IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 26 July 2023, accepted 10 August 2023, date of publication 18 August 2023, date of current version 25 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3306621

RESEARCH ARTICLE

Ensemble Synthesized Minority Oversampling-Based Generative Adversarial Networks and Random Forest Algorithm for Credit Card Fraud Detection

FUAD A. GHALEB[1], FAISAL SAEED[2], (Member, IEEE), MOHAMMED AL-SAREM[3], SULTAN NOMAN QASEM[4], (Senior Member, IEEE), AND TAWFIK AL-HADHRAMI[5]

[1]Department of Computer Science, Faculty of Computing, Universiti Teknologi Malaysia, Johor Bahru 81310, Malaysia
[2]DAAI Research Group, Department of Computing and Data Science, School of Computing and Digital Technology, Birmingham City University, B4 7XG Birmingham, U.K.
[3]College of Computer Science and Engineering, Taibah University, Medina 41477, Saudi Arabia
[4]Computer Science Department, College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh 11432, Saudi Arabia
[5]Department of Computer Science, School of Science and Technology, Nottingham Trent University, NG11 8NS Nottingham, U.K.

Corresponding author: Fuad A. Ghaleb (abdulgaleel@utm.my)

This work was supported and funded by the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU) through Research Partnership Program no RP-21-07-09.

ABSTRACT The recent increase in credit card fraud is rapidly has caused huge monetary losses for individuals and financial institutions. Most credit card frauds are conducted online by illegally obtaining payment credentials through data breaches, phishing, or scamming. Many solutions have been suggested to address the credit card fraud problem for online transactions. However, the high-class imbalance is the major challenge that faces the existing solutions to construct an effective detection model. Most of the existing techniques used for class imbalance overestimate the distribution of the minority class, resulting in highly overlapped or noisy and unrepresentative features, which cause either overfitting or imprecise learning. In this study, a credit card fraud detection model (CCFDM) is proposed based on ensemble learning and a generative adversarial network (GAN) assisted by Ensemble Synthesized Minority Oversampling techniques (ESMOTE-GAN). Multiple subsets were extracted using under-sampling and SMOTE was applied to generate less skewed sets to prevent the GAN from modeling the noise. These subsets were used to train diverse sets of GAN models to generate the synthesized subsets. A set of Random Forest classifiers was then trained based on the proposed ESMOTE-GAN technique. The probabilistic outputs of the trained classifiers were combined using a weighted voting scheme for decision-making. The results show that the proposed model achieved 1.9%, and 3.2% improvements in overall performance and the detection rate, respectively, with a 0% false alarm rate. Due to the massive number of transactions, even a tiny false positive rate can overwhelm the analysis team. Thus, the proposed model has improved the detection performance and reduced the cost needed for *manual* analysis.

INDEX TERMS Class imbalance, credit card fraud detection, GAN, Random Forest, SMOTE.

I. INTRODUCTION

Recently, credit card fraud has increased exponentially due to the reliance on online services, leading to huge losses of monetary funds from consumers and financial institutions. According to [1], credit card fraud cost $35 billion in 2020. The losses are expected to exceed $400 billion in the next decade [2]. Most credit card frauds are performed with card-not-present (CNP) scenarios, such as payments on the Internet, by phone, or by mail. CNP (online) fraud

The associate editor coordinating the review of this manuscript and approving it for publication was Jiachen Yang.

happens when fraudsters use credentials in an unauthorized manner. The credentials are usually illegally obtained from cardholders through data breaches, phishing, or scamming.

Credit card fraud detection is an essential component of monetary systems to detect and block fraudulent transactions. Transactions are investigated using a detection model for any suspicious activities. The detection model is usually built based on rules *designed* by experts or through rules driven by previous knowledge of fraudulent activities. A *manual* investigation is performed if customers have complained. However, due to the huge number of daily transactions, investigators may be overwhelmed by the number of false alarms. That is even with 99% detection accuracy, 1% can require a massive amount of work. The main challenge in constructing an effective credit card fraud detection model is the number of benign transactions compared to fraudulent cases. Learning the fraudulent patterns among millions of benign patterns looks like finding a needle in a haystack. With such a large amount of data, it is impossible even to craft effective rules for humans to follow.

Machine learning techniques, including deep learning, have been widely employed in the construction of credit card fraud detection models. Existing research in this area predominantly utilizes supervised machine learning methodologies to develop fraud classifiers. These classifiers are built based on the experiences and knowledge gained from previous transactions, which include both legitimate and fraudulent samples. However, many challenges are faced in constructing effective detection models, including class imbalance, concept drift, features engineering, real-time requirements, class overlap, and lack of public datasets. The class imbalance problem, which is the focus of this study, has received much attention from researchers because it leads to biased classification toward the majority class [3]. The minority instances are ignored by the classifiers, leading to a low detection rate and a high number of false alarms.

The problems of learning from an imbalanced dataset have been extensively studied in the literature. The existing solutions can be categorized into two main approaches: imbalanced learning and data resampling [4]. In imbalanced learning, also called cost-sensitive classification [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], the classifier is forced to learn minority class patterns by assigning higher scores for minority class samples than those assigned to majority class instances. Ensemble learning has also been suggested for an imbalanced dataset by many researchers [7], [15], [16]. However, although ensemble learning slightly improves the classification performance, by reducing the overfitting due to of diversity, the decision regarding the correct class flows to the majority win strategy, and thus the majority of the classifiers are biased. Meanwhile, in the data resampling approach [7], [17], [18], [19], the instances belonging to the classes are balanced by either oversampling the minority class or under-sampling the majority class. Although these approaches are effective for some applications, they have several drawbacks for other applications. Credit card transaction datasets are rare because they contain sensitive and private information for customers and monetary institutions. In addition, credit card transaction datasets suffer from a high class imbalance: the percentage of fraudulent transactions is lower than 0.01% in most available datasets [20], [21]. This means the solutions proposed for other domains do not necessarily fit the credit card fraud dataset problem. Accordingly, the cost-sensitive approach suffers from an insufficient amount of data to learn from. Thus, resampling is a commonly reported method for imbalance classification, due to its role in magnifying the representation of the minority class. However, the basic resampling approach leads to either overestimating the minority class or underestimating the majority, class leading to imprecise classification. More advanced resampling strategies have also been studied to improve the representation of the minority class [7], [18], [22], [23], [24]. In some studies, synthetic instances were interpolated based on the minority class and were also generated using the Synthetic Minority Oversampling Technique SMOTE [25] and Adaptive Synthetic Sampling ADASYN [26]. However, such techniques are highly dependent on the synthesis of data for learning. Thus, they are imprecise and produce an unstable performance based on the synthesis samples. Recently, a Generative Adversarial Network (GAN) has been utilized to improve the predictivity of minority instances [27], [28], [29]. However, GAN-predicted instances are noisy and biased, due to the insufficient number of samples of the minority class to learn from.

Although the problem of imbalanced data exists in many real-world data and has been extensively studied in the literature, most of the existing solutions solve moderated and low imbalanced-class problem. Credit card fraud data are highly imbalanced (highly skewed), and there are too few samples in the minority class compared to the majority class. A highly skewed dataset leads to training biased classifiers with poor generalizability and poor accuracy. A few solutions have been proposed to solve the imbalance problem of credit card fraud detection. However, the main drawback of these solutions has been the low detection rate and high false alarms due to either imprecise learning as a result of shifting the decision boundary towards the minority class in the cost-sensitive approaches or unrepresentative samples created by sampling techniques. In this study, the Ensemble Synthesized Minority Oversampling based Generative Adversarial Networks technique called (ESMOT-GAN) is proposed to generate synthesized yet representative instances class instances. Multiple subsets of synthesized instances were created using an ensemble of GAN models. Multiple subsets with less skewed data were generated using the under-sampling technique. The SMOTE technique was then used to generate the multiple training subsets that were less skewed and more diverse. To prevent the GANs from modeling the noise, SMOTE was used to partially oversample the minority class and generate moderately imbalanced subsets. GAN models were trained based on subsets with moderately imbalanced classes generated by the SMOTE technique to

accurately predict fraudulent transactions by removing the noises generated by SMOTE. The SMOTE technique was utilized to produce diverse subsets of dataset with fewer overlapping features. Eventually, the ensemble of SMOTE-GAN further eliminates the impact of overlapped features on the model's performance. Accordingly, the ensemble GAN produces diverse yet less noisy and less overlapped features. The outputs of GAN models are diverse subsets of training-balanced datasets. An ensemble of classifiers was constructed for each SMOTE-GAN-produced subset, using the Random Forest (RF) algorithm. RF was selected based on an investigation of best-performing classifiers in noisy and overlapped features. For generalizability and to reduce the overfitting problem, a weighted probabilistic average scheme is used for the final decision.

The main contributions of this study are as follows.

1) An ensemble-based data augmentation technique called ESMOTE-GAN is proposed to address the problem of a highly skewed dataset for credit card fraud detection. SMOTE and GAN techniques were used to generate diverse subsets of the training dataset with balanced yet less overlapped features. To prevent the GAN from modeling the noise generated by SMOTE, SMOTE first was used to partially oversample the minority class and generate moderately imbalanced subsets so that noise could be eliminated. Ensemble sets of GAN networks were trained based on the generated subsets and used to eliminate the noise and improve the representability of the synthesized fraudulent samples.
2) An ensemble-based Credit Card Fraud Detection Model (CCFDM) was designed and developed based on training a diverse set of classifiers using a Random Forest algorithm. In this model, the decision is made based on the weighted probabilistic voting scheme. The performance of the classifiers was used to represent the uncertainty of the model and improve the detection accuracy.
3) Extensive experiments were conducted to evaluate the proposed CCFDM model. The performance of both the augmentation ESMOTE-GAN technique and the detection model (CCFDM) were compared with state-of-the-art techniques and models.

The rest of this paper is organized as follows. The related works are reviewed in Section II. The limitations of the existing solutions and the motivation gap are also discussed. Section III presents the proposed augmentation technique with the corresponding fraud detection model. The experimental details including the used dataset, performance evaluation, and performance measures are presented in Section IV. The results are discussed in Section V and Section 6 presents the conclusions and suggests future work.

II. RELATED WORKS

Learning from imbalanced datasets has been extensively studied in the literature. The solutions provided can be categorized into two main approaches: data resampling and imbalanced learning [4]. The resampling approach works at the data level and can be further classified into three types, namely, under-sampling, oversampling, and combinations of oversampling and under-sampling. Under-sampling is conducted by either randomly removing samples from the majority class [8] or by replacing a group of samples with their cluster centroid. Although under-sampling can be effective in large datasets, removing samples from small datasets results in a loss of potential patterns and causes learning of unrepresentative models and ineffective classification.

Oversampling is achieved by replicating the minority class samples or by generating synthetic samples by interpolating samples from the minority class, such as in the Synthetic Minority Oversampling Technique SMOTE [25] and Adaptive Synthetic Sampling ADASYN [26]. Oversampling has been widely used for credit card fraud detection. Unfortunately, oversampling by duplicating the minority class samples either leads to overfitting (in the case of random resampling) or to amplifying the noise in the data (in the case of a synthetic minority) [30]. In addition, generating synthesized samples without considering the majority class leads to generating overlapping features between majority and minority samples [30]. Moreover, even if the oversampling leads to a balance in the dataset, the internal distribution of the minority class might become unrepresentative, due to the unpredictable behavior of the fraud. Despite these drawbacks, the research community has widely adopted SMOTE. Various extensions and modifications of this technique have been proposed to eliminate its weaknesses [22]. An experimental study that used imbalanced classification approaches for credit card fraud detection [10] compared machine learning algorithms. Random Oversampling was used for data balancing and the C5.0 algorithm, Support Vector Machine (SVM), Naïve Bayes (NB), Artificial Neural Network (ANN), Bayesian Belief Network (BBN), Logistic Regression (LR), K-Nearest Neighbor (KNN), Artificial Immune Systems (AIS), and Negative Selection Algorithm (NSA). In most tested classifiers the number of false alarms is higher than the number of fraudulent samples, which is ineffective indicating the ineffectiveness of these solutions.

Combinations of oversampling and sampling improve the representation, due to the inclusion of more patterns from the majority class and eliminate the noise resulting from the presence of synthesis data with overlapped regions between classes. The authors in [9] proposed a hybrid model that combines oversampling and under-sampling techniques to balance the dataset. SMOTE was used for oversampling while the Spread Subsample was used for under-sampling of the majority class. However, under-sampling the minority class leads to unrepresentative features leading to impractical solutions.

Recently, deep learning techniques such as Generative Adversarial Networks (GAN) and Autoencoders have been utilized to extract distinctive features from the minority class and accordingly predict samples from the same distribution.

The Generative Adversarial Network (GAN) has been used to learn the target distribution and accordingly generate artificial yet plausible samples from the same distribution. GAN has been reported as a useful technique for data argumentation, due to its ability to simulate the distribution of that data [9], [18], [22], [27], [28], [31], [32]. In [17], the authors proposed a solution for data balancing using deep conditional generative models. The minority class samples were oversampled using the GAN model. However, the imprecise prediction by GAN may lead to generating inaccurate samples. Autoencoder-based models have been utilized by many researchers [12], [14], [28], [33], [34] to solve the imbalanced data problem by removing the noise and approximating the distribution of either minority or majority classes. For example, the authors in [12] utilized the autoencoder to address the data skewness problem. In [33], the authors proposed a credit card fraud detection model based on autoencoding to reduce the dimensionality of the feature. The probabilistic Random Forest algorithm was then used to construct the detection classifier. However, the ability to denoise the features may lead to oversampling using autoencoding and hence an overfitting problem. In [28] the autoencoder was integrated with the GAN model to solve the imbalance problem. Although the two models described in [33] and [28] can complement each other, the integration is conducted to solve the sparsity of features in the original dataset. Such integration doesn't help improve the credit card fraud detection model due to the highly imbalanced data and availability of dense features.

The second approach to address the imbalance problem is imbalanced learning, in which resampling is integrated with ensemble learning. In [16], the authors proposed an ensemble learning model by combining different machine learning algorithms. Multiple datasets were created using the random under-sampling technique. These subsets were used to train classifiers using different machine learning algorithms such as SVM, LR, AdaBoost, and NB. For each subset, the best-performing classifier was selected to construct the ensemble model. Their results showed an improvement in detection rates compared to the studied models. However, with the highly unbalanced dataset, there is a potential of losing effective patterns from the dataset and because the heterogeneous classifiers were treated as equal during the decision process, the decision may not be accurate in realistic scenarios. In [35] the Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) was utilized to propose a Credit Card Fraud Detection Model. Oversampling was used for balancing the dataset. However, in addition to the disadvantages of oversampling, in the methodology used oversampling was applied for all datasets including the test dataset, which is not realistic. Thus, the samples that were in the training test also appear in the testing set which leads to inaccurate results. In the present study, the test dataset is separated before applying the argumentation techniques to demonstrate the performance in a realistic situation. The authors in [1] proposed a solution for the class-imbalance problem using bidirectional Long short-term memory (BiLSTM) and bidirectional Gated recurrent unit (BiGRU). An ensemble-based solution was proposed in [7], in which the data imbalance problem was solved using the SMOTE technique with edited k-nearest neighbors. The adaptive boosting (AdaBoost) technique was then used to train a set of classifiers using long short-term memory (LSTM). However, such a model has been evaluated based on the specificity which depends on the majority class which will be always near one. The *precision*, which is an important measure as it evaluates the probability of positive classification among all the samples predicted positive, can be used to generalize performance. A comparative analysis of the performance of several conventional and deep learning techniques was conducted in [5], using the European card dataset to benchmark fraud detection. Algorithms such as DT, KNN, LR, SVM, RF, XGBoost, and CNN were used in the experiments. The overall performance in terms of F-measure showed that the KNN performed better than the other studied models. [36] studied the highly imbalanced class problem and investigated the performance of Naïve Bayes, k-nearest neighbor, and logistic regression, both undersampling and oversampling were used for the skewed problem. The k-nearest neighbors' techniques were reported to achieve the best accuracy. However, all the studied techniques produced high false alarm rates, and their performance was evaluated using the accuracy measure which is not suitable for an imbalanced dataset. Thus the authors of [7] proposed a model that combines ensemble learning and hybrid data resampling methods. An ensemble of long short-term memory (LSTM) is constructed using the adaptive boosting (AdaBoost) technique and trained based on the data generated by the SMOTE with the edited nearest neighbor (SMOTE-ENN). In [37] an ensemble deep-learning model was proposed using LSTM and GRU neural networks as base learners and an MLP as the meta-learner, to address the challenges of credit card fraud detection in the presence of dynamic shopping patterns and class imbalance. The SMOTE-ENN method was used to balance the class distribution in the dataset, aiming to enhance the performance of the classifiers. The authors of [24] investigated the performance of many ensemble-based machine learning algorithms using AdaBoost. However, low-skewed and synthetic datasets were used for training and validating the model which is not realistic. In [38] a comparative study of different machine learning models was conducted. However, the study relied solely on the under-sampling approach to balance the data. Although under-sampling showed high classification performance, removing samples from small datasets results in a loss of potential patterns and causes learning of unrepresentative models and ineffective classification.

In [39] an optimized light gradient boosting machine, OLightGBM, was used for the detection of credit card fraud. By utilizing Bayesian-based hyperparameter optimization, the proposed OLightGBM achieved outstanding performance on real-world credit card transaction datasets, surpassing other approaches. However, the results revealed that the

proposed method trades-off precision with detection rate, leading to low performance. To handle imbalanced data and enhance the performance of the LightGBM method, a study was conducted [40] that involved weight-tuning through class weight-tuning hyperparameters, as well as the utilization of CatBoost, XGBoost and deep learning with Bayesian optimization. The experimental findings indicated that LightGBM and XGBoost exhibited superior performance compared to other approaches. In [41] the authors proposed a loss function called full center loss (FCL), based on both the angle and the distance among neighbors, to maximize the intraclass compactness and separability. However, the problem of the highly skewed dataset in that study made the learner highly dependent on the synthesis data. A multiple Classifiers System (MCS) was proposed by [42]. MCS stacks a sequential set of classifiers such that the output of a classifier is used as input for the subsequent classifier. However, the model suffered from high false alarms, higher than 17% in best-case scenarios. Moreover, such performance is cost-intensive, due to the need for human intervention; thus, it is not suitable for the huge volume of transactions. In [43] a model was proposed using a stacked sparse autoencoder, where SMOTE was used to solve the skewness problem. The results showed that the stacked sparse autoencoder outperformed the conventional machine learning techniques.

Although the problem of imbalanced data exists in many real-world data and has been extensively studied in the literature, credit card fraud data is particularly highly imbalanced. Most of the existing solutions solve the problems of moderately or slightly unbalanced data [5], [7], [11], [15], [20], [31], [33], [44], [45], [46]. These solutions can be grouped under two main concepts: data augmentation and imbalanced learning. Data augmentation techniques include random oversampling, under-sampling, SMOTE, and GAN techniques, while imbalanced learning models comprise different types of ensemble learning and cost-sensitive learning. Each of these solutions has its advantages and limitations. In terms of data augmentation, a combination of different techniques has been reported in many recent studies. Most of the solutions [17], [18], [19], [26], [47] depend on combining basic resampling techniques to improve detection accuracy. However, relying either on sample resampling or unrepresented synthesized samples can create problems. The main drawback of the simple resampling strategies is that they result in a biased posterior probability of the classifiers during the training. Synthetic-based resampling, which is the strategy most used by researchers [7], [9], [18], [22], [23], [24], [25] (such as in SMOTE and ADSYN) is highly dependent on the synthesized data. Although SMOTE and GAN have shown improvement in the detection performance, SMOTE increases the overlapped features between the target classes, leading to a slight decrease in the detection rate and increasing the rate of false alarms, while GAN increases the noise in the training sets, leading to imprecise learning. In addition, the highly imbalanced dataset contains insufficient samples to train the GAN network. Although the detection model using ensemble learning slightly improves the classification performance of deep learning, due to the diversity, and reduces overfitting, the decision regarding the correct class flows from the majority win strategy, and the majority of the classifiers are thus biased.

In this study, SMOTE and GAN are combined so that each technique overcomes the limitations of the other. Firstly, multiple subsets of samples were drawn from the original unbalanced dataset which contained the fraudulent sample in the training dataset with a larger sample from normal samples. SMOTE was then applied for each subset to create synthesized fraudulent samples. The k-nearest neighbor algorithm was utilized to generate a synthetic point in the feature space around the minority class. Drawing random samples in each subset makes the generated synthesis more diverse than applying the method to the whole dataset. The generated subsets were used to train an ensemble of diverse GAN models. In addition, the selection of random samples from the majority class makes distinguishing between the two classes easier. That is, every subset will have fewer overlapped features between the synthesized samples and the majority class. Thus, fewer noisy samples were predicted using GAN with fewer features overlapping. A base classifier was investigated to construct more diverse and cost-sensitive learning. More details of the proposed model are given in the following section.

III. METHODOLOGY

The proposed credit card fraud detection model (CCFDM) was constructed in three phases: the features extraction and pre-processing phase, the data resampling phase, and the model construction phase. Figure 1 shows the flowchart of the methodology used to construct the proposed CCFDM model.

A. DATA ACQUISITION AND PRE-PROCESSING PHASE

In general, the features that can be extracted from the transaction information include card-related features, transection-related features, and customer-related features [5], [48]. Card-related features include card numbers, card limits, and card expiry dates. Examples of transection-related features include the account number, transaction amount, transaction date and time, merchant ID, merchant location, point of sale, and category code among many others. Customer-related features comprise customer profile features, including cardholder ID, spending behavior such as average daily, weekly, and monthly spending, frequency of using credit cards, duration between transactions, and the time of the last transaction. Customer-related features are usually derived from the transaction history of the customer's benign transactions. That is, each customer has a profile constructed based on the user's past spending habits using statistical and probabilistic techniques. These features can be found in different data types such as numerical, categorical, and timestamp records. Thus, the common procedure is to preprocess the data to render it in numerical form. Data may need to be normalized based on the

techniques used for constructing the data models to be used for classification.

This study used the UBL dataset [5] with pre-augmented features and principal component analysis (PCA) was calculated for most of the features, to preserve the privacy and security of both customers and merchants. Due to concerns regarding the confidentiality of consumer transaction details, the majority of the features in the dataset were subjected to principal component analysis (PCA) to reduce dimensionality. PCA is a well-established and widely utilized method in the literature, which enables such datasets to become more interpretable while also minimizing the loss of information. The process of PCA involves generating independent variables that are uncorrelated from each other and maximizing the variance progressively. The dataset used consists of 31 features, including the time feature which is the time lapse between the current and first transaction. 28 of these features are the results of PCA dimensionality reduction and anonymization for protecting privacy (denoted by V1, V2, . . . , V28 in the dataset). The other features are the amount feature, which contains the amount of transactions, and the class label which is either 1 for fraudulent transactions or 0 for a normal transaction. A more detailed description of this dataset can be found in Section IV Part A. In this study, the features in the dataset were normalized to ensure that all variables in the dataset are on the same scale. Variables that are on vastly different scales may have a disproportionate impact on the analysis, leading to incorrect results. Normalization is essential to ensure that the new variables created by PCA have equal variances. Because PCA seeks to maximize variance, variables with larger variances may dominate the analysis, leading to inaccurate results.

B. DATA AUGMENTATION PHASE

The second phase aimed to prepare the training set of the data by solving the imbalanced class problem. The distribution is skewed towards the majority class. Credit card fraud samples are rarely available and contain sensitive information, due to the privacy issue of the transaction information. Among millions of transactions, few fraudulent transactions will occur (less than 1% in most cases). This unbalanced distribution of data makes constructing an effective unbiased detection model a challenging task. Therefore, it is important to increase the number of fraudulent transactions in the training set to handle imbalanced class problems. As mentioned earlier in the Related Work section, many techniques are used to increase the number of fraudulent samples. The synthetic Minority Oversampling Technique (SMOTE) was the most used in the literature. However, SMOTE has four main disadvantages: it oversamples the noisy sample; the accuracy of selecting the nearest neighbor depends on the data in hand (fraudulent transactions have overlapped features); it oversamples uninformative samples, and it focuses on local information, which results in a less diverse set of samples. On the other hand, a generative adversarial network (GAN) can learn the target distribution and accordingly generate artificial yet plausible samples from the same distribution. GAN has been reported as a useful technique for data augmentation due to its ability to simulate the distribution of real data [9], [18], [22], [27], [28], [31], [32].

GAN comprises two deep learning networks that are synergized to produce *realistic* samples (also called fake samples). The first network is called the generator, which is trained to generate samples (fake samples) from specific classes. The second network is known as the discriminator and is designed to determine if the generated fake sample is similar to a real sample. A high similarity between a real and fake sample means that the generator has successfully generated a fake sample while a low similarity value indicates that the generator needs more training. Thus, the models are recursively trained together, and training is stopped when the discriminator is fooled by at least half of the samples. This is done by rewarding the discriminator when correct classification is achieved and at the same time penalizing the generator by updating the model parameters. Alternatively, the generator is rewarded when it fools the discriminator and penalized if the discriminator correctly classifies the samples. That is when the discriminator cannot distinguish clearly between the fake and real samples and if the accuracy of the discriminator falls to around 50%, the training is stopped, indicating that the generative network is ready.

In this study, an ensemble technique based on SMOTE and GAN, called ESMOTE-GAN, is proposed. Because GAN needs a considerable amount of data to learn from, applying GAN for the fraudulent data set is not effective, due to the lack of enough samples to train the GAN. Although SMOTE can be used to generate the initial set of samples. it creates a less diverse set of data that does not well represent the frauds. To achieve the necessary diversity multiple unbalanced subsets of the dataset were created. The fraudulent samples were extracted from the original dataset and a random but larger sample size of normal transactions was extracted, such that the fraudulent samples comprised 10% of the subset. The 10% was selected to the trade-off between the SMOTE performance and the separability of the dataset. That is if the ratio between benign and fraudulent increases, SMOTE generates low-quality samples due to the increases in the overlapping features between samples. On the other hand, if the ratio between benign and fraudulent instances decreases, the fraudulent samples become unrepresentative, and overfitting will occur during the training. For each subset of the created dataset, SMOTE was used to balance the subset. Multiple GAN models were then, constructed based on the fraudulent samples generated using SMOTE. For training the generator, latent space is created from random vectors with lengths the same as those of the problem space. The generator is trained to learn the mapping between the latent space and problem space. Thus, a point in the latent space can be given as input to the generator to produce a realistic sample. Meanwhile, the discriminator tries to learn the mapping between feature vectors from either the problem or latent space and the sample label.

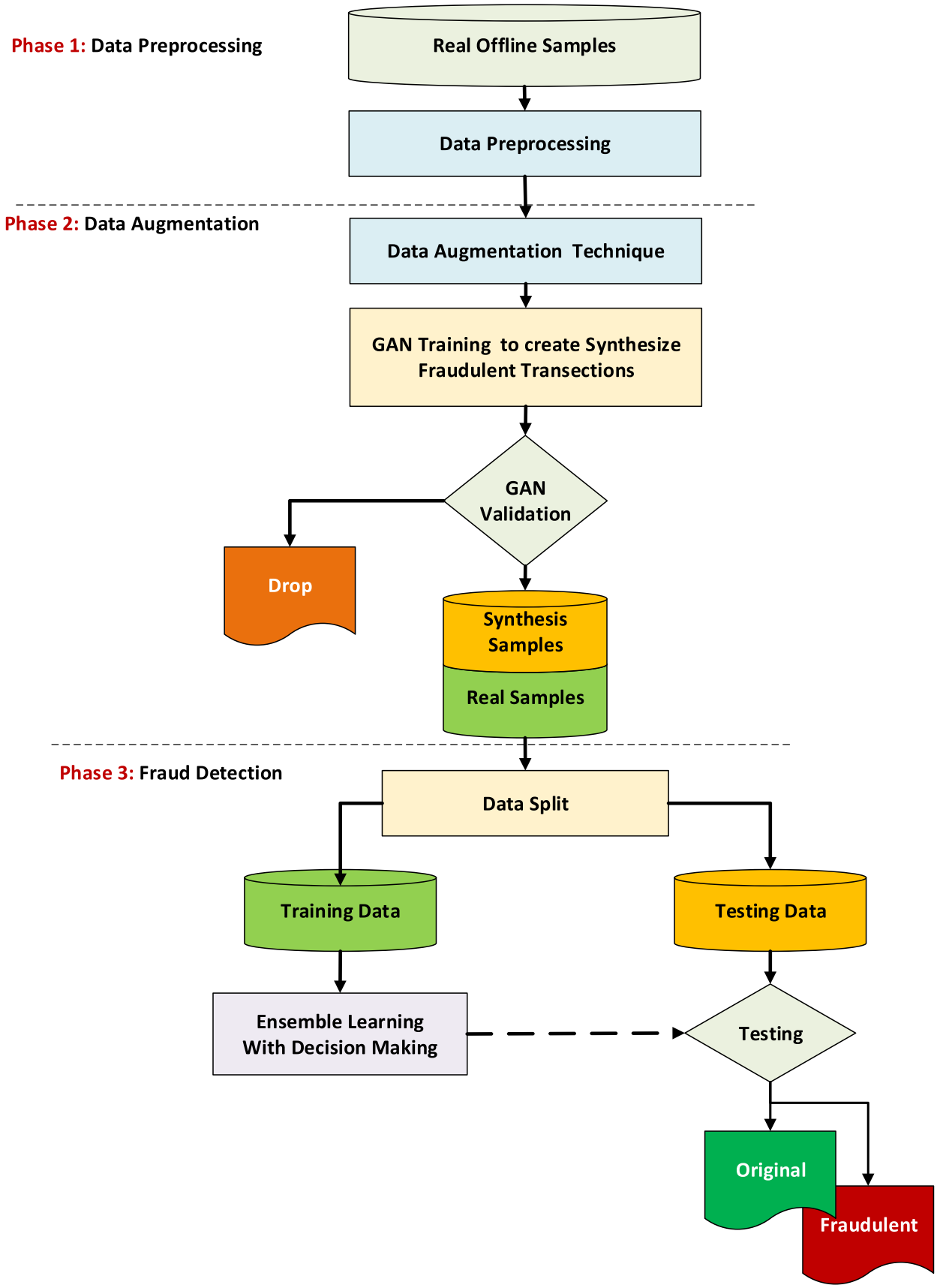


FIGURE 1. The methodology of the proposed CCFDM-based ESMOTE-GAN.

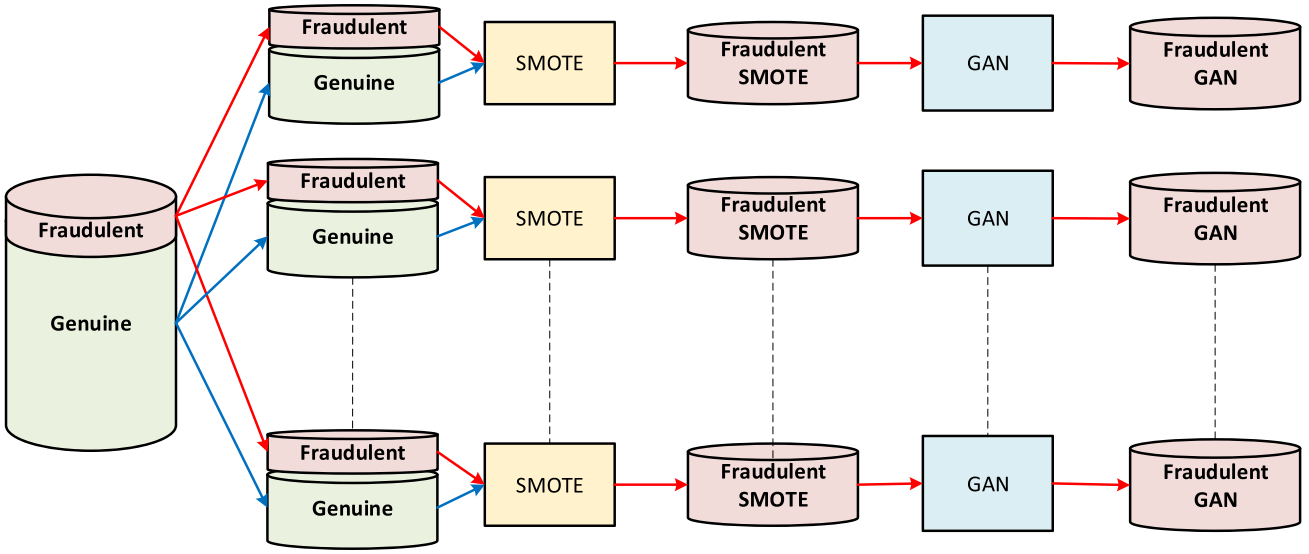


FIGURE 2. Ensemble SMOTE-GAN Model.

Figure 2 shows the architecture of the proposed ESMOTE-GAN Model, while Figure 3 shows the architecture of the GAN model. As can be seen in Figure 2, the ESMOTE-GAN model consists of two main layers: the SMOTE layer and the GAN layer. In the SMOTE layer, the random subsets resemble the dataset. Each set consists of the minority class samples and larger samples of the majority class, randomly resampled with a ratio of 10% of fraudulent transactions and 90% of benign transactions. In the second layer, multiple GAN models were trained based on the SMOTE outputs. As can be seen in Figure 3, the fraudulent transactions have been extracted and used for learning so that the GAN is trained on the imbalanced dataset generated by SMOTE with the fraudulent class as the target output. The generator network of the GAN is trained to produce synthetic samples that look similar to the fraud samples, while the discriminator network is trained to distinguish between the original and the synthetic samples. For the learning, let G and D denote the generator and the discriminator, respectively, and let $Z = \{z1, z2, \ldots, zn\}$ and $X = \{x1, x2, \ldots, xn\}$ denote the distribution of latent and problem space, respectively. G and D G(z) are the output of the generator (the fake sample) and $D(G(z))$ is the output of the discriminator, which is the probability of getting $G(z)$ belonging to real data. The error $e = log(1 - D(G(z))$ should be minimized to generate a fake sample that is drawn from the distribution of the real data. The error e is also used to penalize the generator G and thus to minimize $\log(D(x))$. Thus, the following min-max game must be played by G and D to minimize the generator error and maximize the divergence.

$$\underbrace{min}_{G}\ \underbrace{max}_{D} V(G, D) = E_x(\log(D(x))) + E_z(\log(1 - D(G(z)))) \quad (1)$$

The training of the GAN model continues until the generator can fool the discriminator into believing that the generated samples are real, namely when adversarial loss converges, indicating that the generator is producing realistic fraudulent samples. Algorithm 1 shows the steps used in the proposed data augmentation algorithm ESMOTE-GAN. Table 1 lists the description of the symbols presented in the algorithm. As shown in Algorithm 1, the decision to include the synthetic samples in the new dataset depends on the average of the probabilistic output of the discriminator's predictions of the fraudulent transactions(See Algorithm 1 Lines 17 and 18).

C. FRAUD DETECTION PHASE

In this phase, an ensemble model was constructed and trained based on the generated subsets of the generated dataset. For each synthetic sample generated by GANs, a normal sample is extracted from the dataset. Thus, multiple balanced sets were created for training. For each set, a diverse set of machine learning classifiers was trained, including KNN, CART, NB, ANN, SVM, LR, RF, XGBoost, and SDL. The best-performing classifier was selected as a base classifier for the proposed model. The probabilistic outputs of the trained classifiers were then used to construct the voting ensemble for the final decision.

$$p(y \mid x) = \frac{\sum_{i=1}^{n} w_i p_i(y \mid x)}{n} \quad (2)$$

where n is the number of classifiers and w_i is the weight of the classifier i and $p_i(y \mid x)$ is the probabilistic output of the classifier to predict the class y given the features x. The weight w_i is calculated based on the overall performance in terms of the F-measure of the trained classifiers.

IV. EXPERIMENTAL DESIGN

This section describes and discusses the experimental design used to validate and evaluate the proposed CCFDM model based on the proposed ESMOTE-GAN technique, including

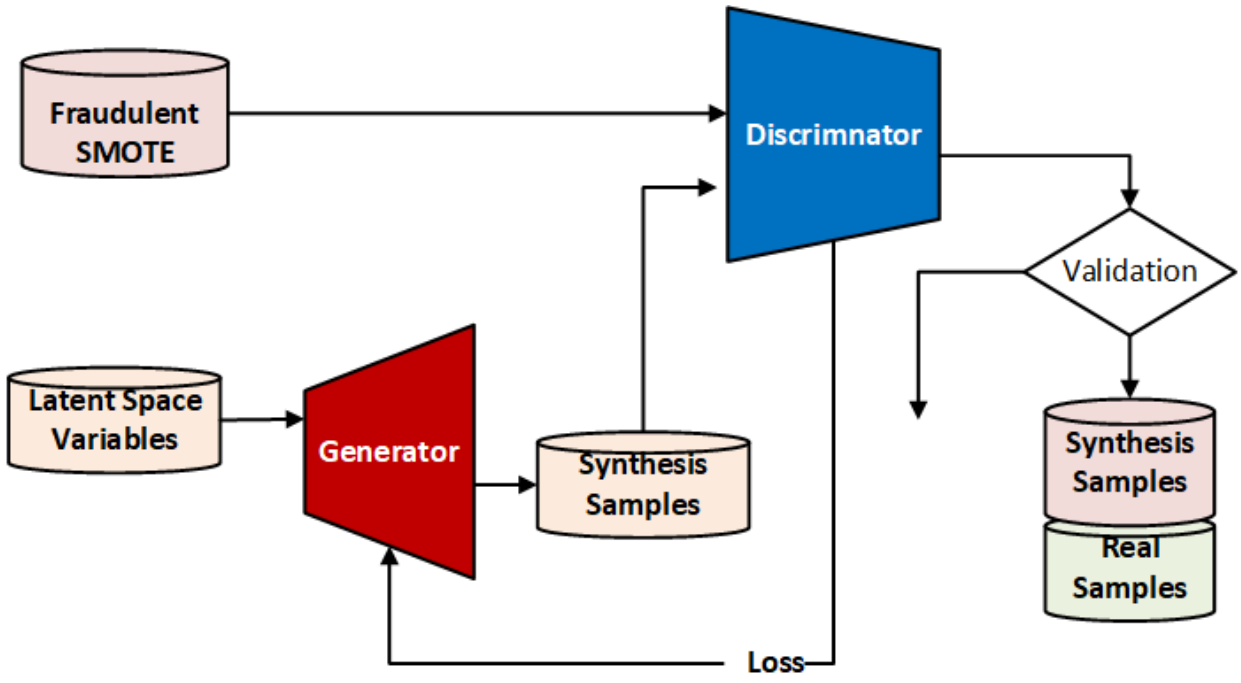


FIGURE 3. GAN Model.

Algorithm 1 Data Augmentation Technique ESMOTE-GAN

1: $for(i = 0, i < ensemble\ size, i++)$
2: $Df_1 \overset{extract}{\longleftarrow} DataSet[C == 1]$
3: $Df_{0,i} \overset{extract}{\longleftarrow} Randome\,(DataSet\,[C = 0]\,, n) : n > len(Df_1)$
4: $Df_i \overset{concatenate}{\longleftarrow} (Df_{0,i} Df_1)$
5: $Df_i' \overset{extract}{\longleftarrow} SMOTE(Df_i)$
6: $Df_{1,i}' \overset{extract}{\longleftarrow} Df'[C == 1]$
7: $for\ number\ of\ training\ iteration$
8: $Generate\ latent\ points\ z \epsilon Z$
9: $for\ k\ steps\ do$
10: $Sample\ batch\ Z_m\ from\ Z$
11: $Sample\ batch\ Df_{1,i,m}'\ from\ Df_{1,i}'$
12: $\underbrace{min}_{G}\ \underbrace{max}_{D}\ V\,(G, D) = E_x\,(\log\,(D\,(\mathrm{x}))) + E_z\,(\log\,(1 - D(G\,(\mathrm{z}))))$
13: $End\ loop$
14: $End\ loop$
15: $Z_{1,i}' \overset{predict\ using\ GAN}{\longleftarrow} G(Z)$
16: $C_i' \overset{predict\ using\ GAN}{\longleftarrow} Z_{1,i}' D(Z_{1,i}')$
17: $\forall c \in C_i'\ and\ z_{1,i}' \in Z_{1,i}' if\ \left(c < mean\,\left(C_i'\right)\right)$
18: $drop\ the\ sample\ z_{1,i}'\ from\ Z_{1,i}'$
19: End loop

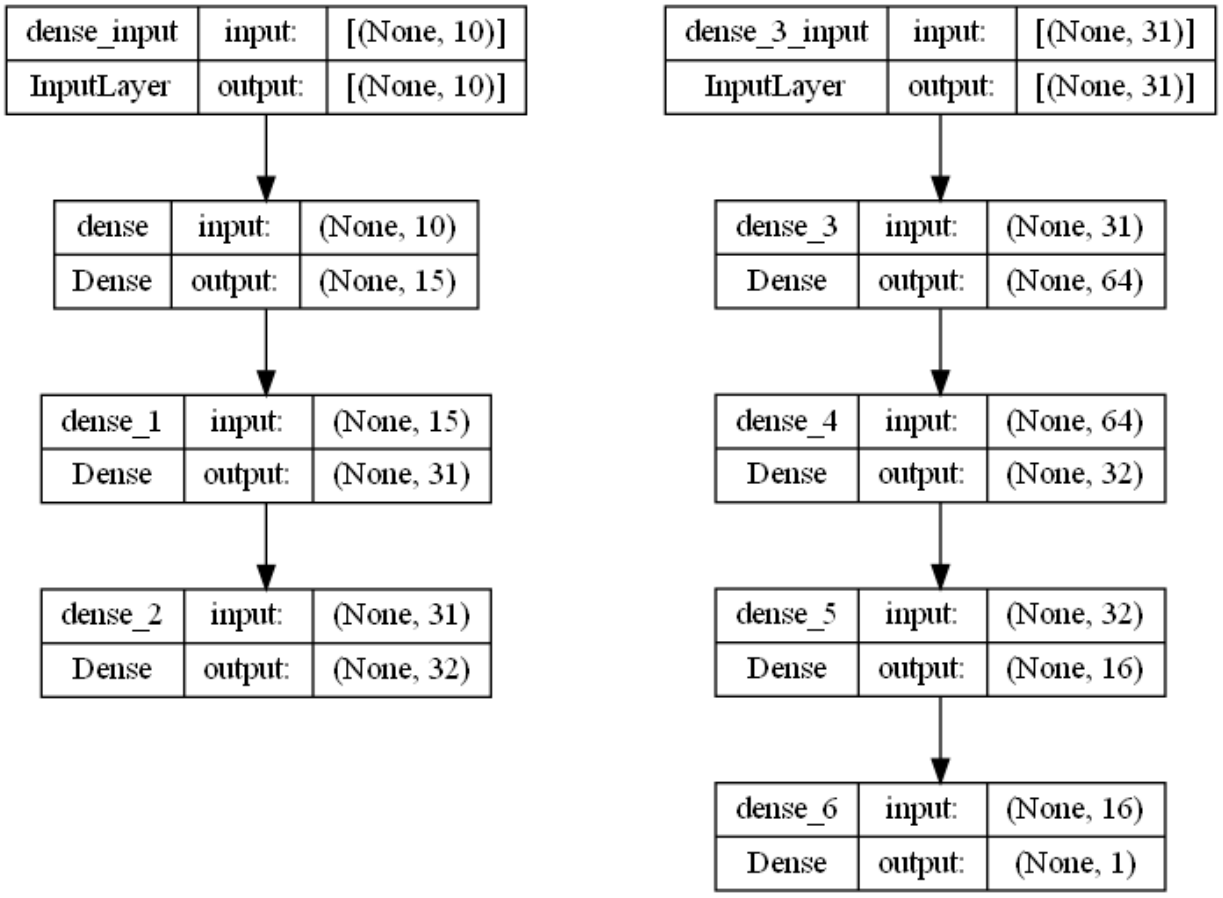
the dataset used, the performance measures, and the performance evaluation.

A. DATASET

The source of the dataset used in this study is the ULB Machine Learning Group (https://mlg.ulb.ac.be/wordpress/portfolio_page/defeatfraud-assessment-and-validation-of-deep-feature-engineering-and-learning-solutions-for-fraud- detection/). It can be downloaded directly from Kaggle repository (https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud). This repository contains data for 284,807 credit card transactions, as well as 492 samples of

TABLE 1. Description of the symbols.

Symbol	Description	Symbol	Description
Df_1	Set of the fraud transactions	G	The generator
C	Class label	D	The discriminator
$Df_{0,i}$	A random subset of benign transactions, i is the subset ID	Z	Random set of latent points
Df_i	An unbalanced random subset containing both fraud and benign transactions	$Z'_{1,i}$	Predicted fraudulent transactions using the generator
Df'_i	Balanced random subset containing both fraud and benign samples	C'_i	The discriminator output is the probability of predicting the synthetic fraud transactions between 0-1



(a) The Generator

(b) The Discriminator

FIGURE 4. GAN Architecture.

fraudulent transactions collected from European cardholders. The dataset is highly imbalanced, and the percentage of the fraud dataset is 0.172% out of all transactions. The dataset contains a total of 30 features, in addition to one column for the class labels, of which 28 are anonymized features, and two features are for the time and amount of the transaction. The 28 features denoted by V1, ... V28 in the dataset are the results of PCA dimensionality reduction and anonymization for protecting privacy. The time feature is the time elapsed between the current and first transaction while the amount features comprise the amount of the transaction and the class label, which is either 1 for fraudulent transactions or 0 for genuine transactions. The dataset has been commonly used by many researchers [11], [47], [49], [50].

B. PERFORMANCE MEASURES

The proposed model was validated using the commonly used performance metrics in the literature, which include, accuracy, precision, recall, F-Measure, and the false alarm rate. The recall is the measure of the false positive rate and can be used as a measure of investigation cost. The F measure is an effective performance indicator for unbalanced data which combines the precision and recall measures to evaluate the overall performance of the model. These metrics can be

calculated using the following formulae.

$$Recall = \frac{\#\ fradulent\ samples\ correctly\ classified}{number\ of\ actual\ fradulent\ samples} \tag{3}$$

$$False\ Alarms = \frac{\#\ normal\ samples\ wrongly\ classified}{total\ number\ of\ normal\ samples} \tag{4}$$

$$Precision = \frac{\#fradulent\ samples\ correctly\ classified}{number\ of\ samples\ classified\ as\ fradulent} \tag{5}$$

$$F - Measure = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{6}$$

C. PERFORMANCE EVALUATION

The proposed model was evaluated by comparing its performance with the models from related work. First, the related data augmentation techniques (Oversampling [17], [19], SMOTE [7], [22], [23], [25], and GAN [28]) were used to benchmark the performance of the proposed SMOTE-GAN and ESMOTE-GAN techniques. Three strategies were implemented to evaluate the proposed SMOTE-GAN and ESMOTE-GAN data augmentation techniques: cost-sensitivity, resampling, and ensemble.

For cost-sensitive multiple machine and deep learning, the classifiers were trained based on the original unbalanced dataset. The probabilistic output of each classifier was tuned to obtain the best performance. For the resampling strategy, oversampling, SMOTE, and GAN techniques were selected to generate the synthesized dataset. The set of common machine learning-based classifiers with training was then tested on the unbalanced test set. The datasets generated by the related resampling and the proposed augmentation techniques were also used to train ensemble-based learning classifiers such as RF and XGBoost for comparison. The algorithms used in this study were implemented using Python 3.9.16. The algorithms KNN, CART, SVM, ANN, LR, RF, and XGBoost were implemented using the Scikit-learn (Sklearn) 1.0.2 library, while the deep learning was implemented using TensorFlow Keras models 2.9.1. Table 2 shows the parameters used for the experiments in this study.

D. RESULTS AND DISCUSSION

Table 3 and Figures 5, 6, 7, and 8 show the overall performance in terms of the F-Measure of the proposed model CCFDM as compared to the related models. The proposed CCFDM is referred to as the ESMOTE-GAN augmentation technique. To ensure the practical applicability of the model on real-world data, we evaluated its performance using the unbalanced test data. As shown in Figures 5, 6, 7, and 8, the performance varies based on the sampling technique and machine learning classifier used. To provide a meaningful comparison, we selected four recent and top-performing models for evaluation: the CNN proposed by the authors in [5], ACL and FCL proposed by the authors in [41], and H-KNN proposed by the authors in [36]. By leveraging these comparable models, we obtained valuable insights into the effectiveness of our proposed CCFDM model.

TABLE 2. Learning parameters.

Algorithm	Parameters
SMOTE	$k\_neighbors = 5$
KNN	$n_{neighbors} = 5, algorithm = auto,$ $uniform\ weights\ leaf_{size} = 30, euclidean\_distance,$
CART	$Split = Gini\ impurity,$ min $\_samples\_split = 2$ $min\_samples\_leaf = 1$
SVM	Regularization parameter C =1, kernel=rfp, gamma=scale
ANN	$solver = lbfgs, alpha = 1e-5, hidden\_layer\_sizes =$ $(15,), random\_state = 1$
LR	
RF	$n\_estimators = 100, random\_state = 42$
XGBoost	$n\_estimators = 100, pred\_leaf = True, 0$ $scale\_pos\_weight = 0.5$
SDL	$Dense\ Layers = 4, Size(128,64,32,16)$ $activation = relu, output\_activation = sigmoid$

As depicted in Table 3, the proposed ESMOTE-GAN augmentation method demonstrated superior performance across various types of classifiers. Particularly, the RF and XGBoost algorithms outperformed other classifiers in terms of overall performance. In terms of F-Measure, RF achieved an overall performance of 92.31, while XGBoost achieved a slightly higher performance of 92.44. On the other hand, LR exhibited the lowest performance among the classifiers. This discrepancy in performance can be attributed to the nature of the dataset and the characteristics of the classifiers. Tree-based classifiers, such as RF and XGBoost, are well-suited for handling non-linear and noisy datasets. They can capture complex relationships and patterns in the data, which leads to improved performance. In contrast, LR attempts to learn a linear decision boundary in the dataset, which may cause a degradation in performance when dealing with non-linear data. Additionally, LR is sensitive to extreme outliers, which can further lower its performance. It is noteworthy that the LR model performed relatively well without the need for resampling techniques and achieved comparable performance even when the dataset was unbalanced (as shown in Table 3). This can be interpreted as demonstrating LR's ability to handle imbalanced datasets and adapt to the inherent class distribution.

Although artificial neural networks (ANN) and deep learning algorithms can effectively handle non-linear datasets, it is worth noting that tree-based classifiers often outperform neural network-based algorithms in high-dimensional and noisy data scenarios. The decision trees employed in tree-based classifiers are capable of capturing complex relationships and handling noisy features, making them well-suited for such datasets.

The subset dataset generated by applying SMOTE and GAN augmentation techniques may contain some noise. This is because the SMOTE algorithm introduces approximations in the first layer, and the GAN model further adds complexity and variability in the second layer. Consequently, the

TABLE 3. Performance comparison in terms of f-measure.

Base Model	Unbalanced	Data Augmentation Technique				
		Oversampling	SMOTE	GAN	SMOTE-GAN	ESMOTE-GAN
KNN	80.15	80.48	85.71	81.33	85.71	85.25
CART	75.86	80.74	76.47	68.67	75.00	78.74
SVM	78.71	79.27	89.47	65.18	85.45	87.50
ANN	78.49	84.13	84.55	82.86	85.22	85.71
LR	83.33	66.30	69.72	46.88	69.72	49.66
RF	72.43	89.46	90.43	88.12	91.38	92.31
XGBoost	83.15	90.14	91.53	87.18	91.53	92.44
SDL	86.87	86.87	83.46	90.10	83.97	88.19

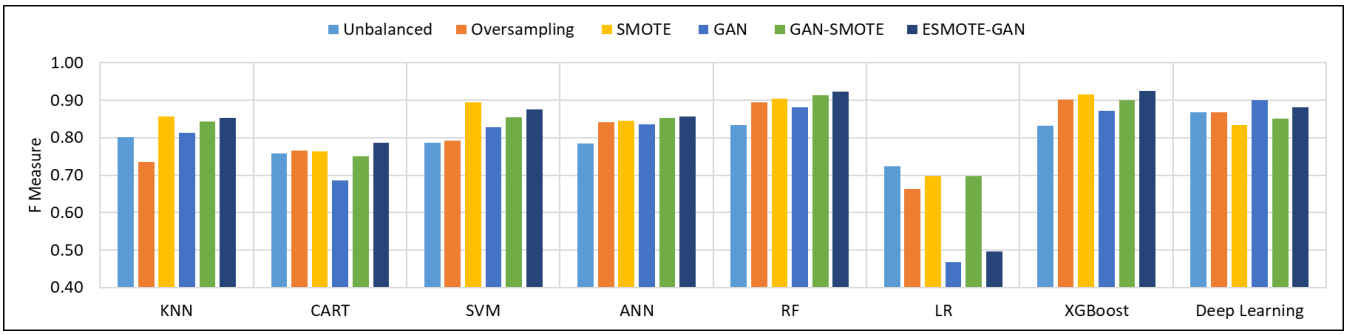


FIGURE 5. Overall Performance Comparison in Terms of F-Measure.

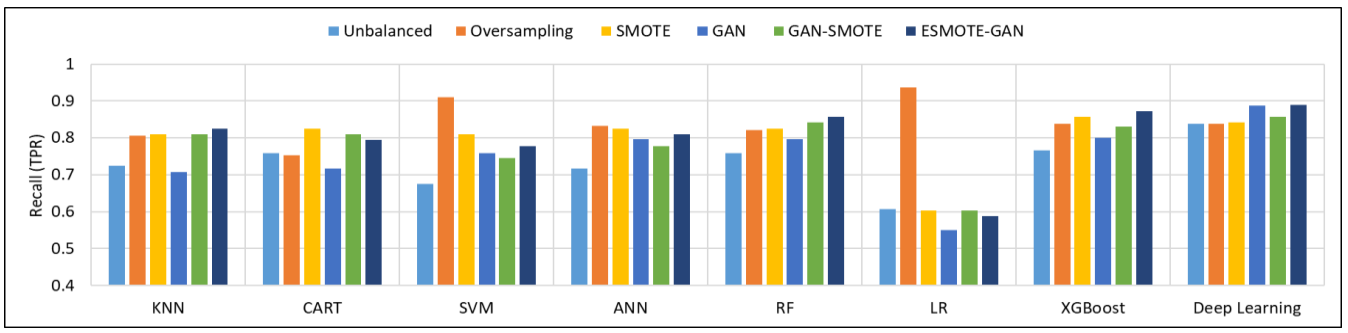


FIGURE 6. Performance Comparison in Terms of True Positive Rate (Recall).

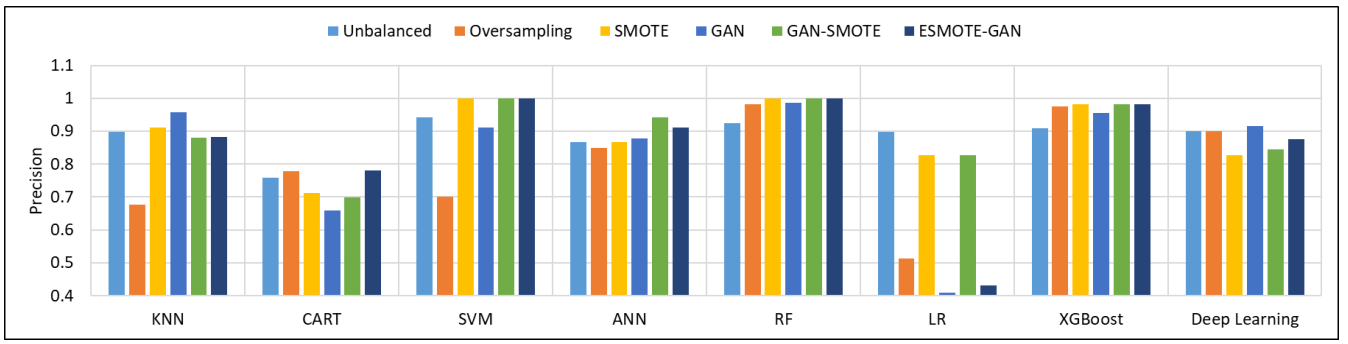


FIGURE 7. Performance Comparison in Terms of Precision.

resulting dataset may exhibit noisy characteristics. It is important to consider this aspect when analyzing the performance of the models. GAN treats the output of SMOTE as a realistic dataset to learn from. However, such a treatment does not hold for highly imbalanced datasets. The combination of SMOTE and GAN augmentation techniques can introduce

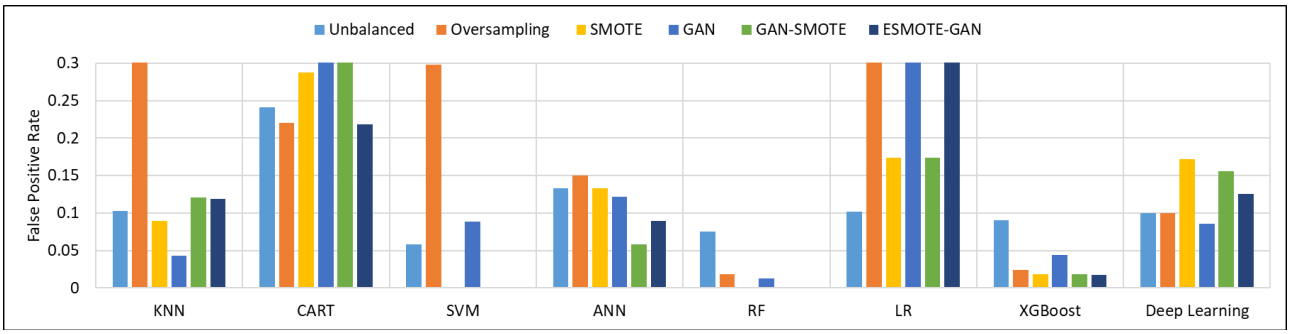


FIGURE 8. Performance Comparison in Terms of False Positive Rate (FPR).

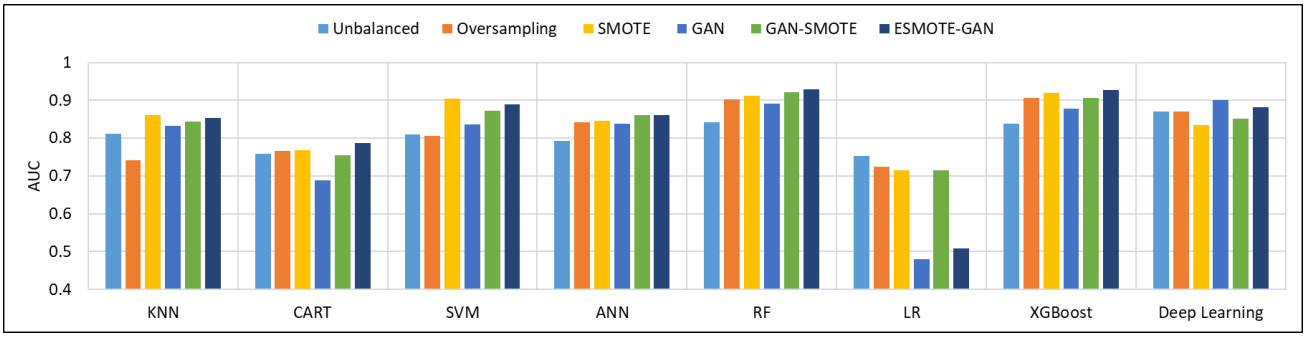


FIGURE 9. Performance Comparison in Terms of Area Under the Curve (AUC).

additional noise and variations that might affect the learning process and overall performance, especially in the context of highly imbalanced data. This interoperates why the proposed ESMOTE-GAN outperformed the traditional stacked SMOTE-GAN. The use of ensemble GAN classifiers which learned from a diverse set of data sub-sets improved the diversity and further improved the classification performance.

Figures 6 and 7 show the detection rate in terms of true positive rate (recall) and precision. In terms of the detection rate, the proposed ESMOTE-GAN model is better than most of the other studied models. Although the LR and SVM models that were built using the oversampling technique achieved better detection rates than ESMOTE-GAN, both SVM, and LR achieved lower precision than the others. This could be due to the complexity of the decision boundary that separates the classes in some subsets of the synthetic data. Logistic regression and SVM are both linear models, meaning they can only learn linear decision boundaries. If the decision boundary in the synthetic data is highly non-linear, these models may struggle to fit it accurately, resulting in lower precision. The choice of hyperparameters for these models can also impact their precision on synthetic data. For example, the choice of the regularization parameter in logistic regression or the choice of kernel function and kernel parameters in SVM can affect their performance. RF and XGBoost algorithms strike a balance between detection rate and precision. While the recall is an indication of the predictability of the fraudulent samples, the precision is a measure of model performance in terms of discriminating between fraudulent and genuine or truthful transactions. The higher the recall, the lower the number of undetected frauds, while the higher the precision, the lower the number of wrongly classified genuine transactions.

Figure 8 shows the false positive rates. SVM and RF both reduced the false positive rate to zero, due to their ability to achieve a 100% precision rate, as compared to the others. However, RF achieves better performance than SVM in terms of detection rate. The false positive rate measure is important because it indicates the intensity of human intervention needed in verifying the alarms. Although the proposed ESMOTE-GAN using XGBoost achieved higher recall compared with that using RF, the false positive rate produced by XGBoost is the disadvantage of such a model. Even a small rate of false positives means a heavy workload is required for human investigation. Given millions of benign transactions every day, XGBoost, which attains 1.7% FPR, can lead to 17,000 false alarms that need to be analyzed.

Figure 9 presents the overall performance score, in the form of the area under the curve (AUC) measure using the Receiving Operating Characteristic (ROC) curve. The process of generating the ROC curve involves the calculation of TPR and FPR for all the feasible probabilities of fraud produced by a classifier [51]. As can be seen in Figure 9, the proposed SMOTE-GAN and ESMOTE-GAN models achieved 92.8% and 92.9% AUC with XGBoost and RF classifiers, respectively. This means that there is more than a 93% probability of correctly classifying fraudulent transactions.

TABLE 4. The improvement gains in terms of f-measure.

Base Model	Unbalanced	Data Augmentation Technique			
		Oversampling	SMOTE	GAN	SMOTE-GAN
KNN	5.1	4.77	-0.46	3.92	-0.46
CART	2.88	-2	2.27	10.07	3.74
SVM	8.79	8.23	-1.97	22.32	2.05
ANN	7.22	1.58	1.16	2.85	0.49
LR	-33.67	-16.64	-20.06	2.78	-20.06
RF	19.88	2.85	1.88	4.19	0.93
XGBoost	9.29	2.3	0.91	5.26	0.91
DSL	1.32	1.32	4.73	-1.91	4.22

TABLE 5. The improvement gains in terms of f-measure.

Model	Sampling Method	Classification Model	F-Measure (%)
Alarfaj, Malik, Khan, Almusallam, Ramzan and Ahmed [5]	-	CNN	85.7
Li, Liu, and Jiang [41]	SMOTE-ENN	ACL with deep learning	81.9
Li, Liu, and Jiang [41]	SMOTE	FCL with deep learning	82.2
Awoyemi, Adetunmbi and Oluwadare [36]	Hybrid	K-NN	84.3
The proposed Model	ESMOTE-GAN	Ensemble RF with a weighted Voting Scheme	92.31

The ESMOTE-GAN model achieved the best performance in terms of AUC, followed by the SMOTE-GAN model. The best performance achieved by SMOTE was 92% and 92.3% AUC with XGBoost and RF, respectively. Oversampling with XGBoost achieved 90.7%, while deep learning achieved the best performance, 87%, using the unbalanced dataset, However, classifiers with the unbalanced dataset achieved the lowest performance among the studied classifiers.

Table 4 shows the gains in terms of F-measure by the proposed ESMOTE_GAN model compared to the state-of-the-art models. It can be noticed that the proposed ensemble-based integration of SMOTE and GAN outperformed all other related models with most machine learning classifiers. The most significant improvement was achieved using the ensemble sequential deep learning technique (SDL) while LR failed to learn from the created synthesized datasets. The overall performance of KNN also slightly dropped using synthesized datasets created by the proposed technique as compared to SMOTE, GAN, and SMOTE-GAN techniques. The reason for such failure is that LR and KNN are very sensitive to noise: they are dependent on the quality of the dataset. Tree-based classifiers such as CART, RF, and XGBoost achieve consistent improvement concerning the compared techniques.

Figure 10 and Table 5 compare the performance of the proposed model with that of the related work. As can be seen in Figure 10 and Table 5, the proposed model outperforms the other studied models. In the model proposed by Alarfaj, Malik, Khan, Almusallam, Ramzan, and Ahmed [5], the CNN was trained based on highly skewed data in which the features of the minority class are insufficient to learn from. The ACL with deep learning proposed by Li, Liu, and Jiang [41] achieved the lowest detection performance. The FCL with deep learning proposed by Li and Liu [38] achieved slightly higher performance compared to the ACL

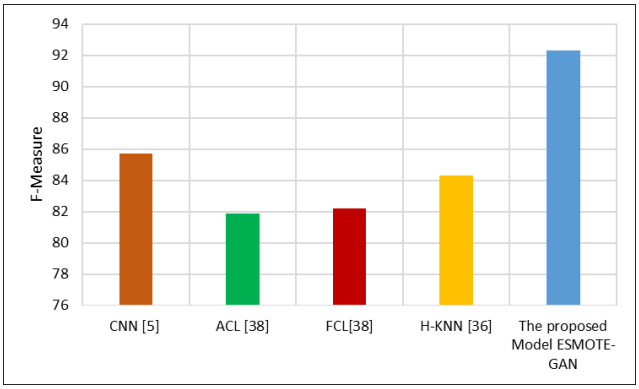


FIGURE 10. Performance Comparison with the related work.

because it is an improvement of the SMOTE, where the angle was used with the distance to generate the synthetic samples for oversampling. The hybrid resampling techniques using both oversampling for the minority class and under-sampling for majority class achieved better performance compared to FCL. Although such a technique can be effective, it suffers from the overfitting problem. The proposed ESMOTE-GAN solves this problem by creating multiple sets of the balanced datasets, using SMOTE with an under-sampling technique. Thus, the trained ensemble GANs together produced more representative features with less noise and fewer overlapped features.

E. THREATS TO VALIDITY

In this subsection, the factors that affect the performance of the proposed fraud detection model are discussed. The imbalanced datasets where the number of fraudulent transactions is much lower than the number of legitimate transactions leads to biased training models. As a result, the trained models

tend to be more accurate in detecting legitimate transactions but less effective in identifying fraudulent ones, due to the lack of representative features during training. To address this issue, our proposed model adopts a two-step approach. First, we reduce the class imbalance by employing the under-sampling technique, which helps narrow the gap between the classes. SMOTE is then utilized to increase the size of the minority class within each subset. By doing so, we can extract diverse patterns from these subsets, resulting in fewer overlapping features. To prevent the GAN from modeling the noise, we apply partial oversampling, using SMOTE on the minority class. This approach allows the trained GAN networks to generate more accurate samples by reducing the noise introduced by SMOTE. Ultimately, the ensemble of SMOTE-GAN further mitigates the impact of overlapped features on the model's performance. Eventually, the proposed ensemble of SMOTE and GAN (ESMOTE-GAN) further eliminates the impact of overlapped features on the model's performance. However, factors such as GAN architecture and the ratio between the classes in the training set impact the performance. In this study, the architecture and size of the sample sets were selected based on trial and error, and the best-performing combination was reported.

Another critical factor that can impact the performance of the fraud detection model is the selection of an appropriate learning algorithm that can effectively learn from diverse sets of samples. Ensemble learning techniques have often been employed to address the overfitting problem and enhance overall performance by leveraging the strengths of multiple classifiers. Deep learning techniques, such as CNN, LSTM, and Autoencoders, have shown promising results in various domains and may potentially yield better performance compared to the proposed ensemble RF model. However, their specific applicability and performance in the context of fraud detection warrant further investigation, which can be explored as part of future work.

V. CONCLUSION

In this study, a feature augmentation technique has been proposed to address the problem of highly class-imbalanced datasets. Multiple sets of samples were created using under-sampling techniques and used as input for SMOTE. The SMOTE technique and GAN model were then cascaded to generate multiple balanced subsets with fewer overlapping and noisy features. An ensemble of diverse random forest-based classifiers was trained to develop a credit card fraud detection model based on the proposed ESMOTE-GAN model. The results show the effectiveness of the proposed argumentation technique and the detection model. The proposed SMOTE-GAN-based model achieved a 3.2% improvement in the detection rate and a 1.9% improvement in the overall detection performance with a 0% false alarm rate. This result implies a reduction of the cost needed for analysis by humans.

One of the limitations of the proposed model is that some synthetic samples generated by SMOTE are too similar to existing minority class samples, leading to the overfitting problem. One possible solution is to use the ENN technique (the edited nearest neighbor technique) to remove the noisy samples from both classes. In addition, the performance of deep learning techniques such as autoencoding with the proposed ESMOTE-GAN augmentation technique needs to be further investigated. A study of the concept drift problem is needed on how the proposed diverse class can be automatically updated using the gradual replacement of the bias classifier based on the new arrival of wrongly classified fraudulent transactions. Further investigation is required to determine if the proposed solutions can be generalized for other domains with rare events, such as anomaly detection, medical diagnosis (e.g. cancer detection), or predicting rare natural disasters such as earthquakes.

ACKNOWLEDGMENT

The authors extend their appreciation to the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU) for funding and supporting this work through Research Partnership Program no RP-21-07-09.

REFERENCES

[1] Y. Fang, Y. Zhang, and C. Huang, "Credit card fraud detection based on machine learning," *Comput., Mater. Continua*, vol. 61, no. 1, pp. 185–195, 2019.

[2] C. Mullen. (Dec. 24, 2022). *Card Industry Faces $400B in Fraud Losses Over Next Decade*. [Online]. Available: https://www.paymentsdive.com/news/card-industry-faces-400b-in-fraud-losses-over-next-decade-nilson-says/611521/

[3] F. Kamalov and D. Denisov, "Gamma distribution-based sampling for imbalanced data," *Knowl.-Based Syst.*, vol. 207, Nov. 2020, Art. no. 106368.

[4] G. Rekha, A. K. Tyagi, N. Sreenath, and S. Mishra, "Class imbalanced data: Open issues and future research directions," in *Proc. Int. Conf. Comput. Commun. Informat. (ICCCI)*, Jan. 2021, pp. 1–6.

[5] F. K. Alarfaj, I. Malik, H. U. Khan, N. Almusallam, M. Ramzan, and M. Ahmed, "Credit card fraud detection using state-of-the-art machine learning and deep learning algorithms," *IEEE Access*, vol. 10, pp. 39700–39715, 2022.

[6] J. R. Dorronsoro, F. Ginel, C. Sgnchez, and C. S. Cruz, "Neural fraud detection in credit card operations," *IEEE Trans. Neural Netw.*, vol. 8, no. 4, pp. 827–834, Jul. 1997.

[7] 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.

[8] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning From Imbalanced Data Sets*. Cham, Switzerland: Springer, 2018.

[9] S. Ganguly and S. Sadaoui, "Classification of imbalanced auction fraud data," in *Proc. Can. Conf. Artif. Intell.*, 2017, pp. 84–89.

[10] 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.

[11] A. Dal Pozzolo, G. Boracchi, O. Caelen, C. Alippi, and G. Bontempi, "Credit card fraud detection: A realistic modeling and a novel learning strategy," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 8, pp. 3784–3797, Aug. 2018.

[12] C. Shen, S.-F. Zhang, J.-H. Zhai, D.-S. Luo, and J.-F. Chen, "Imbalanced data classification based on extreme learning machine autoencoder," in *Proc. Int. Conf. Mach. Learn. Cybern. (ICMLC)*, vol. 2, Jul. 2018, pp. 399–404.

[13] H. Wang, W. Wang, Y. Liu, and B. Alidaee, "Integrating machine learning algorithms with quantum annealing solvers for online fraud detection," *IEEE Access*, vol. 10, pp. 75908–75917, 2022.

[14] E. Wu, H. Cui, and R. E. Welsch, "Dual autoencoders generative adversarial network for imbalanced classification problem," *IEEE Access*, vol. 8, pp. 91265–91275, 2020.

[15] S. Bagga, A. Goyal, N. Gupta, and A. Goyal, "Credit card fraud detection using pipeling and ensemble learning," *Proc. Comput. Sci.*, vol. 173, pp. 104–112, Jan. 2020.

[16] 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.

[17] V. A. Fajardo, D. Findlay, C. Jaiswal, X. Yin, R. Houmanfar, H. Xie, J. Liang, X. She, and D. B. Emerson, "On oversampling imbalanced data with deep conditional generative models," *Expert Syst. Appl.*, vol. 169, May 2021, Art. no. 114463.

[18] Y.-Y. Hsin, T.-S. Dai, Y.-W. Ti, M.-C. Huang, T.-H. Chiang, and L.-C. Liu, "Feature engineering and resampling strategies for fund transfer fraud with limited transaction data and a time-inhomogeneous modi operandi," *IEEE Access*, vol. 10, pp. 86101–86116, 2022.

[19] Y.-J. Lee, Y.-R. Yeh, and Y. F. Wang, "Anomaly detection via online oversampling principal component analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 7, pp. 1460–1470, Jul. 2013.

[20] Y. Lucas and J. Jurgovsky, "Credit card fraud detection using machine learning: A survey," 2020, *arXiv:2010.06479*.

[21] European Central Bank. (Dec. 24, 2022). *Sixth Report on Card Fraud*. [Online]. Available: https://www.ecb.europa.eu/pub/cardfraud/html/ecb.cardfraudreport202008 521edb602b.en.html#toc2

[22] Q. Chen, Z.-L. Zhang, W.-P. Huang, J. Wu, and X.-G. Luo, "PF-SMOTE: A novel parameter-free SMOTE for imbalanced datasets," *Neurocomputing*, vol. 498, pp. 75–88, Aug. 2022.

[23] E. B. Fatima, B. Omar, E. M. Abdelmajid, F. Rustam, A. Mehmood, and G. S. Choi, "Minimizing the overlapping degree to improve class-imbalanced learning under sparse feature selection: Application to fraud detection," *IEEE Access*, vol. 9, pp. 28101–28110, 2021.

[24] E. Ileberi, Y. Sun, and Z. Wang, "Performance evaluation of machine learning methods for credit card fraud detection using SMOTE and AdaBoost," *IEEE Access*, vol. 9, pp. 165286–165294, 2021.

[25] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[26] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proc. IEEE Int. Joint Conf. Neural Netw. (IEEE World Congr. Comput. Intell.)*, Jun. 2008, pp. 1322–1328.

[27] W. A. Al-Khater, S. Al-Maadeed, A. A. Ahmed, A. S. Sadiq, and M. K. Khan, "Comprehensive review of cybercrime detection techniques," *IEEE Access*, vol. 8, pp. 137293–137311, 2020.

[28] H. Yang and Y. Zhou, "IDA-GAN: A novel imbalanced data augmentation GAN," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 8299–8305.

[29] A. Sharma, P. K. Singh, and R. Chandra, "SMOTified-GAN for class imbalanced pattern classification problems," *IEEE Access*, vol. 10, pp. 30655–30665, 2022.

[30] H. Ali, M. N. M. Salleh, K. Hussain, A. Ahmad, A. Ullah, A. Muhammad, R. Naseem, and M. Khan, "A review on data preprocessing methods for class imbalance problem," *Int. J. Eng. Technol.*, vol. 8, pp. 390–397, Jan. 2019.

[31] 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.

[32] H. Tingfei, C. Guangquan, and H. Kuihua, "Using variational auto encoding in credit card fraud detection," *IEEE Access*, vol. 8, pp. 149841–149853, 2020.

[33] 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.

[34] F. Ogme, A. G. Yavuz, M. A. Guvensan, and M. E. Karsligil, "Temporal transaction scraping assisted point of compromise detection with autoencoder based feature engineering," *IEEE Access*, vol. 9, pp. 109536–109547, 2021.

[35] J. F. Roseline, G. Naidu, V. S. Pandi, S. A. A. Rajasree, and D. N. Mageswari, "Autonomous credit card fraud detection using machine learning approach?" *Comput. Electr. Eng.*, vol. 102, Sep. 2022, Art. no. 108132.

[36] J. O. Awoyemi, A. O. Adetunmbi, and S. A. Oluwadare, "Credit card fraud detection using machine learning techniques: A comparative analysis," in *Proc. Int. Conf. Comput. Netw. Informat. (ICCNI)*, Oct. 2017, pp. 1–9.

[37] 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.

[38] S. Rajora, D. L. Li, C. Jha, N. Bharill, O. P. Patel, S. Joshi, D. Puthal, and M. Prasad, "A comparative study of machine learning techniques for credit card fraud detection based on time variance," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Nov. 2018, pp. 1958–1963.

[39] 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.

[40] S. K. Hashemi, S. L. Mirtaheri, and S. Greco, "Fraud detection in banking data by machine learning techniques," *IEEE Access*, vol. 11, pp. 3034–3043, 2023.

[41] Z. Li, G. Liu, and C. Jiang, "Deep representation learning with full center loss for credit card fraud detection," *IEEE Trans. Computat. Social Syst.*, vol. 7, no. 2, pp. 569–579, Apr. 2020.

[42] 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.

[43] S. A. Ebiaredoh-Mienye, E. Esenogho, and T. G. Swart, "Artificial neural network technique for improving prediction of credit card default: A stacked sparse autoencoder approach," *Int. J. Electr. Comput. Eng. (IJECE)*, vol. 11, no. 5, p. 4392, Oct. 2021.

[44] C. Jiang, J. Song, G. Liu, L. Zheng, and W. Luan, "Credit card fraud detection: A novel approach using aggregation strategy and feedback mechanism," *IEEE Internet Things J.*, vol. 5, no. 5, pp. 3637–3647, Oct. 2018.

[45] 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.

[46] A. Srivastava, A. Kundu, S. Sural, and A. K. Majumdar, "Credit card fraud detection using hidden Markov model," *IEEE Trans. Depend. Secure Comput.*, vol. 5, no. 1, pp. 37–48, Jan. 2008.

[47] A. D. Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi, "Calibrating probability with undersampling for unbalanced classification," in *Proc. IEEE Symp. Ser. Comput. Intell.*, Dec. 2015, pp. 159–166.

[48] I. Benchaji, S. Douzi, and B. E. Ouahidi, "Credit card fraud detection model based on LSTM recurrent neural networks," *J. Adv. Inf. Technol.*, vol. 12, no. 2, pp. 113–118, 2021.

[49] F. Carcillo, Y.-A. Le Borgne, O. Caelen, Y. Kessaci, F. Oblé, and G. Bontempi, "Combining unsupervised and supervised learning in credit card fraud detection," *Inf. Sci.*, vol. 557, pp. 317–331, May 2021.

[50] F. Carcillo, A. Dal Pozzolo, Y.-A. Le Borgne, O. Caelen, Y. Mazzer, and G. Bontempi, "SCARFF: A scalable framework for streaming credit card fraud detection with spark," *Inf. Fusion*, vol. 41, pp. 182–194, May 2018.

[51] Y.-A. Le Borgne, W. Siblini, B. Lebichot, and G. Bontempi, *Reproducible Machine Learning for Credit Card Fraud Detection—Practical Handbook*. Brussels, Belgium: Université Libre de Bruxelles, 2022. [Online]. Available: https://github.com/Fraud-Detection-Handbook/fraud-detection-handbook and https://fraud-detection-handbook.github.io/fraud-detection-handbook/Foreword.html

FUAD A. GHALEB received the B.Sc. degree in computer engineering from the Faculty of Engineering, Sana'a University, Yemen, in 2003, and the M.Sc. and Ph.D. degrees in computer science (information security) from the Faculty of Engineering, School of Computing, Universiti Teknologi Malaysia (UTM), Johor, Malaysia, in 2014 and 2018, respectively. He is currently a Senior Lecturer with the Faculty of Engineering, School of Computing, UTM. His research interests include vehicular network security, cyber security, intrusion detection, data science, data mining, and artificial intelligence. He was a recipient of many awards and recognitions, such as the Postdoctoral Fellowship Award, the Best Postgraduate Student Award, the Excellence Awards, and the Best Presenter Award from the School of Computing, Faculty of Engineering, UTM, and the best paper awards from many international conferences.

FAISAL SAEED (Member, IEEE) received the B.Sc. degree in computers (information technology) from Cairo University, Egypt, and the M.Sc. degree in information technology management and the Ph.D. degree in computer science from Universiti Teknologi Malaysia (UTM), Malaysia. He is currently a Senior Lecturer with the Computing and Data Science Department, School of Computing and Digital Technology, Birmingham City University (BCU), U.K., where he is also leading the Smart Health Laboratory, Data Analytics and AI Research Group. Previously, he was an Assistant/Associate Professor with Taibah University, Saudi Arabia, from 2017 to 2021, and a Senior Lecturer with the Department of Information Systems, Faculty of Computing, UTM, from 2014 to 2017. He has published several papers in indexed journals and international conferences. His research interests include data mining, artificial intelligence, machine learning, information retrieval, and health informatics.

MOHAMMED AL-SAREM received the B.Sc. and M.Sc. degrees in information technology and computer engineering from the Faculty of Informatics and Computer Engineering, Volgograd State Technical University, Volgograd, Russia, in 2005 and 2007, respectively, and the Ph.D. degree from the Faculty of Informatics, University of Hassan II Casablanca, Mohammedia, Morocco, in 2014. He is currently an Associate Professor with the Department of Information Systems, Taibah University. He is also a researcher with more than eight years of experience teaching courses at the bachelor's and master's levels. He has received many grants/funding from local and international parties. He has published more than 40 articles in peer-reviewed reputed journals and about 29 conference papers. His research interests include artificial intelligence, big data analytics, natural language processing (NLP), social network analysis, educational data mining, and machine learning/deep learning. He is a Senior Fellow of the Higher Education Academy (SFHEA), U.K., and a Certified Google Data Scientist. In addition, he has managed and organized many international conferences.

SULTAN NOMAN QASEM (Senior Member, IEEE) received the B.Sc. degree in computer science from Mustansiriyah University, Iraq, in 2002, and the M.Sc. and Ph.D. degrees in computer science from Universiti Teknologi Malaysia (UTM), Malaysia, in 2008 and 2011, respectively. From 2012 to 2020, he was an Assistant Professor with Imam Mohammad Ibn Saud Islamic University, Saudi Arabia. From May 2011 to October 2012, he was a Senior Lecturer with the Faculty of Computing, UTM. He is currently an Associate Professor with the Department of Computer Science, College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University. He has directed many funded research projects. He has published in several reputed peer-reviewed journals and contributed book chapters and conference proceedings. His research interests include applied artificial intelligence, machine learning, data science, multi-objective evolutionary algorithms, metaheuristic optimization algorithms, and health informatics. He has served as a program committee member for various international conferences and a reviewer for various international journals. He is a technical editor of other journals.

TAWFIK AL-HADHRAMI received the M.Sc. degree in IT/applied system engineering from Heriot-Watt University, Edinburgh, U.K., and the Ph.D. degree in wireless mesh communication/IoT from the University of the West of Scotland, Glasgow, U.K., in 2015. He is currently a Senior Lecturer with Nottingham Trent University (NTU), U.K. He is also a member of the Cyber Security Research Group (CSRG), NTU, and the National Cyber Security Center (NCSC), which is an organization of the U.K. Government that provides advice and support for the public and private sectors on how to avoid computer security threats. He has been involved in research with the IoT and Networking Group, University of the West of Scotland, U.K., and also involved in different projects with industries, such as Ofgem, Catapult, and Sustainable Healthcare. He has established collaborations with different international institutions over the world. He has published more than 50 papers in peer-reviewed journals and conferences. His research interests include cybersecurity IC, the Internet of Things (IoT) applications, the IoT/artificial intelligence (AI) in healthcare and behavior applications, machine learning, smart cities, network infrastructures and emerging technologies, computational intelligence, and 5G/6G wireless communications. He is an Associate/Assistant Editor for several decent journals, such as IEEE Access, *International Journal of RF and Microwave Computer-Aided Engineering* (Hindawi), *Peer J Computer Science*, *Frontiers in Communications and Networks (IoT and Sensor Networks)*, and *International Journal of Cyber Forensics and Advanced Threat Investigations*.