L. Gao, and Z. Liu, "A heterogeneous ensemble learning model based on data distribution for credit card fraud detection," *Wireless Commun. Mobile Comput.*, vol. 2021, pp. 1–13, Jul. 2021, doi: [10.1155/2021/2531210](https://doi.org/10.1155/2021/2531210).
- [99] H. Ahmad, B. Kasasbeh, B. Aldabaybah, and E. Rawashdeh, "Class balancing framework for credit card fraud detection based on clustering and similarity-based selection (SBS)," *Int. J. Inf. Technol.*, vol. 15, no. 1, pp. 325–333, Jan. 2023, doi: [10.1007/s41870-022-00987-w](https://doi.org/10.1007/s41870-022-00987-w).
- [100] *UCI Machine Learning Repository: Default of Credit Card Clients Data Set*. Accessed: Dec. 5, 2022. [Online]. Available: <https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients>



IBOMOYE DOMOR MIENYE (Member, IEEE) received the B.Eng. degree in electrical and electronic engineering and the M.Sc. degree (cum laude) in computer systems engineering from the University of East London, in 2012 and 2014, respectively, and the Ph.D. degree in electrical and electronic engineering from the University of Johannesburg, South Africa. He is currently a Post-doctoral Research Fellow with the Department of Electrical and Electronic Engineering Science, University of Johannesburg. His research interests include machine learning and deep learning for finance and 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 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.

• • •