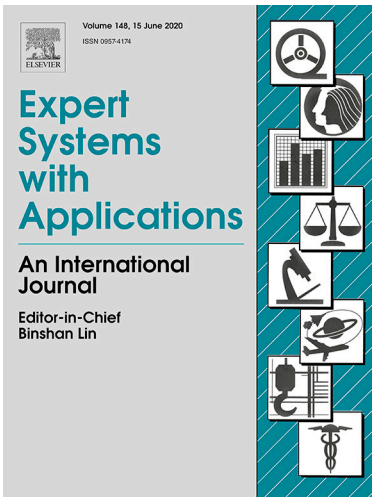
Journal Pre-proofs



Fraud detection in financial statements using data mining and GAN models

Seyyede Zahra Aftabi, Ali Ahmadi, Saeed Farzi

PII: S0957-4174(23)00646-2
DOI: https://doi.org/10.1016/j.eswa.2023.120144
Reference: ESWA 120144

To appear in: *Expert Systems with Applications*

Received Date: 11 December 2022
Revised Date: 18 March 2023
Accepted Date: 11 April 2023

Please cite this article as: Zahra Aftabi, S., Ahmadi, A., Farzi, S., Fraud detection in financial statements using data mining and GAN models, *Expert Systems with Applications* (2023), doi: https://doi.org/10.1016/j.eswa.2023.120144

Fraud detection in financial statements using data mining and GAN models

Seyyede Zahra Aftabi[1]

Faculty of Computer Engineering, K. N. Toosi University of Technology, Tehran, Iran

Email: szaftabi@email.kntu.ac.ir

ORCID: https://orcid.org/0000-0002-3651-9723

Ali Ahmadi

Faculty of Computer Engineering, K. N. Toosi University of Technology, Tehran, Iran

Email: ahmadi@kntu.ac.ir

ORCID: https://orcid.org/0000-0003-4211-6258

Saeed Farzi

Faculty of Computer Engineering, K. N. Toosi University of Technology, Tehran, Iran

Email: saeedfarzi@kntu.ac.ir

ORCID: https://orcid.org/0000-0002-5531-8134

Fraud detection in financial statements using data mining and GAN models

Seyyede Zahra Aftabi [a], Ali Ahmadi [b], Saeed Farzi [c]

[a] Faculty of Computer Engineering, K. N. Toosi University of Technology, Tehran, Iran, szaftabi@email.kntu.ac.ir

[1] **Corresponding author:**
Seyyede Zahra Aftabi, Faculty of Computer Engineering, K. N. Toosi University of Technology, Tehran, 16317-14191, Iran.

[b] Faculty of Computer Engineering, K. N. Toosi University of Technology, Tehran, Iran, ahmadi@kntu.ac.ir

[c] Faculty of Computer Engineering, K. N. Toosi University of Technology, Tehran, Iran, saeedfarzi@kntu.ac.ir

Abstract

Financial statements are analytical reports published periodically by financial institutions explaining their performance from different perspectives. As these reports are the fundamental source for decision-making by many stakeholders, creditors, investors, and even auditors, some institutions may manipulate them to mislead people and commit fraud. Fraud detection in financial statements aims to discover anomalies caused by these distortions and discriminate fraud-prone reports from non-fraudulent ones. Although binary classification is one of the most popular data mining approaches in this area, it requires a standard labeled dataset, which is often unavailable in the real world due to the rarity of fraudulent samples. This paper proposes a novel approach based on the generative adversarial networks (GAN) and ensemble models that is able to not only resolve the lack of non-fraudulent samples but also handle the high-dimensionality of feature space. A new dataset is also constructed by collecting the annual financial statements of ten Iranian banks and then extracting three types of features suggested in this study. Experimental results on this dataset demonstrate that the proposed method performs well in generating synthetic fraud-prone samples. Moreover, it attains comparative performance with supervised models and better performance than unsupervised ones in accurately distinguishing fraud-prone samples.

Keywords: Fraud in financial statements, anomaly detection, outlier generation, generative adversarial networks, ensemble models, banking sector

1. Introduction

Today, the incremental growth of fraud in business, especially in financial services, has become an earnest and costly problem. There exists no single definition for the concept of fraud in scientific sources. One of the clearest definitions available is the one provided by the Association of Certified Fraud Examiners (ACFE) in 2008. According to this definition, individuals and organizations may commit illegal actions such as deception or betrayal of trust for specific reasons, such as obtaining money, property, or individual or collective benefits, which are interpreted as fraud (Hashim et al., 2020; Sadgali et al., 2019; Syahria, 2019). The American Institute of Certified Public Accountants (AICPA) has also attributed the concept of fraud to any type of fraud, including minor employee theft, unproductive performance, embezzlement, misappropriation of assets, and fraudulent financial reporting (Hashim et al., 2020).

As it can be understood from the above definitions, there are variants of fraud, among which this study is focused on fraud in financial statements. Financial statements are reports that detail an organization's business activities and financial performance from various perspectives (Ashtiani & Raahemi, 2021; Jan, 2018). The most important contents of these reports include expenses, incomes, received or granted loans, profits, and losses (Ashtiani & Raahemi, 2021). These large amounts of numbers and figures provide an opportunity for profit seekers to cheat. Among the most common forms of fraud in financial statements are premature revenue recognition, spurious entries of incomes or profits, overstating

assets, understating expenses, and concealment or false disclosure of expenses (Craja et al., 2020; Gray & Debreceny, 2014). According to the ranking provided by ACFE, financial statement fraud is the third most prevalent type of occupational fraud, after corruption and embezzlement (Hashim et al., 2020; Petković et al., 2021; Syahria, 2019). However, it has taken first place regarding the financial costs and the amount of loss it incurs (Omidi et al., 2019). Hence, early detection of this type of fraud can prevent its exorbitant financial consequences.

Traditional approaches for fraud detection are often founded on a theory called the fraud triangle, which includes three aspects: incentive (or pressure), rationalization, and opportunity (Ashtiani & Raahemi, 2021; Ravisankar et al., 2011; Syahria, 2019). Based on this theory, auditors try to identify fraudulent behaviors by inspecting financial institutions, evaluating the incentives of people to commit fraud, and analyzing fraud patterns (Craja et al., 2020). Nevertheless, the report published by ACFE in 2022 stated that the share of internal and external auditors in identifying suspected cases of fraud was only 16 and 4 percent, respectively (ACFE, 2022). The roots of this inefficiency can be found in the rapid development of modern technologies. To be more precise, in addition to being very expensive and time-consuming, traditional audit approaches are inaccurate and impractical due to reasons such as the frequent change in fraud methods by fraudsters, the rare fraud-prone samples, and the lack of data mining knowledge. Hence, they are defeated by new fraud methods (Huang et al., 2017; Omidi et al., 2019; Petković et al., 2021; West & Bhattacharya, 2016).

In recent years, creating and developing intelligent systems capable of detecting fraud in financial statements has become one of the popular research areas. These systems can facilitate decision-making for auditors through early warning notification (Craja et al., 2020). The literature review indicates that most researchers in this field are inclined to use various data mining methods. This tendency can be attributed to the efficiency of these methods, both in detecting fraud in financial statements and in detecting other types of financial crimes, such as check fraud, loan fraud, and credit card fraud (Gupta & Mehta, 2021). Meanwhile, the role of machine learning algorithms as the most important data mining tool for discovering and extracting relationships and truths hidden in big data is undeniable. However, these approaches may encounter problems such as the high dimensionality of the data, the lack of fraud-prone samples, or the data imbalance that should be addressed carefully (Al-Hashedi & Magalingam, 2021; J. I. Z. Chen & Lai, 2021; Fiore et al., 2019; Jeragh & AlSulaimi, 2018; Mohammadi et al., 2020; Sethia et al., 2018).

In this article, the occurrence of fraud in the annual financial statements of Iranian banks, which are the most vital components of the country's economy, has been studied. To this end, a novel approach based on outlier detection and ensemble models is proposed, which consists of three major components. The first component uses a specific type of generative adversarial networks (GANs) to generate synthetic outliers so that they can hardly be distinguished from the real samples. The second component produces a standard labeled dataset by combining generated and real samples as positive and negative classes, respectively. After employing a feature subset selection method in the third component, a binary classification model is trained to detect fraudulent financial statements. Thus, the main innovation of the proposed approach is its adaptability to challenges such as small data size, high-dimensional feature spaces, and the absence of samples susceptible to fraud.

Another achievement of this study is compiling a set of audited financial statements of Iranian banks to create a standard dataset. Since the collected data is completely raw, different preprocessing techniques are used to convert it into a useful dataset. In the current study all data items are contributed in the construction of feature vectors, as opposed to previous studies that only used

specialized financial ratios and ignored the original data items. In particular, new features, such as growth rate, horizontal ratio, and vertical ratio, are defined, which can be analyzed even by non-specialists.

In order to evaluate the performance of the proposed approach, four scenarios are conducted over the real-world dataset. The results of the first two scenarios highlight the efficiency of the proposed features as well as the effective performance of GAN in creating fake outliers. In addition, the outcomes of the last two scenarios indicate the generalizability of the ensemble model and its competitive performance with several well-known classification methods, such as SVM (Boser et al., 1992) and Logistic regression (Cramer, 2003).

The contributions of this paper are as follows:

- Modeling data of banks' financial statements and building a standard dataset containing the information of 49 real-world financial statements
- Overcoming the lack of fraud-prone samples by generating synthetic samples susceptible to fraud, utilizing generative adversarial networks
- Reducing the high dimensionality of the feature space and providing a rich representation of data by using the correlation matrix and employing unsupervised models
- Running plenty of experimental analysis in four segments, including loans, deposits, incomes, and costs, to evaluate the performance of the proposed approach and compare it with other competitors

The rest of the paper is organized as follows. In Section 2, recent related work is presented. In Section 3, the description of some prerequisites and foundations is provided. Section 4 discusses the problem definition and the proposed approach in detail. Experimental studies are then reported in Section 5. Finally, conclusions and suggestions for future work are discussed in Section 6.

2. Background and related work

In recent years, many researchers have focused on designing decision-support systems to detect fraud in financial statements. These studies can be categorized from five different perspectives (Ashtiani & Raahemi, 2021; Craja et al., 2020): (1) technique type, (2) data distribution, (3) feature type, (4) learning model, and (5) evaluation criteria.

Technique type: Generally, data mining techniques for solving the problem of fraud detection in financial statements can be divided into two categories based on the utilized data: binary classification and anomaly detection[2]. In the first category, once there is a sufficient number of fraud-prone samples in the dataset, the statistical models or machine learning methods are applied to classify the samples into two classes, "fraud-prone" and "fraud-free" (Craja et al., 2020; Karlos et al., 2017; Patel et al., 2019; Ravisankar et al., 2011; Temponeras et al., 2019). The second category is when the dataset contains no fraud-prone samples, or the fraud-prone samples are scarce. In this situation, fraud is treated as an anomaly that should be sought to be identified (Lokanan et al., 2019; Noels et al., 2022). The word anomaly refers to samples that do not follow the general behavior of normal samples or have a significant deviation from others.

Data distribution: Most researchers that have used the classification approach have first identified a set of fraudulent and non-fraudulent companies in the stock market and then collected financial statements

[2] also referred to as outlier detection

of these companies in a specific period of time (Craja et al., 2020; Dutta et al., 2017; Hajek, 2019; Karlos et al., 2017; Patel et al., 2019; Ravisankar et al., 2011; Temponeras et al., 2019; Yao et al., 2018). One of the main challenges of this approach is the imbalance between fraud-prone and fraud-free samples. For example, in studies (Temponeras et al., 2019) and (Karlos et al., 2017), the ratio of fraudulent to non-fraudulent samples was 1 to 4 and 1 to 3, respectively. To address this challenge, some studies proposed models well-compatible with imbalanced data, and a few studies tried to balance the dataset by oversampling or generating artificial samples (Dutta et al., 2017; Fiore et al., 2019; Gangwar & Ravi, 2019; Sethia et al., 2018).

Feature type: Concerning the type of features used, the studies can be classified into five categories: (1) original data, (2) financial ratios, (3) financial and non-financial ratios, (4) linguistic features, and (5) financial ratios and linguistic features. Every financial statement consists of a large number of items that are mostly numerical and expressed in the currency of each country. As the direct use of these items may expose the model to the curse of dimensionality, many experiments are limited to using only a certain number of financial ratios. Financial ratios refer to statistical indicators that, based on a specific definition, calculate the relative magnitude of two numerical values taken from financial statements. They make it possible to quantitatively analyze financial statements and compare the performance of similar financial firms (Ashtiani & Raahemi, 2021; Karlos et al., 2017; Kingsley & Patrick, 2021; Omidi et al., 2019; Patel et al., 2019; Temponeras et al., 2019).

In addition to financial ratios, some works have also used non-financial features such as the institution size, the quality of management, and the board of directors or supervisors (Ravisankar et al., 2011; Yao et al., 2018). Besides, since the managers in some financial institutions have to prepare the analysis of financial statements, the decisions made, and the outlook for the coming years in the form of a text report, therefore, a small number of studies have benefited from the information in this part of the reports to detect fraud (Ashtiani & Raahemi, 2021; Hajek, 2019; Hajek & Henriques, 2017; Throckmorton et al., 2015).

In the following, an overview of the latest research on fraud detection in financial statements using data mining and machine learning models is provided.

Ravisankar et al. (Ravisankar et al., 2011) aimed to apply data mining techniques to a collection of financial statements of Chinese companies. Each financial statement consisted of 35 items, 28 of which were financial ratios reflecting the company's liquidity, profitability, efficiency, and security. After normalizing the features, the most effective ones were selected using the t-statistic. Then, as reported in Table 1, different machine learning models were utilized to identify fraudulent cases. According to the reported results, the Probabilistic Neural Network (PNN) has shown the best performance, and the Genetic Programming (GP) has brought the worst performance.

Karlos et al. (Karlos et al., 2017) investigated the active learning (AL) mechanism as a solution to deal with the lack of labeled data. They collected 164 financial statements described with 25 nominal predictors. Then, the ability of each predictor to resolve the ambiguity of similar samples was evaluated using the RelifF criterion. Eventually, the top 8 predictors were selected. Performance evaluation of various machine learning models specified the efficacy of active learning, especially for small datasets.

Dutta et al. (Dutta et al., 2017) assessed different machine learning approaches with the aim that the designed system could distinguish not only fraudulent cases from non-fraudulent ones but also intentional fraud from inadvertent mistakes. Each financial statement had 116 features, brought down

to 15 after a feature selection stage. The results pointed out the high efficiency of Artificial Neural Networks (ANN). Additionally, the Support Vector Machine (SVM) has revealed the weakest performance.

Yao et al. (Yao et al., 2018) investigated the usefulness of combining feature selection with machine learning algorithms. In the first step, various models performance was compared in the presence of all features, and SVM achieved the best accuracy. In the second step, the features were ranked using Principal Component Analysis (PCA) (Pearson, 1901) and Extreme Gradient Boosting (XGBoost) algorithms and then presented to models in descending order of importance. Results demonstrated that using 2 to 5 features could be apt for fraud detection. Besides, Random Forest (RF) has outperformed the others.

Hajek et al. (Hajek & Henriques, 2017) used not only financial data but also textual data from managerial reports. Their idea was based on the fact that the most significant frauds are committed by senior managers. Accordingly, as the managers' comments are reflected in textual reports, analyzing reports would make it possible to predict their potential of committing fraud. They extracted linguistic features, such as the occurrence rate of uncertain, constraining, and litigation words, in addition to financial attributes. Afterward, dimensionality reduction was performed utilizing the correlation-based filter and the best-first search method. Finally, 14 different machine learning techniques were surveyed to establish fraud early warnings. The results revealed that ensemble models and Bayesian Belief Networks (BBN) had performed best in correctly detecting fraud-prone and fraud-free statements, respectively.

In another study (Hajek, 2019), Hajek focused on the interpretability of detected frauds. In this way, after removing irrelevant features using a steady-state genetic algorithm, a fuzzy rule-based detection system was optimized by various evolutionary and non-evolutionary algorithms. Comparative analysis with the state-of-the-art fuzzy rule-based systems indicated the competitive accuracy of his proposed model. However, the interpretability of predictions for auditors was cited as its main strength.

Temponeras et al. (Temponeras et al., 2019) proposed a deep dense neural network for predicting fraud. The model was tested using 164 financial statements, each described by 23 numerical attributes. Experimental results showed its superiority in comparison to other machine learning models reported in Table 1.

Patel et al. (Patel et al., 2019) compared the performance of 42 different machine learning models. The features extracted from financial statements were 31 financial ratios, which were narrowed down to the ten most significant ones using a t-test. Based on the reported results, RF has surpassed other competitors in terms of accuracy.

Lokanan et al. in (Lokanan et al., 2019) proposed an anomaly detection approach to identify fraud-prone financial reports. They accumulated the annual and quarterly financial statements of Vietnamese companies. Each financial statement held 31 financial ratios, of which seven ratios were removed due to high correlation with others. After data normalization using the multivariate normal distribution, Mahalanobis distance was used to compute each sample's proximity to the centroid of the distribution. According to the results, most Vietnamese firms were trusted, and only about a quarter was suspected of fraud.

Inspired by (Hajek & Henriques, 2017), Craja et al. (Craja et al., 2020) focused on fusing the text embedding vectors produced by deep learning models into the fraud detection process. To achieve this goal, nine textual features and 47 numerical features were extracted from the content of each financial

statement. They assessed the proficiency of several machine learning and deep learning models alongside their proposed model called Hierarchical Attention Network (HAN) in the presence and absence of textual features. The presented results pointed out the good potential of textual information in improving accuracy. Furthermore, deep learning models have performed better than traditional machine learning models in detecting fraud-prone cases.

Noels et al. (Noels et al., 2022) used a graph distance metric named Earth Mover's Distance (EMD) to estimate the similarity between the financial statements of different companies. More specifically, they used a weighted tree to represent the hierarchy and relationships between data items. Then, the total cost of shifting weights over the edges of one tree to become identical to another was measured as the distance. The results showed the effectiveness of their proposed approach. An overview of related work is provided in Table 1.

Table 1. An overview of related work in fraud detection in financial statements from different aspects

Ref		Dataset			Feature type					Utilized models										Evaluation metric						
		Data source	Number of fraudulent	Number of non-fraudulent financial statements	Financial data items																					
(Ravisankar et al., 2011)	C	Chinese SE	101	101	28	7				✓	✓	✓		✓		✓		✓			✓	✓	✓	✓			
(Karlos et al., 2017)	C	Greek companies on the Athens SE	41	123	25					✓		✓	✓	✓		✓		✓	✓		✓						
(Dutta et al., 2017)	C	Audit Analytics database, COMPUSTAT database	3513	60720	116					✓		✓	✓	✓	✓			✓	✓		✓	✓	✓	✓	✓		
(Yao et al., 2018)	C	China securities regulatory commission	120	120	17	5				✓	✓	✓		✓	✓			✓			✓						
(Hajek & Henriques, 2017)	C	New York SE	311	311	32				✓	✓	✓	✓	✓	✓	✓	✓			✓	✓		✓	✓	✓	✓		✓	✓

(Hajek, 2019)	R	New York SE	311	311		32	✓											✓		✓						✓
(Temponeras et al., 2019)	C	Greek companies on the Athens SE	41	123		25		✓	✓	✓			✓			✓		✓		✓	✓	✓			✓	
(Patel et al., 2019)	C	Bombay SE	172	184		31		✓	✓	✓	✓	✓	✓	✓			✓	✓		✓						
(Lokanan et al., 2019)	A	Vietnam SE	22488 (unlabeled)			31												✓								✓
(Craja et al., 2020)	C	United States SE	208	7341		47	✓	✓	✓	✓				✓			✓		✓	✓	✓	✓	✓		✓	
(Noels et al., 2022)	A	Silverfin[3]	100 (unlabeled)		✓											✓										✓

C: Classification, A: Anomaly Detection, R: Rule-based, SE is the abbreviation of Stock Exchange

In addition to the above studies, research in other branches of fraud detection can also provide ideas for developing decision support systems for financial statement fraud detection. For example, by treating banks' financial statements as streaming time series, one can use the method proposed by Lee et al. (Lee et al., 2020) to determine whether an anomaly is likely to happen in the near future. They proposed a proactive and real-time LSTM-based model whose primary advantage was requiring minimal human intervention and offline learning time. Indeed, after a short initial training, it was able to start detecting anomalies while dynamically adjusting its detection threshold and re-training its LSTM model to accommodate changing patterns.

Saia and Carta (Saia & Carta, 2019) also proposed a proactive approach for detecting anomalies in transaction data. The idea was to transform the sequence of values of each transaction feature into new domains using Fourier and Wavelet transformations. Apart from reducing data heterogeneity and computational complexity, it allowed the authors to address the cold-start issue by including only legitimate transactions in the model definition. Yet, it met with some limitations when the data distribution was balanced or there were fraudulent samples available.

Bagga et al. (Bagga et al., 2020) evaluated the performance of nine various techniques, including but not limited to KNN, SVM, RF, pipelining and ensemble learning to identify potential fraud cases in credit card transactions. The pipelining technique was found to be the most suitable.

Carcillo et al. (Carcillo et al., 2021) combined unsupervised and supervised learning techniques for credit card fraud detection. First, they defined several unsupervised outlier scores in different levels of

[3] A Belgian accountancy cloud service

granularity, ranging from the card level to the global level. Then, they assessed the added value of these scores once integrated as features in a RF model. Overall, besides learning from past fraudulent behaviors, this approach offers the advantage of detecting new types of fraud in light of outlier scores.

3. Basic concepts and prerequisites

In this section, first, the definition of Generative Adversarial Networks (GANs) is given and a special type of them called MO-GAAL is presented (Section 3.1). Then, in Section 3.2, the ensemble models are introduced.

3.1 Generative Adversarial network

Generative Adversarial Networks (GANs) are relatively new deep learning models built based on a competitive game between two neural networks, named generator and discriminator. Generator $G(z;\theta_G)$ consistently engages in generating fake samples that resemble real data. In contrast, taking an input sample, discriminator $D(s;\theta_D)$ tries to correctly estimate the probability that it belongs to each of the two classes, real or fake. The training continues until reaching a Nash equilibrium where the discriminator is deceived more than half the time. In this situation, it can be claimed that the generator has succeeded in producing believable fake samples (Paper, 2021; Shahriar, 2022; Sim et al., 2021; C. Zhao et al., 2022). Thus, the optimization process of a GAN network can be formulated using Eq. 1. (Goodfellow et al., 2020).

$$\min_{\theta_g} \max_{\theta_D} V(D;G) = E_{s\sim p_{data}}[log\ D(s)] + E_{z\sim p_z}[log(1 - D(G(z)))] \quad (1)$$

where θ_G and θ_D are the parameters of model G and D, respectively. Furthermore, s denotes a real sample from data generating distribution p_{data} and z represents a noise variable sampled from a certain distribution p_z.

Traditional GANs may encounter two issues during the training process, namely mode collapse and gradient vanishing. To alleviate these problems, various solutions are presented in the literature, including modification of objective functions, using multiple generators or discriminators, designing different architectures, dynamically adjusting the training strategy using evolutionary techniques, and proposing phased evolutionary GANs (Xue et al., 2022). Among these, Liu et al. (Y. Liu et al., 2019) introduced Multi-Objective Generative Adversarial Active Learning (MO-GAAL), consisting of one discriminator model and more than one generator model. In this network, the data $\mathbb{S}$ is first equally partitioned into k subsets (i.e., $\bigcup_{i=1}^{k}\mathbb{s}_i = \mathbb{S}$) according to the outputs of the discriminator $D(s)$. Each subset contains samples most similar to each other in the sample space. Then, for each subset $\mathbb{s}_i$, the sub-generator g_i learns to generate potential fake samples having outputs close to $\mathbb{s}_i$. Therefore, the discriminator output should also be changed from 1 to T_i, where T_i represents a statistical representation of the samples in subset $\mathbb{s}_i$. Accordingly, the objective function reformulates, as shown by Eq. 2.

$$\max_{\theta_D} V(D) = \frac{1}{2|\mathbb{S}|}\sum_{j=1}^{|\mathbb{S}|} log\ D(s^j) + \frac{1}{2|\mathbb{S}|}\sum_{i=1}^{k}\sum_{j=1}^{n_i} log(1 - D(g_i(z_i^j))) \quad (2)$$

$$\min_{\theta_{g_i}} V(g_i) = \frac{-1}{|\mathbb{S}|}\sum_{j=1}^{|\mathbb{S}|}\left[T_i log\ D(g_i(z_i^j)) + \frac{-1}{|\mathbb{S}|}\sum_{i=1}^{k}(1-T_i) log(1 - D(g_i(z_i^j)))\right]$$

In the simple case, each generator g_i produces $n_i = |\mathbb{S}|/k$ number of fake samples. Thus, all generators have an equal share. Once the training stops, the fake samples from all the generators are combined.

The success of GANs in different image processing applications, such as generating realistic-looking images, has gradually drawn researchers' attention to using these networks to mitigate data imbalance in anomaly detection problems. For example, Oh et al. (Oh et al., 2019) proposed an oversampling model named OD-GAN in which the generator component sought to generate synthetic data through learning the minority class distribution. Sethia et al. (Sethia et al., 2018) also applied various GANs to produce diverse samples prone to credit card fraud and compared their effectiveness in downstream classification tasks. Other studies, including (El Kafhali & Tayebi, 2022), (Strelcenia & Prakoonwit, 2022), (Gangwar & Ravi, 2019), and (Fiore et al., 2019), also used GANs to over-sample fraudulent data in the problem of credit card fraud detection. Moreover, they compared the effectiveness of GAN with other popular oversampling models and reported the superiority of GAN.

3.2 Ensemble models

In the field of machine learning, there are two main learning strategies: supervised learning and unsupervised learning (Ashtiani & Raahemi, 2021; Omidi et al., 2019; Sadgali et al., 2019). Supervised learning aims to model the relationship between the input features and the target labels using a well-labeled dataset. So far, many studies in financial fraud detection, have used supervised approaches (Al-Hashedi & Magalingam, 2021; Ashtiani & Raahemi, 2021). In contrast, unsupervised learning comes into play when no prior knowledge is available regarding the target labels. This strategy makes it possible to train more complex models by discovering patterns, similarities, and relationships in large data (Omidi et al., 2019; Y. Zhao & Hryniewicki, 2018).

Several studies provided ensemble models to improve generalization and reduce bias and variance (Li et al., 2020, 2022; Tin Kam Ho, 1995; Y. Zhao & Hryniewicki, 2018; Y. Zhao et al., 2018). Their primary idea is to form a more solid learning model by combining numerous weak learners. XGBOD, presented in 2018, is one of the semi-supervised ensemble models proposed to better learn the complex patterns in problems having imbalanced data, such as outlier detection (Y. Zhao & Hryniewicki, 2018). It classifies samples through three phases. In the first phase, multiple unsupervised models are applied separately to the original data. Each model estimates the outlier score per sample. Accordingly, for every sample s_i, a feature vector is determined as formulated by Eq. 3.

$$\mathrm{F}(s_i) = \{f_u(s_i)\}_{u=1}^{U} = \{f_1(s_i), f_2(s_i), \ldots, f_U(s_i)\} \quad (3)$$

where U is the number of models, and f_u denotes the scoring function of the uth model. In the second phase, as demonstrated by Eq. 4, the score vectors obtained from the first phase are augmented to the original d-dimensional feature vectors to form new representations.

$$s_i^{new} = [s_i, \mathrm{F}(s_i)] \in [0,1]^{1 \times (d+U)} \quad (4)$$

Note that it is possible to pick only p scores from $\mathrm{F}(.)$ for combining with the original feature vector. In this case, the dimension of the newly generated vectors would be $(d + p)$. Finally, the s_i^{new} vectors are submitted to an XGBoost algorithm (T. Chen & Guestrin, 2016) to predict the labels in the third phase. A key point that remains is that in selecting unsupervised models, it is necessary to observe a trade-off between diversity and accuracy. In other words, the more diverse the base detectors, the better the ability to learn distinct data characteristics. Meanwhile, all models should be accurate as well (Y. Zhao & Hryniewicki, 2018).

4. Proposed Method

In this section, first, the problem of fraud detection in bank financial statements is formulated in Section 4.1. Next, in Section 4.2, the proposed approach and all its components are described in detail.

4.1 Problem formulation

Let $\mathbb{S}$ be a set of audited financial statements in which every statement $s \in \mathbb{S}$ reports the financial performance of one bank, containing its assets, liabilities, profits, and losses in a specific period. Thus, according to Eq. 5, each financial statement s can be considered a vector of numerical features.

$$s = \{x_1, x_2, \cdots, x_d, \ell\} \quad (5)$$

where d is the number of features and ℓ represents the label so that if s is fraudulent, the label is 1; otherwise, the label is 0. Provided that the standard data $\mathbb{S}$ is available, a model like $\Phi(\cdot)$ can be trained in such a manner that by taking a feature vector corresponding to one financial statement, it predicts the label, as formulated by Eq. 6.

$$\ell' = \Phi(s) \in \{0,1\}, \forall s \in \mathbb{S} \quad (6)$$

The model performs optimally if it correctly marks non-fraudulent samples with 0 labels and, in the meantime, has a high recall in detecting potentially fraudulent samples. In this way, the loss function can be defined based on Eq. 7.

$$\mathcal{L}(\Phi(\mathbb{S})) = \frac{1}{|\mathbb{S}|} \sum_{s \in \mathbb{S}} |\ell_s - \ell'_s| \quad (7)$$

To sum up, the objective is to train a classifier by minimizing the loss function.

4.2 Approach

As indicated in Figure 1, the proposed approach in this study consists of three main components: (1) outlier generation, (2) data augmentation, and (3) fraud detection. The first component generates a set of outlier samples by getting the original data and a noise variable. In the second component, all data, including real samples and generated outliers, are combined and labeled to produce a uniform dataset. Finally, feeding this dataset to the third component, an ensemble model is trained to classify samples into one of the two classes, fraud-prone and fraud-free. Indeed, the model is a decision support system that can later pinpoint even unseen samples prone to fraud. In the following, these three components are explained in detail.

Outlier generation. As discussed in Section 4.1, an audit decision support system must be able, upon receiving a new financial statement, to classify it into one of two categories, fraud-prone or fraud-free. Training such a model requires a labeled dataset containing proper distribution of both classes. However, there is no fraud-prone sample in the present study considering that the utilized data is a collection of financial statements of Iranian banks that have passed the audit process and were found to be free of fraud. To tackle the data deficiency problem, a component titled outlier generation is included in the proposed method, which generates outlier samples using the MO-GAAL network. Among the most important reasons for preferring MO-GAAL to other solutions are the high-dimensionality of data, the unknown distribution of samples, and the very similar behavior of bank fraudsters and auditors, respectively, to generator and discriminator components in GANs.

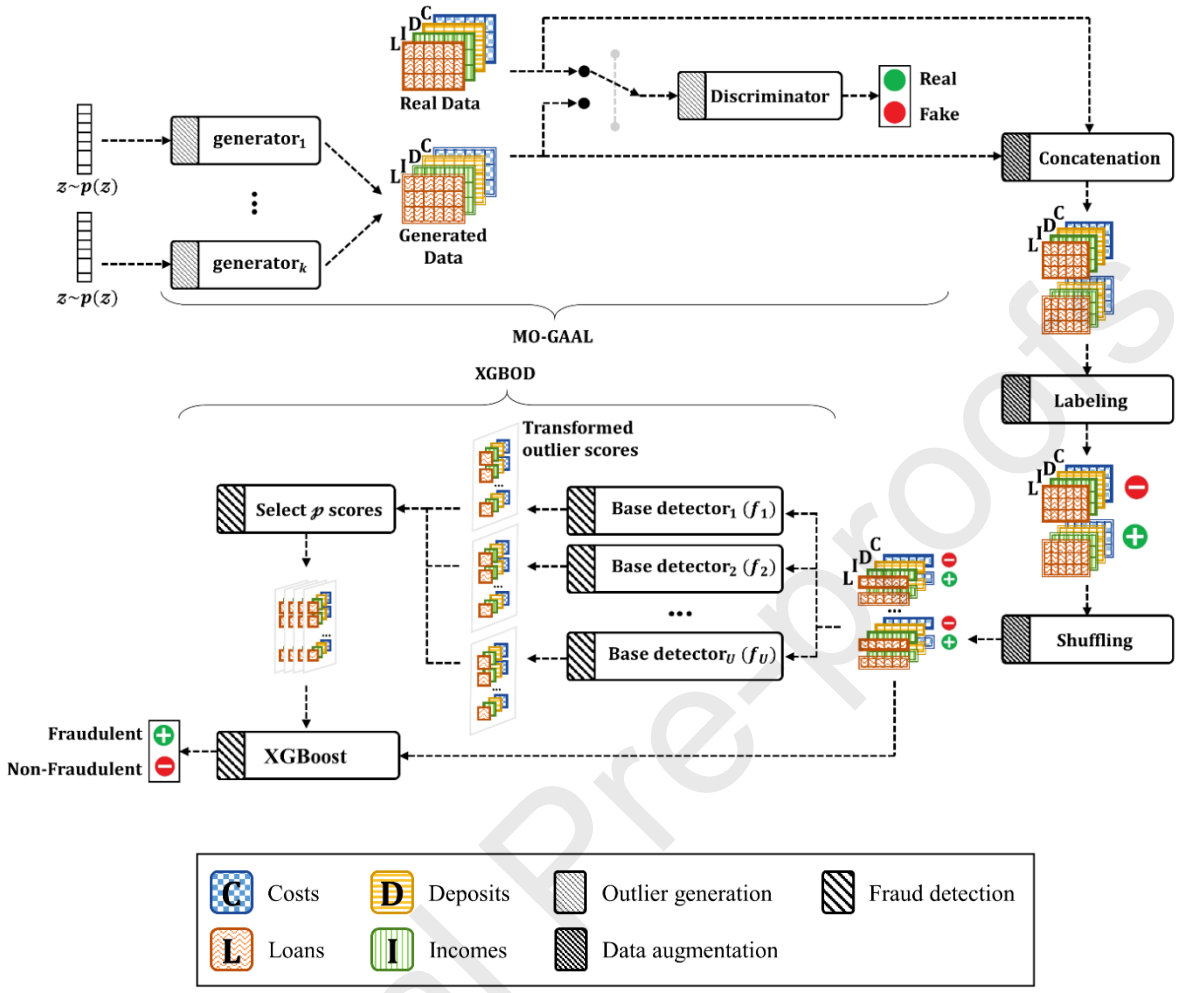


Figure 1. The proposed approach for fraud detection in banks' financial statements

As shown in Figure 1, first, all feature vectors and a noise variable z sampled from a uniform distribution $p(z)$ are taken as input. Then, the generators cooperate to produce diverse outliers scattered in different parts of the feature space and hardly distinguishable from the real ones. Next, the artificial outliers and the real vectors are fed to the discriminator to be trained to detect outliers. The generators and the discriminator are successively trained until convergence is met. At this time, the generators have fooled the discriminator more than half the time by producing high-quality synthetic samples.

Data augmentation. The data augmentation component is responsible for creating a standard labeled dataset, accomplished in three steps. In the first step, the outlier samples produced by the previous component are added to the end of the real data. It is worth noting that the number of artificial outlier samples is considered equal to the number of real samples for data equilibrium.

In the second step, all samples are labeled following one of the two possible approaches: (1) using machine learning algorithms to estimate a pseudo-target label for samples, or (2) assigning fraud-prone labels for synthetic samples and fraud-free labels for samples in the primary data. In this study, the

second approach is pursued because the audit organization has already reviewed all collected financial statements. In the last step, all samples are shuffled to ensure the appropriate distribution of samples in training and test data splits, leading to more generalization and avoidance of overfitting.

Fraud detection. In the fraud detection component, given the constructed labeled dataset, a machine learning model learns to predict the label of samples, whether potentially fraudulent or fraud-free. As noted in Section 3.2, this paper uses a model called XGBOD, an ensemble of supervised and unsupervised models, for feature engineering and classification. It is of note that selecting a proper set of base estimators highly affects the accuracy of predictions.

5. Experimental study

The key research questions (RQ) of this paper are as follows:

RQ1: How can data elements extracted from real-world financial statements be converted into feature vectors without relying on financial ratios? Which features are more efficient?

RQ2: To what extent are the generative adversarial networks suitable for generating synthetic outliers?

RQ3: How is the performance of XGBOD in discovering fraud-prone samples in real-world data?

RQ4: How is the proficiency of XGBOD in discriminating fraudulent samples from non-fraudulent ones compared to other known models?

Following this section, the research data, setting parameters, and evaluation metrics are explained. Thereafter, research questions are addressed by carrying out plenty of experiments in four main scenarios: (1) extracting feature vectors, (2) examining the performance of generative adversarial networks in simulating outliers, (3) investigating the performance of XGBOD in fraud detection, and (4) comparing the performance of XGBOD with several well-known models.

5.1 Data

The proposed approach is evaluated on a dataset consisting of 49 annual financial statements collected in 2014-2019 from ten different banks in Iran. This data is prepared in three stages. First, a list of the most valuable data items and the relationships between them is drawn up in four segments, namely loans, deposits, incomes, and costs. Second, considering the hierarchy of data items, a data model is designed using the SQL server. It is contrived so that all data items related to a specific financial statement can be retrieved from the database by querying only the bank name and the fiscal year. Third, the data is imported into the database. To this end, all financial statements are first converted to a uniform excel template formerly designed, and then, the data items related to the four mentioned segments are automatically extracted from the excel files and inserted into their corresponding fields in the database.

5.2 Configuration and setting parameters

All algorithms are implemented in python on the Google Colab platform with 12.68 GB RAM and 1 Tesla T4. Table 2 reports the value of all the parameters introduced in this study, along with a brief description of each. In addition, the SVM kernel is RBF.

Table 2. The values of setting parameters

Model	Parameter	Description	Value
MO-GAAL	n	total number of samples	39[l], 49[dci]
	lr_d	learning rate of discriminator	0.01
	lr_g	learning rate of generators	1e-4[ldc], 2e-4[i]
	k	number of sub-generators	3[l], 7[dci]
	ep	number of epochs to train	120
XGBOD	lr	learning rate of all detectors	0.2
	U	number of base detectors*	100

l: loans, d: deposits, c: costs, i: incomes, *: including KNN, LOF, HBOS, OCSVM, and IForest

5.3 Evaluation metrics

In this research, three well-known metrics in the field of data mining, namely accuracy, precision, and recall, are used. The F1-score is also reported to address the compromise that must be struck between recall and precision (Bagga et al., 2020; De Rossi et al., 2020). Note that the more accuracy and F1-score are close to 1, the better the model performance.

5.4 Extracting feature vectors

This section addresses **RQ1** (*How can data elements extracted from real-world financial statements be converted into feature vectors without relying on financial ratios? Which features are more efficient?*).

It is evident that all data items reported in financial statements provide auditors with valuable information to realize the financial status of institutions. However, these values are effective for fraud detection only when scrutinized in conjunction with each other. In other words, what discloses fraud is the inconsistency of the ratio between data items or sudden changes in the pattern of growth or decline in a data item over a short term. Therefore, instead of using the data items themselves, most previous research benefited from a set of financial ratios that have specific definitions and interpretations in

auditing (Ashtiani & Raahemi, 2021; Karlos et al., 2017; Kingsley & Patrick, 2021; Omidi et al., 2019; Patel et al., 2019; Temponeras et al., 2019).

Although financial ratios are developed by audit experts, we believe that training a fraud detection system exclusively with financial ratios is like being limited to rule-based approaches in artificial intelligence. Indeed, such systems suffer from non-scalability and ignorance of latent relationships between data items in spite of their interpretability, speed, and high accuracy. Apart from this, over time and with changing fraud tricks, the financial ratios should be updated, and new ratios should be defined. In this paper, a hybrid approach for feature extraction is proposed, in which all data items contribute to the construction of the feature vector.

Exploiting the original data values in the form of instructive features made us define three novel features, hereafter called vertical ratio, horizontal ratio, and growth rate. The vertical ratio, computed based on Eq. 8, is the ratio of the net amount of a data item x_i to the total net amount of items within the same column.

$$Vertical\_ratio(x_i) = \frac{x_i}{\sum_i x_i} \quad (8)$$

In tables presenting a specific categorization in the form of multiple columns, usually, there is a row whose name starts with the words "Net" or "Total." In these tables, as formulated by Eq. 9, the horizontal ratio of each category c_j can be calculated by dividing the total of that category by the total sum of all categories, which both are reported in the aforementioned row.

$$Horizontal\_ratio(c_i) = \frac{\sum_{x \in c_j} x}{\sum_{c_j} \sum_{x \in c_j} x} = \frac{c_j}{\sum_j c_j} \quad (9)$$

where x represents one of the data items in the category c_j.

The third feature is the growth rate, defined based on Eq. 10, in which the value of a data item in a specific fiscal year ${x_i}^t$ is first subtracted from its corresponding value in the previous fiscal year denoted by ${x_i}^{t-1}$. Next, the obtained result is divided by the value of the penultimate period. Indeed, this feature measures an item's increment or decrement rate after one fiscal cycle.

$$Growth\_rate(x_i) = \frac{-\left({x_i}^{t-1} - {x_i}^t\right)}{{x_i}^{t-1} + \varepsilon} \qquad (\varepsilon > 0) \quad (10)$$

where the additional term, ε, is added to prevent the denominator from being zero.

In this study, as depicted in Figure 2, three packages of SQL queries are defined to extract these features from four data segments, including loans, incomes, deposits, and costs. Each query retrieves a feature

vector from the database given a bank name and a fiscal year. Then, the resulting vectors are concatenated per segment to construct a comprehensive feature vector. Repeating this process for various banks and fiscal years and appending the results together, our primary dataset with the dimensions of $n \times d$ is formed, where n represents the total number of financial statements and d denotes the number of features.

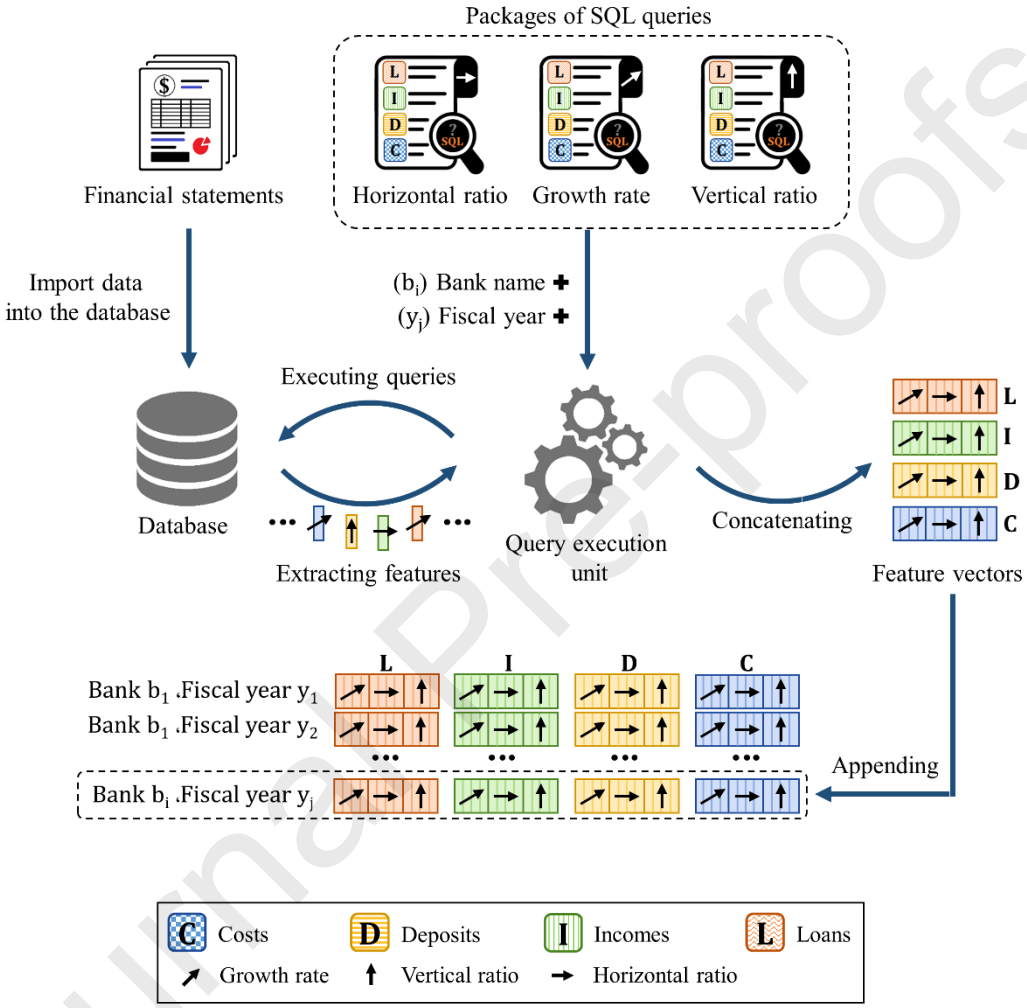


Figure 2. The overall process of extracting features and constructing the primary data

Yet, this primary dataset needs to be revised and standardized for four reasons. First, regarding the numerator of Eq. 10, when the value of an item in the previous fiscal year is zero, the growth rate becomes very high. Second, the data is very sparse because of banks' non-disclosure of some financial details or the monopolization of certain data items to one or more specific banks. Third, unlike the vertical and horizontal ratios, which always lie in a range of $[-1, 1]$, the growth rate has no definite range. Consequently, comparing its values is practically useless. Fourth, the high-dimensionality of feature vectors not only increases the required processing time and computational resources but may also cause noisy information to be fed into the decision-making model.

To address the challenges mentioned above, the following four actions are taken: (1) Adding conditional statements to SQL queries related to growth rate so that, if the value of a data item in the last fiscal year is zero, the growth rate should also be zero, (2) Replacing all empty features with zero values, followed by eliminating all-zero columns, which caused feature space reduction about 54%, 27%, 6%, and 15%, respectively in loans, incomes, deposits, and costs segments, (3) Scaling all values in each column to the range $[-1, 1]$, using a min-max normalization technique, and (4) Reducing dimensions utilizing the correlation matrix. To explain the fourth action more, firstly, the pairwise Pearson's correlation coefficient formulated by Eq. 11 is computed between feature pairs. Secondly, the feature pairs that are reported in the same subject notes, having an identical type (e.g., are both a kind of collaterals), and their correlation coefficient is greater than a predefined threshold τ are identified and grouped. Lastly, in each group, the best feature having the lowest average correlation with all other features is left, and the remaining are removed.

$$\rho_{X,Y} = \frac{COV(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \xrightarrow{n\ sample} \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad (11)$$

Here, $COV(.)$ gives the covariance between two random variables, X and Y, and $E(.)$ represents the expected value. μ_X and μ_Y denote the mean of the variables X and Y, and σ_X and σ_Y are their standard deviation. In the current study, the threshold τ is set to 0.7 by try and error. To illustrate the decline in the number of features after taking each action, we resort to Figure 3.

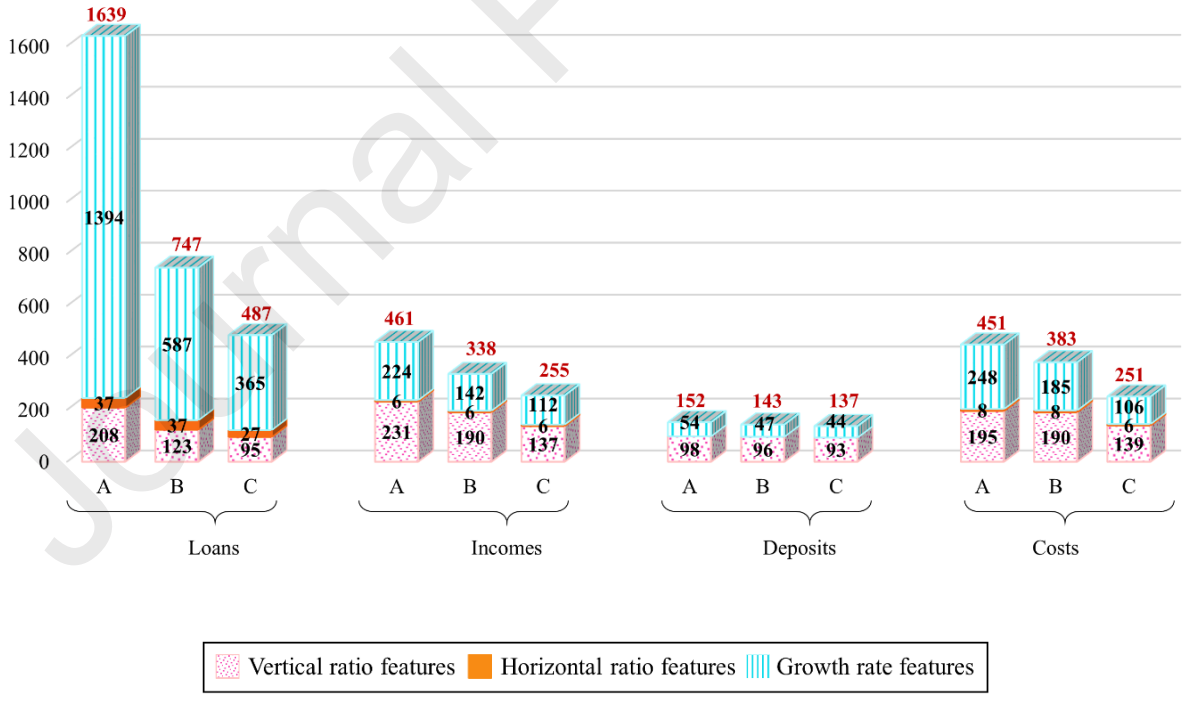


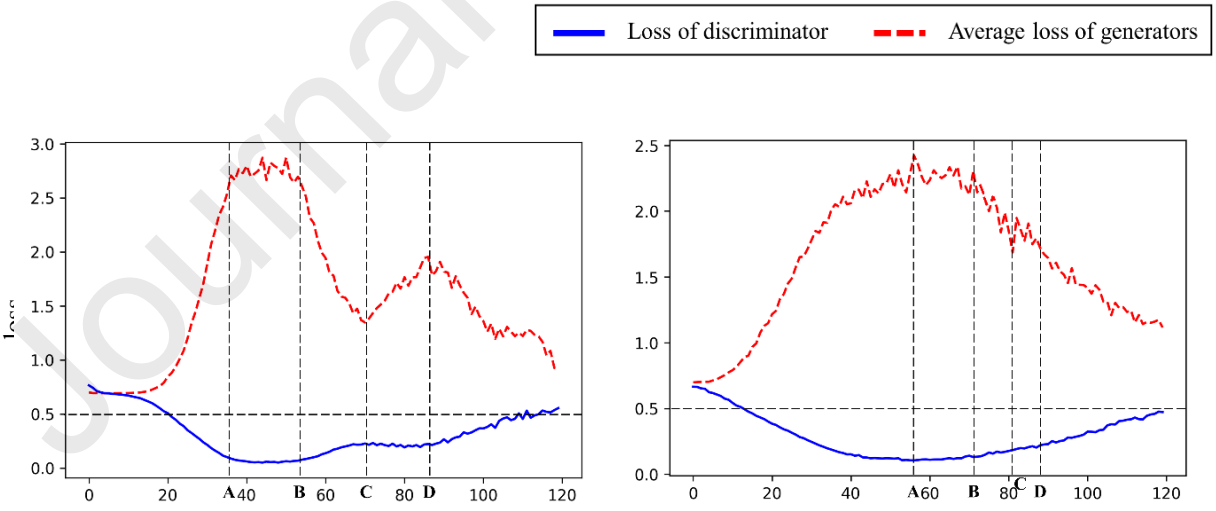
Figure 3. The number of various features in (A) the primary dataset, (B) the normalized dataset after taking the second action, and (C) the final dataset resulting from correlation-based dimension reduction in the fourth action.

5.5 Analysis of generative adversarial networks in outlier generation

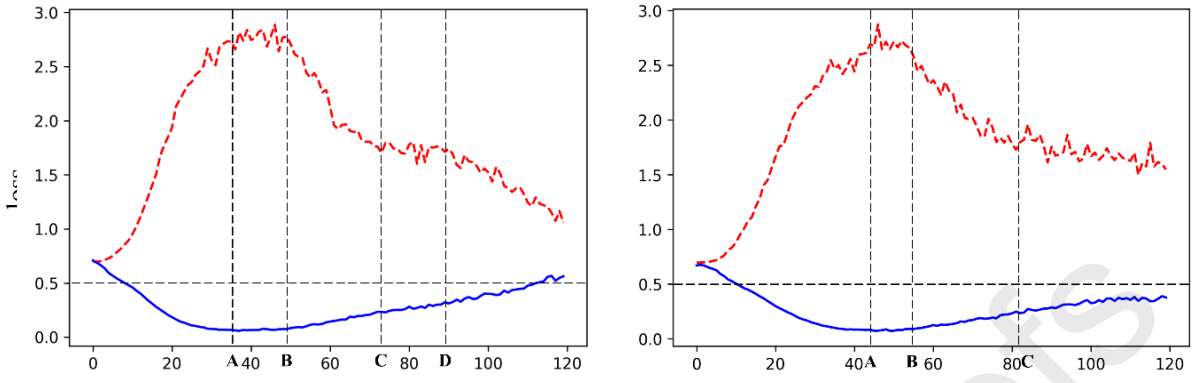
In response to **RQ2** (*To what extent are the generative adversarial networks suitable for generating synthetic outliers?*), we provide the final feature vectors obtained for loans, deposits, incomes, and costs, as input to the first component of the proposed approach, i.e., the outlier generation.

Figure 4 illustrates both the discriminator loss and the average loss of generators for each of the four segments. The vertical axis represents the loss values, and the horizontal axis shows the iteration. As subplots (a) to (d) of Figure 4 show, during early iterations in the range zero to A, the average loss of the generators is significantly increased. In particular, the performance of generators is unsatisfactory as they are not yet sophisticated enough to generate outlier samples around and near normal ones. This weakness causes the discriminator to easily unravel normal samples from outliers, and thus, the discrimination loss is reduced in this interval.

During iterations A to B, the generators gradually attain the required proficiency in producing outliers and deceive the discriminator. Hence, in the interval between iterations B and C, an increase in the discriminator loss and a significant decrease in the average loss of generators occur. A few iterations later (i.e., in the range of C to D), the discriminator tries to improve its performance by better defining the decision boundaries. This competition ends once the discriminator loss is raised above 0.5, owing to producing credible outliers by generators. Eventually, the final generated outliers related to the four segments are stored separately.



(a) loans (b) deposits

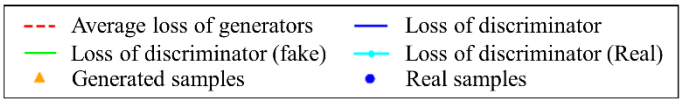


(c) incomes (d) costs

Figure 4. The losses of the generators and the discriminator in different iterations. Each plot displays the convergence of the generative adversarial model for one of the four segments.

To verify that a Nash equilibrium is obtained and the dataset is large enough to adequately train the GAN model, a detailed view of the discriminator loss and the average loss of generators for the loan data segment is demonstrated in Figure 5. The scatter plot of all generated and real samples is also presented in several arbitrary iterations. Note that the *d*-dimensional vector space is transformed to a 2-dimensional space using the t-SNE model (Van der Maaten & Hinton, 2008).

According to Figure 5, a Nash equilibrium is achieved once the discriminator loss reaches around 0.5 (i.e., the discriminator is deceived about half the time). In addition, the generated data are first aggregated around one or more randomly selected points and then gradually moved to the areas where the real data are located. The final generated outliers are expected to be placed around the real samples and be as similar as possible to them to be hardly distinguishable.



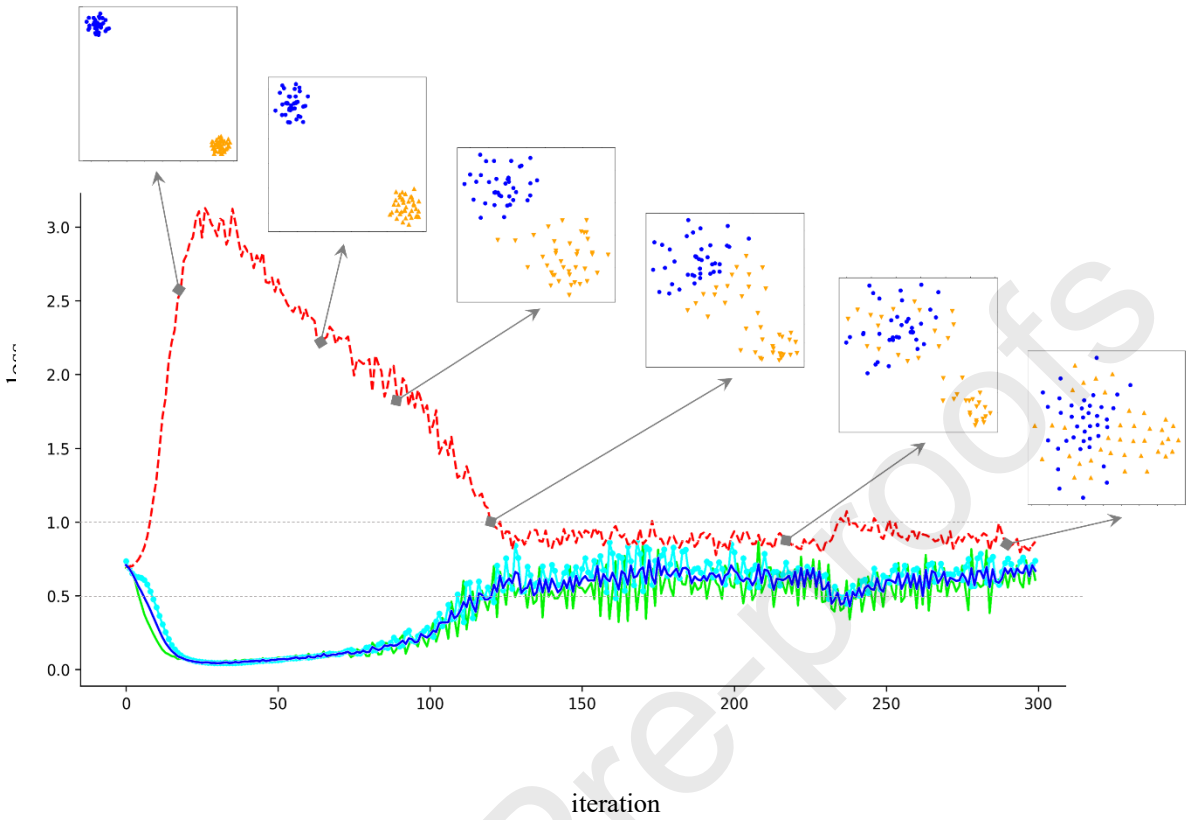


Figure 5. The data distribution and the convergence of the generative adversarial model to a Nash equilibrium during training on the loan data segment.

5.6 Analysis of XGBOD in fraud detection

Concerning **RQ3** (*How is the performance of XGBOD in discovering fraud-prone samples in real-world data?*), we take two steps. In the first step, for each segment of the data, the feature vectors of both the original samples and the generated outliers are combined through the data augmentation component to form a labeled dataset. The number of fraud-prone samples is considered the same as the number of fraud-free ones. In the second step, each data segment is separately used for training and testing the outlier detection component. Table 3 reports the evaluation results. It should be noted that all experiments are conducted using a 5-fold cross-validation technique, and the mean and variance of metrics are also reported.

The evidence in Table 3 implies that the model performs reasonably during training in all data segments and for all folds. It also denotes that XGBOD works well in testing in terms of mean recall, which conveys that the model is proficient in correctly identifying fraud-prone samples. However, some declines in the recall values are observed in some folds, which are mainly attributed to the data imbalance or the small data size. Moreover, the simultaneous high values of accuracy and recall indicate that the model successfully detects fraud-free samples as well. It is noteworthy that the recall is more valuable than the accuracy because detecting fraud-prone samples is more critical than fraud-free ones. Additionally, the

insignificant variance values confirm that the model is well trained and has a stable performance despite the small size of training data; ergo, it is promising to be generalizable to unseen samples.

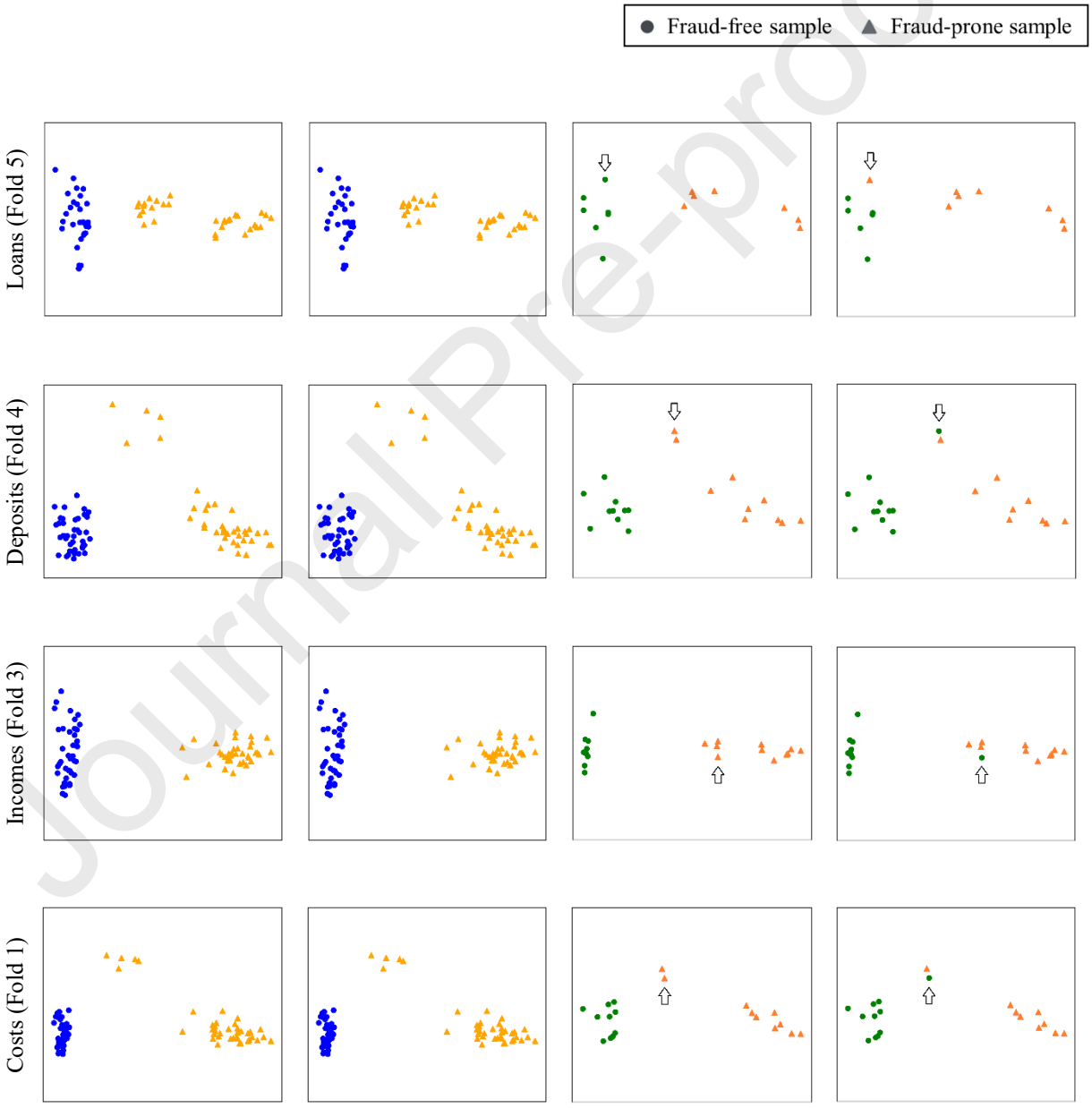
Table 3. Experimental results of XGBOD in each segment of the data, showing the precision, recall, and accuracy metrics in detail.

Segment	Fold	Train					Test				
		#Fraud-prone samples	#Fraud-free samples	Precision	Recall	Accuracy	#Fraud-prone samples	#Fraud-free samples	Precision	Recall	Accuracy
Loans	1	32	30	1	1	1	7	9	1	1	1
	2	33	29	1	1	1	6	10	1	1	1
	3	29	33	1	1	1	10	6	1	1	1
	4	30	32	1	1	1	9	7	1	0.89	0.94
	5	32	32	1	1	1	7	7	0.88	1	0.93
	Mean			1	1	1	Mean		0.976	0.978	0.974
	Variance			0	0	0	Variance		0.0023	0.0019	0.0010
Deposits	1	38	40	1	1	1	11	9	1	1	1
	2	37	41	1	1	1	12	8	0.92	1	0.95
	3	43	35	1	1	1	6	14	1	0.83	0.95
	4	39	39	1	1	1	10	10	1	0.90	0.95

	5	39	41	1	1	1	10	8	1	1	1
		Mean		1	1	1	Mean		0.984	0.946	0.970
		Variance		0	0	0	Variance		0.0010	0.0049	0.0006
Incomes	1	39	39	1	1	1	10	10	1	1	1
	2	40	38	1	1	1	9	11	1	1	1
	3	38	40	1	1	1	11	9	1	0.91	0.95
	4	41	37	1	1	1	8	12	1	1	1
	5	38	42	1	1	1	11	7	1	1	1
		Mean		1	1	1	Mean		1	0.982	0.990
		Variance		0	0	0	Variance		0	0.0013	0.0004
Costs	1	39	39	1	1	1	10	10	1	0.90	0.95
	2	44	34	1	1	1	5	15	1	1	1
	3	34	44	1	1	1	15	5	1	1	1
	4	39	39	1	1	1	10	10	1	1	1
	5	40	40	1	1	1	9	9	1	1	1
		Mean		1	1	1	Mean		1	0.980	0.990

	Variance	0	0	0	Variance	0	0.0016	0.0004

Figure 6 illustrates the distribution of labels, in which each row contains plots related to one of the quad segments. The plots in columns A-I and B-I represent the distribution of target labels, respectively, in train and test data. Likewise, columns A-II and B-II contain the distribution plots of predicted labels. As the test samples change in each iteration of cross-validation, only one of the five folds is displayed, whose number is written in front of the segment name. Non-fraudulent samples are marked with circles, and fraudulent samples with triangles. Note that as visualizing a *d*-dimensional vector is impossible, the feature space is reduced to two dimensions by PCA.



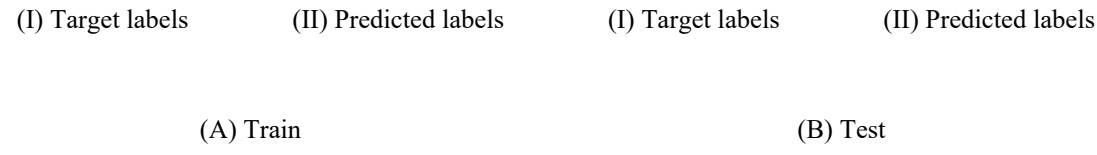


Figure 6. A two-dimensional visualization of the target labels and the predicted labels for one fold of each data segment. Symbol ⇨ shows the samples which are incorrectly labeled by the model.

Figure 6 clearly shows that the outlier generation component is perfectly capable of producing deceptive fraud-prone samples. Nonetheless, the fraud detection component also discriminates two classes well, promising to work even better by injecting larger datasets.

5.7 comparison of XGBOD with several well-known models

In this section, the **RQ4** (*How is the proficiency of XGBOD in discriminating fraudulent samples from non-fraudulent ones compared to other known models?*) is answered.

In addition to the semi-supervised XGBOD model, there are other well-known anomaly detection models, including KNN (Peterson, 2009), iForest (F. T. Liu et al., 2008), LSCP (Y. Zhao et al., 2018), ECOD (Li et al., 2022), and COPOD (Li et al., 2020) in the category of unsupervised models and SVM (Boser et al., 1992) and Logistic regression (Cramer, 2003) among the supervised models. Table 4 reports the performance of these models along with XGBOD, separately for each segment of the data. Remind that all models are assessed using a 5-fold cross-validation technique, and the mean and variances are reported. Moreover, a detailed analysis of the accuracy metric through 5-fold cross-validation is illustrated by boxplots in Figure 7.

Table 4 implies that SVM and Logistic Regression, which are both supervised models, perform best in classifying different segments of the research data. The former determines strict decision boundaries during training and then predicts the labels of newly arrived samples based on these boundaries. The latter is a statistical algorithm that calculates the probability of the occurrence of a specific event by discovering linear relationships between features (West & Bhattacharya, 2016). Figure 7 also confirms this finding by showing their superb accuracy in all four data segments. Nevertheless, for three reasons, none of these two models seem to be smart choices for the problem of fraud detection in financial statements in the real world. First, if the data is imbalanced, as in the problem under study (Moepya et al., 2014), the SVM may bias toward the majority class, which in our case is the class of fraud-free samples. Second, the learning complexity of SVM strongly depends on the data size and is therefore not suitable for large-size datasets such as the archives of financial statements, whose numbers increase yearly (Nguyen et al., 2020; Song et al., 2014). Third, fraudulent financial statements may not be linearly separable from non-fraudulent ones.

From the viewpoint of anomaly detection, the model capability to infer patterns, irregularities, and dependencies within features becomes doubly important. Therefore, unsupervised models are expected to be better suited for this kind of real-world application. However, taking a simultaneous look at the results of Table 4 and Figure 7 reveals that they have a low recall despite having relatively high accuracy. Further, the KNN model ranks lowest. This weakness in the performance of unsupervised models is somewhat comprehensible. On the one hand, these models typically use density- or proximity-based

approaches to recognize similar patterns. On the other hand, the fraud-prone samples are produced by generators as similar and close as possible to the fraud-free ones. Therefore, due to mixing two classes' samples, these models wrongly identify many fraudulent financial statements as non-fraudulent. As a result, using an unsupervised model individually is not sufficient for the current study.

Table 4 (a). The performance comparison of several outlier detection models in each segment of the data, in terms of precision, recall, and accuracy.

Model	Train					Test				
	Mean precision (variance)		Mean recall (variance)		Mean accuracy (variance)	Mean precision (variance)		Mean recall (variance)		Mean accuracy (variance)
	–	+	–	+		–	+	–	+	
iForest	0.560 (0.0004)	0.971 (0.0033)	0.994 (0.0001)	0.219 (0.0004)	0.606 (0.0006)	0.592 (0.0009)	1 (0)	1 (0)	0.292 (0.0140)	0.652 (0.0010)
LSCP	0.554 (0.0003)	0.943 (0.0049)	0.988 (0.0002)	0.212 (0.0003)	0.599 (0.0005)	0.556 (0.0013)	0.800 (0.1600)	1 (0)	0.187 (0.0217)	0.600 (0.0025)
KNN	0.478 (0.0001)	0.256 (0.0078)	0.880 (0.0001)	0.044 (0.0005)	0.461 (0.0001)	0.474 (0.0029)	0.233 (0.0844)	0.844 (0.0153)	0.0694 (0.0082)	0.460 (0.0059)
ECOD	0.543 (0.0007)	1 (0)	1 (0)	0.166 (0.0069)	0.583 (0.0019)	0.600 (0.0048)	1 (0)	1 (0)	0.295 (0.0367)	0.652 (0.0087)
COPOD	0.546 (0.0003)	0.971 (0.0033)	0.994 (0.0001)	0.179 (0.0032)	0.586 (0.0008)	0.558 (0.0014)	0.900 (0.0400)	0.976 (0.0023)	0.220 (0.0136)	0.600 (0.0025)
SVM	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)

	Logistic Regression	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)
	XGBOD	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	0.976 (0.0023)	0.976 (0.0023)	0.972 (0.0031)	0.978 (0.0019)	0.974 (0.0010)
	iForest	0.558 (0.0014)	1 (0)	1 (0)	0.205 (0.0001)	0.602 (0.0011)	0.570 (0.0196)	1 (0)	1 (0)	0.264 (0.0180)	0.628 (0.0175)
	LSCP	0.558 (0.0014)	1 (0)	1 (0)	0.205 (0.0001)	0.602 (0.0011)	0.556 (0.0165)	0.800 (0.1600)	1 (0)	0.220 (0.0153)	0.610 (0.0104)
	KNN	0.526 (0.0013)	0.739 (0.0180)	0.948 (0.0008)	0.142 (0.0005)	0.546 (0.0011)	0.518 (0.0208)	0.633 (0.1378)	0.944 (0.0051)	0.136 (0.0075)	0.546 (0.0187)
	ECOD	0.560 (0.0016)	1 (0)	1 (0)	0.205 (0.0017)	0.602 (0.0017)	0.564 (0.0190)	0.800 (0.1600)	1 (0)	0.239 (0.0275)	0.618 (0.0175)
	COPOD	0.560 (0.0016)	1 (0)	1 (0)	0.195 (0.0011)	0.597 (0.0016)	0.550 (0.0162)	0.800 (0.1600)	1 (0)	0.189 (0.0197)	0.598 (0.0138)
	SVM	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)
	Logistic Regression	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)
	XGBOD	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	0.968 (0.0016)	0.984 (0.0010)	0.976 (0.0023)	0.946 (0.0049)	0.970 (0.0006)

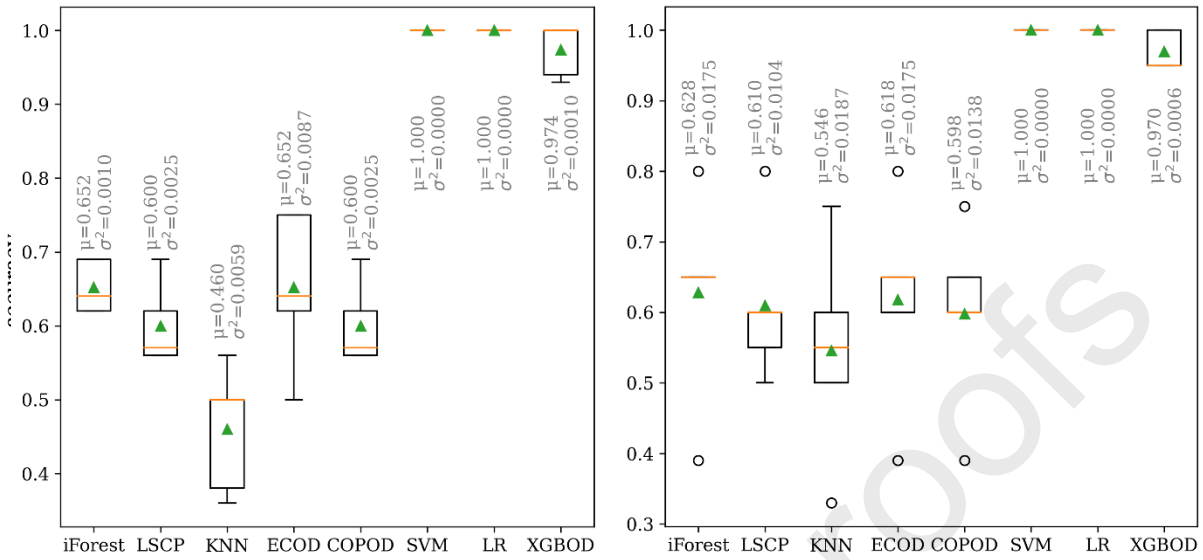
−: non-fraudulent class, +: fraudulent class

Table 4 (b). The performance comparison of several outlier detection models in each segment of the data, in terms of precision, recall, and accuracy.

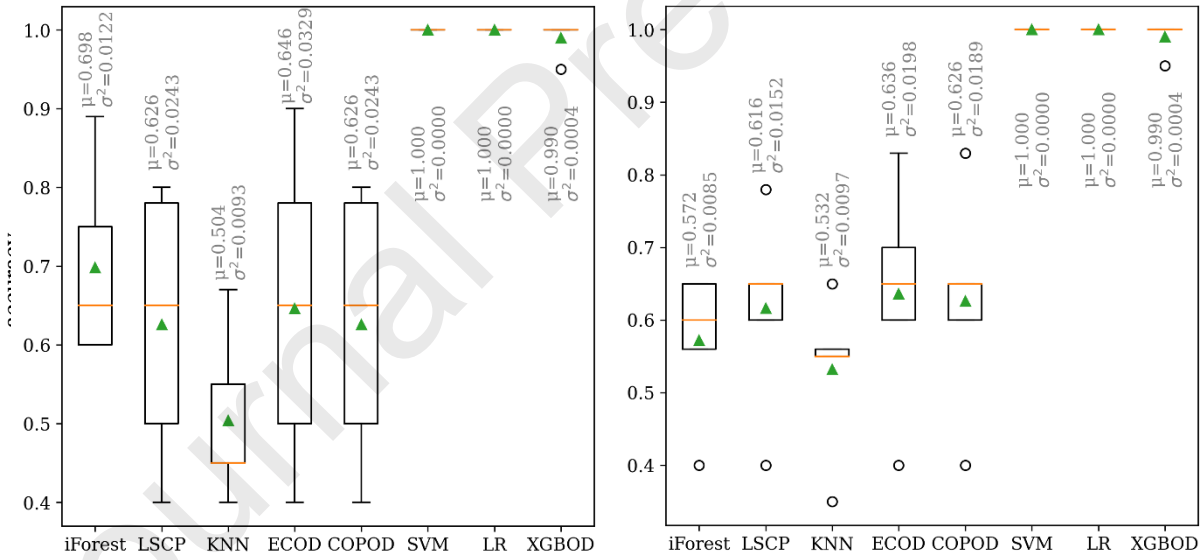
Model	Train					Test				
	Mean precision (variance)		Mean recall (variance)		Mean accuracy (variance)	Mean precision (variance)		Mean recall (variance)		Mean accuracy (variance)
	–	+	–	+		–	+	–	+	
iForest	0.552 (0.0011)	0.950 (0.0100)	0.990 (0.0004)	0.195 (0.0007)	0.592 (0.0014)	0.632 (0.0146)	1 (0)	1 (0)	0.397 (0.0414)	0.698 (0.0122)
LSCP	0.558 (0.0006)	1 (0)	1 (0)	0.205 (0.0001)	0.602 (0.0005)	0.584 (0.0193)	0.800 (0.1600)	1 (0)	0.275 (0.0433)	0.626 (0.0243)
KNN	0.502 (0.0010)	0.494 (0.0280)	0.914 (0.0011)	0.087 (0.0015)	0.500 (0.0019)	0.504 (0.0082)	0.533 (0.1600)	0.906 (0.0069)	0.110 (0.0066)	0.504 (0.0093)
ECOD	0.562 (0.0009)	1 (0)	1 (0)	0.215 (0.0028)	0.607 (0.0013)	0.606 (0.0285)	0.800 (0.1600)	1 (0)	0.325 (0.0758)	0.646 (0.0329)
COPOD	0.560 (0.0009)	1 (0)	1 (0)	0.206 (0.0023)	0.602 (0.0013)	0.584 (0.0193)	0.800 (0.1600)	1 (0)	0.275 (0.0433)	0.626 (0.0243)
SVM	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)
Logistic Regression	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)

	XGBOD	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	0.980 (0.0016)	1 (0)	1 (0)	0.982 (0.0013)	0.990 (0.0004)
	iForest	0.530 (0.0014)	0.775 (0.0088)	0.954 (0.0004)	0.160 (0.0007)	0.556 (0.0018)	0.540 (0.0114)	0.900 (0.04)	0.960 (0.0064)	0.186 (0.0057)	0.572 (0.0085)
	LSCP	0.556 (0.0008)	1 (0)	1 (0)	0.205 (0.0002)	0.602 (0.0007)	0.568 (0.0163)	1 (0)	1 (0)	0.236 (0.0221)	0.616 (0.0152)
	KNN	0.520 (0.0013)	0.693 (0.0032)	0.938 (0.0002)	0.140 (0.0004)	0.538 (0.0011)	0.522 (0.0113)	0.733 (0.0511)	0.922 (0.0055)	0.152 (0.0042)	0.532 (0.0097)
	ECOD	0.564 (0.0017)	1 (0)	1 (0)	0.222 (0.0029)	0.610 (0.0020)	0.588 (0.0213)	1 (0)	1 (0)	0.286 (0.0335)	0.636 (0.0198)
	COPOD	0.562 (0.0010)	1 (0)	1 (0)	0.216 (0.0008)	0.607 (0.0011)	0.580 (0.0202)	1 (0)	1 (0)	0.261 (0.0378)	0.626 (0.0189)
	SVM	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)
	Logistic Regression	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)
	XGBOD	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	0.982 (0.0013)	1 (0)	1 (0)	0.980 (0.0016)	0.990 (0.0004)

−: non-fraudulent class, +: fraudulent class



(a) loans (b) deposits

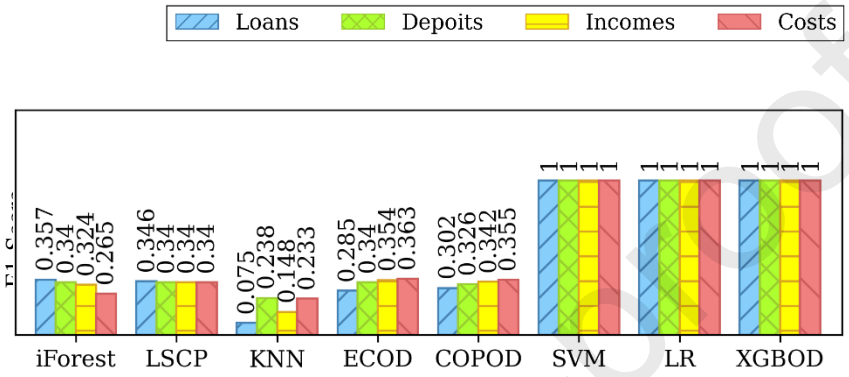


(c) incomes (d) costs

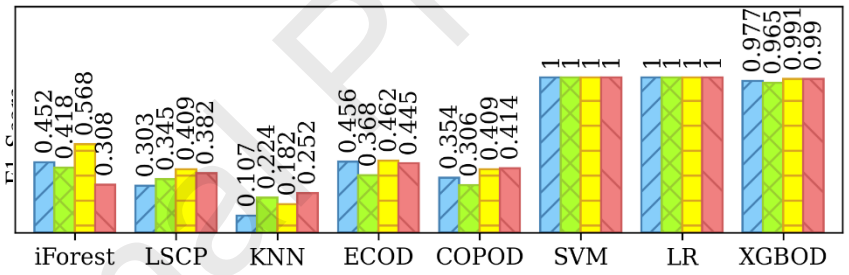
Figure 7. The performance comparison of outlier detection models in each segment of the data, in terms of accuracy. All boxplots are drawn based on 5-fold cross-validation results. The mean and variance of 5-folds are also reported using μ and σ^2 notations.

According to the results presented in Table 4 and Figure 7, the XGBOD algorithm used in the proposed method is ranked first, right after the supervised algorithms. It has a more stable behavior with variance

values close to zero than other unsupervised models. In addition, this model is able to address the challenge of extensive and high-dimensional data by involving multiple unsupervised models to distill a set of profitable features. Besides, it also takes advantage of the strength of supervised models by using XGBoost as its final decision-maker. In order to verify these findings, a performance comparison between different models in terms of F1-score is also made which is presented in Figure 8. Figure 8 (a) demonstrates the obtained results on the training data, while the test results are provided in Figure 8 (b).



(a) Train



(b) Test

Figure 8. The F1-Score comparison of outlier detection models in each segment of the data.

6. Conclusion and Future Work

In this paper, a new approach has been proposed to detect fraud in bank financial statements. The basic idea is to adopt generative adversarial networks instead of over-sampling, under-sampling, or one-class classification techniques to make the approach applicable to real-world scenarios whose data is highly imbalanced with no or few fraudulent samples. The second idea is to tackle with high-dimensionality of the feature space by leveraging the ensemble of supervised and unsupervised models. In particular, the proposed approach utilizes a kind of generative adversarial model called MO-GAAL to fabricate a set of fraud-prone samples that have unconventional behavior on the one hand and are difficult to distinguish

from fraud-free samples on the other hand. Further, samples are classified by an ensemble model, namely XGBOD, in which the outlier scores of each sample are first estimated by a collection of unsupervised models, and then these scores form a new feature vector to be classified by a supervised model named XGBoost. In summary, the ability to train an efficient decision-making model even in the absence of actual fraudulent samples is the main advantage of this work.

Another achievement of the current research has been the construction of a real-world dataset by collecting a number of banks' annual financial statements and introducing three new features which can be interpreted even by non-specialists. While the growth rate feature enables the model to analyze the bank's financial behavior relative to the previous year, the vertical and horizontal ratios measure the magnitude of a data item among similar data items in the same fiscal year. The experimental results on this dataset have indicated that the proposed method is capable of generating high-quality fraud-prone samples and making effective detections in terms of accuracy, precision, recall, and F1-Score. In future work, we would like to include financial ratios in feature vectors and specify the exact position of the occurrence of fraud in the financial statement. Moreover, comparing the performance of the proposed approach with one-class classification approaches and assessing them on other real-world datasets sounds to be an absorbing research topic.

References

ACFE. (2022). Occupational Fraud 2022: A Report to the nations. In *Acfe*.

Al-Hashedi, K. G., & Magalingam, P. (2021). Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019. *Computer Science Review*, *40*, 100402. https://doi.org/10.1016/j.cosrev.2021.100402

Ashtiani, M. N., & Raahemi, B. (2021). Intelligent Fraud Detection in Financial Statements using Machine Learning and Data Mining: A Systematic Literature Review. *IEEE Access*, *PP*, 72504–72525. https://doi.org/10.1109/ACCESS.2021.3096799

Bagga, S., Goyal, A., Gupta, N., & Goyal, A. (2020). Credit Card Fraud Detection using Pipeling and Ensemble Learning. *Procedia Computer Science*, *173*(2019), 104–112. https://doi.org/10.1016/j.procs.2020.06.014

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the Fifth Annual Workshop on Computational Learning Theory - COLT '92*, 144–152. https://doi.org/10.1145/130385.130401

Carcillo, F., Le Borgne, Y.-A., Caelen, O., Kessaci, Y., Oblé, F., & Bontempi, G. (2021). Combining unsupervised and supervised learning in credit card fraud detection. *Information Sciences*, *557*, 317–331. https://doi.org/10.1016/j.ins.2019.05.042

Chen, J. I. Z., & Lai, K. L. (2021). Deep Convolution Neural Network Model for Credit-Card Fraud Detection and Alert. *Journal of Artificial Intelligence and Capsule Networks*, *3*(2), 101–112. https://doi.org/10.36548/jaicn.2021.2.003

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Craja, P., Kim, A., & Lessmann, S. (2020). Deep learning for detecting financial statement fraud. *Decision Support Systems*, *139*, 113421. https://doi.org/10.1016/j.dss.2020.113421

Cramer, J. (2003). The Origins of Logistic Regression. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.360300

De Rossi, G., Kolodziej, J., & Brar, G. (2020). A recommender system for active stock selection. *Computational Management Science*, *17*(4), 517–547. https://doi.org/10.1007/s10287-018-0342-9

Dutta, I., Dutta, S., & Raahemi, B. (2017). Detecting financial restatements using data mining techniques. *Expert Systems with Applications*, *90*, 374–393. https://doi.org/10.1016/j.eswa.2017.08.030

El Kafhali, S., & Tayebi, M. (2022). Generative Adversarial Neural Networks based Oversampling Technique for Imbalanced Credit Card Dataset. *2022 6th SLAAI International Conference on Artificial Intelligence (SLAAI-ICAI)*, 1–5. https://doi.org/10.1109/SLAAI-ICAI56923.2022.10002630

Fiore, U., De Santis, A., Perla, F., Zanetti, P., & Palmieri, F. (2019). Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. *Information Sciences*, *479*, 448–455. https://doi.org/10.1016/j.ins.2017.12.030

Gangwar, A. K., & Ravi, V. (2019). WiP: Generative Adversarial Network for Oversampling Data in Credit Card Fraud Detection. In *International Conference on Information Systems Security* (pp. 123–134). https://doi.org/10.1007/978-3-030-36945-3_7

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139–144. https://doi.org/10.1145/3422622

Gray, G. L., & Debreceny, R. S. (2014). A taxonomy to guide research on the application of data mining to fraud detection in financial statement audits. *International Journal of Accounting Information Systems*, *15*(4), 357–380. https://doi.org/10.1016/j.accinf.2014.05.006

Gupta, S., & Mehta, S. K. (2021). Data Mining-based Financial Statement Fraud Detection: Systematic Literature Review and Meta-analysis to Estimate Data Sample Mapping of Fraudulent Companies Against Non-fraudulent Companies. *Global Business Review*. https://doi.org/10.1177/0972150920984857

Hajek, P. (2019). Interpretable Fuzzy Rule-Based Systems for Detecting Financial Statement Fraud. In *IFIP International Conference on Artificial Intelligence Applications and Innovations* (pp. 425–436). https://doi.org/10.1007/978-3-030-19823-7_36

Hajek, P., & Henriques, R. (2017). Mining corporate annual reports for intelligent detection of financial statement fraud – A comparative study of machine learning methods. *Knowledge-Based Systems*, *128*, 139–152. https://doi.org/10.1016/j.knosys.2017.05.001

Hashim, H. A., Salleh, Z., Shuhaimi, I., & Ismail, N. A. N. (2020). The risk of financial fraud: a management perspective. *Journal of Financial Crime*, *27*(4), 1143–1159. https://doi.org/10.1108/JFC-04-2020-0062

Huang, S. Y., Lin, C. C., Chiu, A. A., & Yen, D. C. (2017). Fraud detection using fraud triangle risk

factors. *Information Systems Frontiers*, *19*(6), 1343–1356. https://doi.org/10.1007/s10796-016-9647-9

Jan, C. (2018). An Effective Financial Statements Fraud Detection Model for the Sustainable Development of Financial Markets: Evidence from Taiwan. *Sustainability*, *10*(2), 513. https://doi.org/10.3390/su10020513

Jeragh, M., & AlSulaimi, M. (2018). Combining Auto Encoders and One Class Support Vectors Machine for Fraudulant Credit Card Transactions Detection. *2018 Second World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)*, 178–184. https://doi.org/10.1109/WorldS4.2018.8611624

Karlos, S., Kostopoulos, G., Kotsiantis, S., & Tampakas, V. (2017). Using Active Learning Methods for Predicting Fraudulent Financial Statements. In *International Conference on Engineering Applications of Neural Networks* (Vol. 744, pp. 351–362). https://doi.org/10.1007/978-3-319-65172-9_30

Kingsley, U., & Patrick, A. (2021). *Application Of Neural Network Models In Predicting Fraudulent Financial Reporting In Listed Manufacturing Firms In Nigeria*. *7*, 17–36. https://doi.org/10.46654/ij.24889849.s77602

Lee, M., Lin, J., & Gran, E. G. (2020). RePAD: Real-Time Proactive Anomaly Detection for Time Series. In *Proceedings of the 34th International Conference on Advanced Information Networking and Applications (AINA-2020)* (pp. 1291–1302). https://doi.org/10.1007/978-3-030-44041-1_110

Li, Z., Zhao, Y., Botta, N., Ionescu, C., & Hu, X. (2020). COPOD: Copula-Based Outlier Detection. *2020 IEEE International Conference on Data Mining (ICDM)*, *2020-Novem*(1), 1118–1123. https://doi.org/10.1109/ICDM50108.2020.00135

Li, Z., Zhao, Y., Hu, X., Botta, N., Ionescu, C., & Chen, G. H. (2022). ECOD: Unsupervised Outlier Detection Using Empirical Cumulative Distribution Functions. *IEEE Transactions on Knowledge and Data Engineering*, 1–1. https://doi.org/10.1109/TKDE.2022.3159580

Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008). Isolation Forest. *2008 Eighth IEEE International Conference on Data Mining*, 413–422. https://doi.org/10.1109/ICDM.2008.17

Liu, Y., Li, Z., Zhou, C., Jiang, Y., Sun, J., Wang, M., & He, X. (2019). Generative Adversarial Active Learning for Unsupervised Outlier Detection. *IEEE Transactions on Knowledge and Data Engineering*, *32*(8), 1517–1528. https://doi.org/10.1109/TKDE.2019.2905606

Lokanan, M., Tran, V., & Vuong, N. H. (2019). Detecting anomalies in financial statements using machine learning algorithm: The case of Vietnamese listed firms. *Asian Journal of Accounting Research*, *4*(2), 181–201. https://doi.org/10.1108/AJAR-09-2018-0032

Moepya, S. O., Akhoury, S. S., & Nelwamondo, F. V. (2014). Applying Cost-Sensitive Classification for Financial Fraud Detection under High Class-Imbalance. *2014 IEEE International Conference on Data Mining Workshop*, 183–192. https://doi.org/10.1109/ICDMW.2014.141

Mohammadi, M., Yazdani, S., Khanmohammadi, M. H., & Maham, K. (2020). Financial Reporting Fraud Detection: An Analysis of Data Mining Algorithms. *International Journal of Finance & Managerial Accounting*, *4*(16), 1–12.

Nguyen, T. T., Tahir, H., Abdelrazek, M., & Babar, A. (2020). Deep Learning Methods for Credit Card Fraud Detection. *CoRR Abs/2012.03754*.

Noels, S., Vandermarliere, B., Bastiaensen, K., & De Bie, T. (2022). An Earth Mover's Distance Based Graph Distance Metric For Financial Statements. *2022 IEEE Symposium on Computational Intelligence for Financial Engineering and Economics (CIFEr)*, 1–8. https://doi.org/10.1109/CIFEr52523.2022.9776204

Oh, J., Hong, J. Y., & Baek, J. G. (2019). Oversampling method using outlier detectable generative adversarial network. *Expert Systems with Applications*, *133*, 1–8. https://doi.org/10.1016/j.eswa.2019.05.006

Omidi, M., Min, Q., Moradinaftchali, V., & Piri, M. (2019). The Efficacy of Predictive Methods in Financial Statement Fraud. *Discrete Dynamics in Nature and Society*, *2019*, 1–12. https://doi.org/10.1155/2019/4989140

Paper, D. (2021). Generative Adversarial Networks. In *State-of-the-Art Deep Learning Models in TensorFlow* (pp. 243–263). Apress. https://doi.org/10.1007/978-1-4842-7341-8_10

Patel, H., Parikh, S., Patel, A., & Parikh, A. (2019). An Application of Ensemble Random Forest Classifier for Detecting Financial Statement Manipulation of Indian Listed Companies. In *Recent Developments in Machine Learning and Data Analytic* (Vol. 740, pp. 349–360). https://doi.org/10.1007/978-981-13-1280-9_33

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559–572. https://doi.org/10.1080/14786440109462720

Peterson, L. (2009). K-nearest neighbor. *Scholarpedia*, *4*(2), 1883. https://doi.org/10.4249/scholarpedia.1883

Petković, Z., Milojević, S., Novaković, S., & Trivunović Sajić, Đ. (2021). Fraudulent Financial Reporting from the Managers' Perspective. *International Academic Journal*, *2*(2), 35–39.

Ravisankar, P., Ravi, V., Raghava Rao, G., & Bose, I. (2011). Detection of financial statement fraud and feature selection using data mining techniques. *Decision Support Systems*, *50*(2), 491–500. https://doi.org/10.1016/j.dss.2010.11.006

Sadgali, I., Sael, N., & Benabbou, F. (2019). Performance of machine learning techniques in the detection of financial frauds. *Procedia Computer Science*, *148*(Icds 2018), 45–54. https://doi.org/10.1016/j.procs.2019.01.007

Saia, R., & Carta, S. (2019). Evaluating the benefits of using proactive transformed-domain-based techniques in fraud detection tasks. *Future Generation Computer Systems*, *93*, 18–32. https://doi.org/10.1016/j.future.2018.10.016

Sethia, A., Patel, R., & Raut, P. (2018). Data Augmentation using Generative models for Credit Card Fraud Detection. *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, 1–6. https://doi.org/10.1109/CCAA.2018.8777628

Shahriar, S. (2022). GAN computers generate arts? A survey on visual arts, music, and literary text

generation using generative adversarial network. *Displays*, *73*, 102237. https://doi.org/10.1016/j.displa.2022.102237

Sim, E.-A., Lee, S., Oh, J., & Lee, J. (2021). GANs and DCGANs for generation of topology optimization validation curve through clustering analysis. *Advances in Engineering Software*, *152*, 102957. https://doi.org/10.1016/j.advengsoft.2020.102957

Song, X. P., Hu, Z. H., Du, J. G., & Sheng, Z. H. (2014). Application of Machine Learning Methods to Risk Assessment of Financial Statement Fraud: Evidence from China. *Journal of Forecasting*, *33*(8), 611–626. https://doi.org/10.1002/for.2294

Strelcenia, E., & Prakoonwit, S. (2022). GAN-based Data Augmentation for Credit Card Fraud Detection. *2022 IEEE International Conference on Big Data (Big Data)*, 6812–6814. https://doi.org/10.1109/BigData55660.2022.10020419

Syahria, R. (2019). Detecting financial statement fraud using fraud diamond (A study on banking companies listed on the indonesia stock exchange period 2012-2016). *Asia Pacific Fraud Journal*, *4*(2), 183–190. https://doi.org/10.21532/apfjournal.v4i2.114

Temponeras, G. S., Alexandropoulos, S. A. N., Kotsiantis, S. B., & Vrahatis, M. N. (2019). Financial Fraudulent Statements Detection through a Deep Dense Artificial Neural Network. *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, 1–5. https://doi.org/10.1109/IISA.2019.8900741

Throckmorton, C. S., Mayew, W. J., Venkatachalam, M., & Collins, L. M. (2015). Financial fraud detection using vocal, linguistic and financial cues. *Decision Support Systems*, *74*, 78–87. https://doi.org/10.1016/j.dss.2015.04.006

Tin Kam Ho. (1995). Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, *1*, 278–282. https://doi.org/10.1109/ICDAR.1995.598994

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(11).

West, J., & Bhattacharya, M. (2016). Intelligent financial fraud detection: A comprehensive review. *Computers & Security*, *57*, 47–66. https://doi.org/10.1016/j.cose.2015.09.005

Xue, Y., Tong, W., Neri, F., & Zhang, Y. (2022). PEGANs: Phased Evolutionary Generative Adversarial Networks with Self-Attention Module. *Mathematics*, *10*(15), 2792. https://doi.org/10.3390/math10152792

Yao, J., Zhang, J., & Wang, L. (2018). A financial statement fraud detection model based on hybrid data mining methods. *2018 International Conference on Artificial Intelligence and Big Data (ICAIBD)*, 57–61. https://doi.org/10.1109/ICAIBD.2018.8396167

Zhao, C., Zhu, Y., Du, Y., Liao, F., & Chan, C.-Y. (2022). A Novel Direct Trajectory Planning Approach Based on Generative Adversarial Networks and Rapidly-Exploring Random Tree. *IEEE Transactions on Intelligent Transportation Systems*, *23*(10), 17910–17921. https://doi.org/10.1109/TITS.2022.3164391

Zhao, Y., & Hryniewicki, M. K. (2018). XGBOD: Improving Supervised Outlier Detection with

Unsupervised Representation Learning. *2018 International Joint Conference on Neural Networks (IJCNN)*, 1–8. https://doi.org/10.1109/IJCNN.2018.8489605

Zhao, Y., Nasrullah, Z., Hryniewicki, M. K., & Li, Z. (2018). LSCP: Locally Selective Combination in Parallel Outlier Ensembles. *Proceedings of the 2019 SIAM International Conference on Data Mining*, 585–593. https://doi.org/10.1137/1.9781611975673.66

Seyyede Zahra Aftabi: Methodology, Software, Investigation, Writing - Original Draft, Writing - Reviewing & Editing, Visualization, Data Curation.

Ali Ahmadi: Conceptualization, Methodology, Reviewing & Editing, Supervision.

Saeed Farzi: Conceptualization, Methodology, Reviewing & Editing, Supervision.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: