Computer Science Review 40 (2021) 100402

Contents lists available at ScienceDirect

Computer Science Review

journal homepage: www.elsevier.com/locate/cosrev

Review article

Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019

Khaled Gubran Al-Hashedi *, Pritheega Magalingam

Advanced Informatics Department, Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia

ARTICLE INFO

Article history:
Received 7 April 2020
Received in revised form 9 April 2021
Accepted 15 April 2021
Available online 23 April 2021

Keywords:
Financial fraud
Data mining technique
Credit card fraud
Insurance fraud
Bitcoin fraud
Financial statement fraud

ABSTRACT

This paper gives a comprehensive revision of the state-of-the-art research in detecting financial fraud from 2009 to 2019 inclusive and classifying them based on their types of fraud and data mining technology utilized in detecting financial fraud. The review result yielded a sample of 75 relevant articles (58 conference papers with 17 peer-reviewed journal articles) that are categorized into four main groups (bank fraud, insurance fraud, financial statement fraud, and cryptocurrency fraud). The study shows that 34 data mining techniques were used to identify fraud throughout various financial applications. The SVM is found to be one of the most widely used financial fraud detection techniques that carry about 23% of the overall study, followed by both Naïve Bayes and Random Forest, resulting in 15%. The results of our comprehensive review revealed that most data mining techniques are extensively implemented to bank fraud and insurance fraud with a total of 61 research studies out of 75 that constitute the largest portion equal to 81.33% of the overall number of papers. This review provides a good reference source in guiding the detection of financial fraud for both academic and practical industries with useful information on the most significant data mining techniques used and shows the list of countries that are exposed to financial fraud. Our review contributes by expanding the sample of the reviewed articles that were not included by previous research and presents a summary of the prominent works done by various researchers in the field of financial fraud.

Contents

* Corresponding author.
E-mail address: khaled.gubran1@gmail.com (K.G. Al-Hashedi).

https://doi.org/10.1016/j.cosrev.2021.100402
1574-0137/

1. Introduction

Over the past years, a technological revolution has occurred on the Internet that paved to the emergence of modern services especially in e-commerce and money transfer. E-commerce is one of the many economic domains in information and communications technology that contributed to business improvement, paved the way for managing medium and small companies, reducing costs and saving time, and increasing productivity [1]. The growth in e-commerce enabled most companies and organizations to perform their financial transactions electronically through the adoption of payment systems such as Healthcare Insurance systems, Telecommunication systems, and the Financial Sector. Lately, there is a noticeable rise in the number of financial transactions due to the large adoption of Internet bank services and financial institutions as well as e-commerce. The Internet has played an important role in making an online payment, that in return has become a breeding ground for malicious attackers to exploit these services for performing fraudulent businesses leading to the emergence of cybercrime that targets e-banking services as a result of a significant rise in the number of frauds by fraudsters causing an annual loss with an approximate cost of billions of dollars [2].

Financial fraud is an essential problem that affects both the finance sector and everyday life and plays a critical role in impacting integrities and confidences in financial sectors as well as the individuals' cost of living [3]. Financial fraud is known as financial abuse which is a big concern in economic society causing huge losses to the economy of governments, organizations, corporate sectors or even individuals. It can be defined as an act of wrongful or illegal behavior, resulting in a beneficial gain to either individual or organization from unethical and illegal ways [4]. Fraud detection techniques were introduced to identify abnormal activities, that occurred in past transactions aiming to discover cases that fraudsters intend to violate the values that the organizations make in exchange for supplying services. Various methods have been proposed to detect fraud, but these methods are infeasible due to the constant evolution of new methods developed by fraudsters or in new technologies such as cryptocurrency. According to the fact that any E-commerce system that involves online transactions such as financial services is vulnerable to be compromised by fraudsters [5]. Therefore, anti-fraud has become a topic of interest by many scientists to explore the issues related to this field. The important issues of fraud motivated scientists to develop detection methods or even estimate fraud risk.

Data mining is an approach used in extracting meaningful data from a given dataset using one or more approaches such as statistical, machine learning, mathematical or artificial intelligence techniques. Among these approaches, different kinds of techniques can be applied for financial fraud such as Naïve Bayes (NB), support vector machine (SVM), Logistic Regression (LR), and many more [11]. Generally, data mining is usually used to discover financial frauds that can be classified into six categories such as classification, visualization, outlier detection, clustering, regression, and prediction [4]. Furthermore, it is believed that the last two years have witnessed a large operation of fraud activities targeting 1 out of 3 organizations. But the most unexpected thing is that only 10% of these financial frauds are discovered by chance [14]. Several studies have been conducted on the annual cost of financial fraud in the U.S. and the U.K. Moreover, the figures of these studies show that financial fraud in the U.S. made a loss of $400 billion every year while 1.6 billion pounds to insurers in the United Kingdom. Moreover, A study estimates that fraud activities will be increased especially in online fraud from $10.7 billion in 2015 to $25.6 billion in 2020 [15]. Besides that, financial fraud also has tremendous consequences on society, which can be used for supplying illegal activities such as organized crime and fund terrorism [16]. However, most organizations are interested to take action against fraudulent activities.

1.1. Related surveys

Fraud detection has been studied by researchers and scientists in various surveys and review articles that have emerged in academic publications. In [17–20], for example, conducted surveys across various kinds of fraud on credit card and money laundering based on statistical and data mining techniques. Similarly, Delamaire et al. [21], made a study on several sorts of fraud on credit cards including counterfeit fraud, bankruptcy fraud, and discussed proper methods to prevent them. In contrast, Zhang and Zhou [22] explored data mining techniques on financial applications including the stock market and fraud detection. Raj et al. [23] examined several techniques utilized to uncover fraud on credit cards. Phua et al. [24] surveyed fraud detection data mining (DM) techniques in various kinds that are credit card, insurance, and telecoms subscription fraud. In [6,9,25], investigated several fraud detection techniques in healthcare sectors using statistical methods. Ngai et al. [7] conducted a review and classification of several financial fraud detection techniques that are applied to data mining covering 49 articles ranging from 1997 to 2008. Sithic and Balasubramanian [10] provide an extensive survey of a variety of types of fraud to detect fraud in medical and auto insurance systems based on DM techniques. Popat and Chaudhary [13] analyzed several machine learning classification techniques with their methodology and challenges to detect fraud in credit cards. Ryman-Tubb, Krause et al. [12] surveyed and benchmarked existing methods of payment card fraud detection using transactional volumes in 2017. The survey confirmed that only eight methods have a practical performance to

Table 1
Comparison of the proposed review with existing reviews.

No.	Article	Year	Fraud area	Classification of fraud types	Pros and cons	Datasets	Validation methods
1	Travaille [6]	2011	Healthcare insurance fraud	✔			
2	Ngai et al. [7]	2011	All	✔			
3	Richhariya and Singh [8]	2012	Credit card and online auction	✔			
4	Liu and Vasarhelyi [9]	2013	Healthcare insurance fraud				
5	Sithic and Balasubramanian [10]	2013	Insurance fraud	✔			
6	Albashrawi and Lowell [11]	2016	All	✔		✔	
7	Krause et al. [12]	2018	Credit card fraud			✔	✔
8	Popat and Chaudhary [13]	2018	Credit card fraud	✔			
9	**Proposed review**	–	All	✔	✔	✔	✔

be deployed in the industry. Richhariya and Singh [8] introduced a comprehensive survey and review for different data mining techniques used to detect financial fraud. In the extension work of Ngai et al. Albashrawi and Lowell [11] presented a revision for one decade from 2004 to 2015 covering DM tools to reveal fraud in financial domains. However, this was not comprehensive enough as they left out validation methods, pros, and cons of most data mining techniques. There has been a noticeable rise in the number of financial fraud and fraudulent activities in recent years [16]. Therefore, this motivated us to extend the Albashrawi and Lowell [11] review. Table 1 shows the comparison of the current review with the existing reviews related to financial fraud detection applying data mining techniques. Comparison is performed based on the fraud area, classification of fraud types, pros and cons of data mining techniques, dataset, and evaluation metrics. As shown in Table 1, it can be clearly seen that only one review has covered all types of financial fraud but has not focused on evaluation metrics, pros, and cons of most of the data mining techniques. Thus, a blank cell in the table indicates that the review has not covered all aspects of financial fraud and some information is missing for that specific column and a mark ✔ means that the review has covered that area. For that reason, this review intends to cover that missing information.

1.2. Contribution

The main objective of this paper is to extend the work of Albashrawi and Lowell [11], and therefore the significant contributions of this paper are as follows:

1. This review focuses on expanding the sample of the reviewed articles and presents a comprehensive summary of recent research on data mining techniques in the financial fraud detection domain. It covers several novel works and references that were not presented in previous reviews.
2. It reveals the most frequent financial fraud detection technique used in data mining that can result in disclosing fraud in financial domains with high-performance accuracy and provides a new classification framework for financial fraud as well as creating a roadmap for researchers and practitioners seeking to better understand this field.
3. It provides information regarding various datasets and validation metrics used to estimate the performance of the data mining techniques and it can be used as a reference source for both academic and practical applications. This paper will be very useful for future research.
4. This paper summarizes the pros, cons, business applications and provides a detailed description of the most data mining techniques on financial fraud detection and presents a summary of the findings of each study.
5. It discloses the context of financial fraud types that are mostly exposed for such fraudulent activities and reveals the countries that are highly exposed to financial fraud.

1.3. Paper organization

This paper begins with an introduction of financial fraud in Section 1 that includes related surveys, comparison of existing review, contributions, and organization. Section 2 provides a taxonomy of common financial fraud types and a description of different types of financial frauds such as credit card fraud, mortgage fraud, money laundering, financial statement fraud, securities and commodities fraud, insurance fraud, and cryptocurrency fraud. Section 3 provides a detailed discussion of the review methodology used in this paper. Various data mining methods utilized to detect fraud in financial domains are reviewed in Section 4. The literature related to these techniques has also been covered in this section. The result and discussion about financial fraud detection are illustrated in Section 5. This includes distribution of papers by journal and conference title, classification of financial fraud types, pros and cons of each data mining technique reviewed in this paper, distribution of paper by data mining techniques, and distribution of examined papers by publication year. Lastly, the review and future work are concluded in Section 6.

2. Types of financial fraud

Fraudulent activities vary depending on industry sectors. According to several studies [2,4,15,16,26,27], financial fraud can be categorized into different groups such as bank fraud, insurance, money laundering, financial statement, mortgage, and health care fraud. However, this paper is introducing crypto-currency fraud. However, a credit card is a popular kind of fraud due to its widespread use. Therefore, investigating the financial domain studies can identify the best solution for this paper. Therefore, the following section will discuss the techniques and studies conducted in the financial domains from 2009 to 2019 Fig. 1 depicts the most common areas of financial fraud.

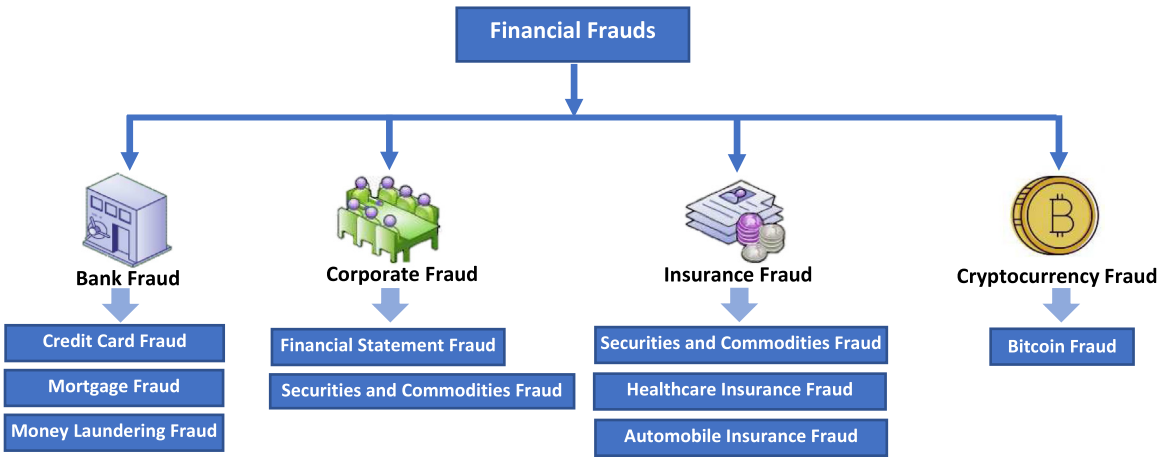

Fig. 1. Common types of financial fraud.

2.1. Credit card fraud

The concept of performing money transactions electronically without the existence of physical money is always referred to as a credit. The credit card is a small and thin piece made of plastic containing customer details and credit service. Credit Card is the most well-known services growing domain in e-banking that is widely used in online money transactions and E-Commerce. In contrast, the fast growth in credit card use has given rise to various forms of fraud. Fraudsters utilize a credit card to perform illegitimate transactions that are resulting in large losses to cardholders and banks [28]. On the other hand, the production of fake cards has assisted fraudsters to conduct transactions easier than ever before. Credit card fraud detection is the act of categorizing suspicious transactions into two categories of ordinary and counterfeit or suspicious transactions. A normal transaction refers to the behavior of the individuals who performed the transactions successfully and normally without showing any false transactions while the fake or suspicious transaction refers to the behavior of the individuals who frequently accessed the system illegally making the system facing repeatedly errors due to the conducting of suspicious transactions [1]. Credit card fraud is considered unlawful to utilize the card without authorization by the owner. Hence, the actual owner does not know that the card is being utilized by fraudsters. Thus, fraudsters can log in to a certain account unlawfully, or making any transactions fraudulently [15].

Credit card fraud can be split into two categories: Online and Offline fraud. In Online fraud, fraudsters perform their transactions especially in online purchases via the Internet, cell phone, or even by using a web browser while in Offline fraud, the fraudsters perform their activities using a stolen credit card as the actual cardholder [2].

2.2. Mortgage fraud

It intentionally targets mortgage documents by modifying or removing information during the process of loan application in a mortgage. It occurs when there is a real estate fraud that misleads the original value of the property with a purpose to gain more material benefits or influence the financier to finance its lend [7].

2.3. Money laundering

It is a way that criminals use to tamper with or hide the origin of illegal money by encouraging fraudsters to legalize and clean the dirty money [15]. Money laundering is the action of fraudsters or organizations that intentionally exchange financial transactions with income earned from some illegal activities to disguise the original source of the money or the nature of income for the purpose of making them appear legitimate. Money laundering has a significant impact on society by which it is the main source of committing other crimes such as funding terrorism and weapons trading [4].

2.4. Financial statement fraud

It is defined as the records created by an organization containing their financial activities and financial conditions that include their expenses, loans, income, profits, etc. These records also cover some remarks made up by the management to explain the business' performance and predicted problems that may appear in the future. The different financial statements provide a clear picture of the organization's condition that shows how successful the organization is, which helps to determine if the organization is bankable. Financial statement fraud acts as deceiving financial statement users by correcting misstatement to enable the organization to look beneficial. The motivation for committing this fraud is to increase share prices, getting personal bank loans, reducing tax liabilities or attract as many investors as possible [16].

2.5. Securities and commodities fraud

It is deceptive practice occurs when an individual is asked to invest in an organization based on given fake information. This includes a variety of methods such as Ponzi Schemes, Foreign Currency Exchange Fraud, Pyramid Schemes, and many more [7].

2.6. Insurance fraud

Insurance fraud refers to the act of committing an unethical crime with the unlawful intent or misuse of an insurance policy to gain illegal profit from an insurance company. Generally, insurance is made to protect the assets and businesses of individuals or organizations from financial loss. It may occur at any stage within the insurance procedure by anyone like clients or agents [4]. This type of fraud takes place when an individual intentionally committing a fake accident or loss assets resulting in an overstated repair and injury costs. Insurance frauds may exist in various types of insurance domains such as healthcare, crop, automobile insurance, etc. In automobile claims, the fraudsters may provide a fake document containing exaggerating bills of a fabricated accident. In health care insurance fraud, the fraudsters may submit a report of false medical services claiming an excessive surgery

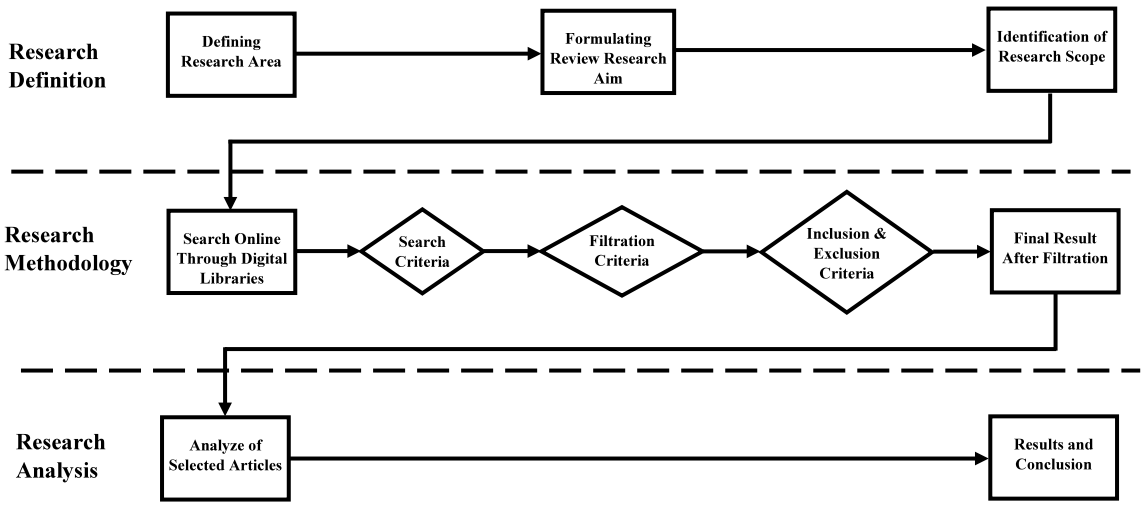

Fig. 2. Research methodology.

cost. Besides, crop insurance fraud may occur when fraudsters increase their losses due to the impact of changes in agricultural prices or a natural disaster [16].

2.7. Cryptocurrency fraud

Cryptocurrency fraud is a type of fraud that deliberately deceives naive users by providing them fake investments or services. The idea of these services is to lure innocent users with the promise of a big amount of profits as a result of their investment. Thus, many naive users are engaged with fake investments due to the absence of their knowledge of online investments e.g. Ponzi schemes [29]. Indeed, Cryptocurrency is designed as a tool for criminals to perform their activities in stealing or illegal use of digital currencies utilizing the benefit of the decentralized nature and the lack of regulations [30]. Hence, these services are intentionally fraudulent by design, they can tempt in unsuspecting many honest victims and thus generate revenues worth millions of dollars. According to Vasek and Moore [31] that scams services such as nefarious cases of btcQuick, CoinOpened, Ubitex, or BTC Promo have gained a profit from Cryptocurrency investors with estimated revenue worth $11 million.

3. Methodology

This study aims to conduct a revision of the latest financial fraud detection techniques and their implementations from 2009 to 2019 inclusive. The research methodology comprises of three main stages: definition of research, research methodology, and the last stage is to achieve research analysis, as shown in Fig. 2.

In stage 1, we began by identifying the research criteria through the definition of the research area, disclosing the main objective of the research, and finally declaring the research scope. However, the research area is specified only to all scientific research in the academic field that includes applications of data mining techniques to detect fraud in the financial domain. The research objective is to review and examine the latest data mining techniques related to financial fraud and classifying them based on their types of fraud. The research scope is to select the range of papers that includes data mining techniques used in financial domains that are published between 2009 and 2019.

In stage 2, we performed the search criteria by using numerous keywords to identify the relevant articles. However, the most acceptable search criteria that achieved relevant results as follows:

- ((‘Financial fraud’ AND ’data mining’))
- ((‘Financial fraud’ OR ‘credit card fraud’ OR ’insurance fraud’ OR ’Bitcoin Fraud’ OR ’Money laundering fraud’ OR ’financial statement fraud’ OR ’mortgage fraud’)

The search process is conducted through several digital libraries. The digital libraries that were selected in this review include IEEE, ACM, Emerald, and Elsevier. However, once the search terms were executed, the initial search result returned 644 articles from the libraries. Due to the duplication of many articles or were irrelevant to our research area, a further step was performed to filter duplication of articles concurrently and independently by the main authors to minimize the potential bias and identify differences. The Filtration process was as follows:

(1) All duplicated articles were ignored.
(2) The identification of relevant articles and the deletion of unrelated articles were performed through inclusion and exclusion criteria.
(3) Performance of quality assessment was performed to make sure that Only high-quality articles were included.

Our inclusion criteria included:

(1) Articles published between the period of 2009 and 2019.
(2) Articles that include fraud in financial fraud using either data mining or machine learning techniques.
(3) The latest version/edition of a paper.
(4) Articles that are only included in peer-reviewed journals or conference papers.

Our exclusion criteria included:

(1) Articles that are related to financial fraud detection but are irrelevant to data mining or machine learning.
(2) Articles that are not classified as peer-reviewed journals or conference papers.
(3) Articles that involve Data mining or machine learning but were not related to financial fraud detection.

Most of the relevant papers were found in Management Information Systems Quarterly related journals, e.g., Expert Systems with Applications but some were found in data science and Computing Technology related journals. Therefore, after the filtration, our search achieved a total of 75 articles (58 conference papers with 17 peer-reviewed journal articles).

In the last stage, we performed an analysis study on the obtained articles to make some conclusions and identify the overall trends of future research. However, further analysis results are given in Section 5.

4. Literature review

Financial frauds have been intensively studied by academic and industrial research due to their importance in diverse critical industries. Therefore, fraud detection has been a sensitive issue in the last years in various areas by many surveys and review articles. This includes fraud types, fraud areas, and fraud detection approaches and methods. Hence, we reviewed the latest research studies and methods to identify fraud in financial domains using data mining tools and tie the existing trends to academic researchers and financial services industries.

In addition, there are some efficient solution techniques that deal with financial fraud and can be applied to detect anomalies. One of these techniques is by employing data mining techniques that explore the underlying patterns in a big dataset and further differentiate between normal and suspicious transactions [4]. This section reviews the latest fraud detection techniques in financial domains from 2009 to 2019.

4.1. Support vector machine (SVM)

The SVM is fundamentally a classification method applied in linear classification by constructing a hyperplane as the decision plane. It works by properly classifying training sets by grouping them into valid categories [32]. Xu and Liu [33] proposed a study to identify fraudulent transactions in an online credit card using an optimized SVM model. The proposed model was examined in a dataset of a commercial bank's business to create a classification training set. The proposed model was compared with ID3+BP hybrid model. As a result, SVM outperformed ID3+BP hybrid model thereby verifying its feasibility. Rajak and Mathai [34] presented an intelligent fraudulent detection method by adopting a hybrid model using SVM and the fusion of Danger theory. The authors have found that this study performed better than previous work using the SVM model along with Danger Theory in the two aspects with respect to F-Measures and Time Complexity.

Zareapoor and Shamsolmoali [35] examined the evaluation of 3 advanced credit card fraud detection techniques. This includes KNN, NB, and SVM. The 3 techniques were evaluated with bagging ensemble classifiers. The study was conducted on a real-life credit card dataset proposed by UCSD-FICO competition. Due to the growing fraud in the bank system. The result revealed that the Bagging classifier based on three methods yielded a good evaluation performance and it is appropriate to detect credit card fraud. Gyamfi and Abdulai [36] in their paper introduced supervised learning methods using a combination of SVM to detect legal and illegal customer behavior in credit card transactions. The authors used a combination of SVM, linear regression, and logic regression to enhance the accuracy detection rate. However, the results revealed that the proposed techniques are effective in detecting banking fraud. Mareeswari and Gunasekaran [37] presented a study to identify credit card fraudulent behavior utilizing a combination of a hybrid support vector machine (HSVM), communal, and spike detection to address the drawbacks of current systems. The authors classified the transaction into two types: fraudulent and legitimate transactions depending on the prediction using a credit card dataset. Thus, the authors proved that the result of the proposed method outperformed the result of the other two existing methods.

Sundarkumar et al. [38] proposed a study for improving the identification of fraud in insurance companies using a one-class support vector machine (OCSVM) based under-sampling method. The authors employed various classification techniques including Logistic Regression, Group Method of Data Handling (GMDH), SVM, and Decision Tree (TD). The result of their study revealed the decision tree (TD) outperformed other classification techniques and decreased the complexity of the system. Francis et al. [39] studied the architecture of an automated medical bill for identifying fraud detection in medical insurance claims using SVM aiming to support experts by fast medical fraud detection. The purpose of this study is to provide a real-world speed up for medical fraud detection experts in their work. The findings improved the accuracy detection rate for detecting fraud in insurance claims.

Although intensive work has been done in discovering credit card fraud in recent years, there are some complications when using machine learning to identify fraud in credit cards including the skewed nature of the training dataset as well as the lack of fraudulent labeled transactions that may affect the detection accuracy [40]. Jeragh and AlSulaimi [41] suggested a novel unsupervised learning model for identifying fraud in credit cards by combining OSVM with an autoencoder. The authors implemented the combined model on the Kaggle credit card dataset. The authors also used measures and metrics due to the skewness in the dataset. However, the combined techniques were able to differentiate between fraudulent and normal transactions. The result of the two combined models achieved promising results, especially when evaluated using metrics performance. Deng [42] presented a model for disclosing fraudulent financial statements using SVM. The authors normalized the given dataset to ensure that the data is not affecting by different dimensions of variables and made different experiments using various variables for each. The results of the suggested model proved that the experiments' findings correspond with previous study findings pointing that issued financial statement information is intentionally falsification.

4.2. Fuzzy Logic based system

Fuzzy Logic (FL) is an effective intellectual frame for handling the problem of representing data in an environment of vagueness and inaccuracy. FL is a logic that indicates that methods of thinking are estimated and not accurate. It is ideal in human thinking patterns, especially in almost logical reasoning. The FL and Fuzzy combinations provide sophisticated methods for dealing with complicated modeling in a new and broader manner [43]. Behera and Panigrahi [44] made a study in discovering fraud in credit card fraud aiming to decrease the creation of false alarms by classifying fraud transaction and non-fraud transaction using a hybrid technique based on two approaches fuzzy c-means clustering and neural network evaluated in a synthetic dataset because there is no real credit card dataset. The authors claimed that using a combination of clustering techniques and learning mechanisms can assist to identify suspicious activities in an efficient way and reduce the creation of false alarms. The result showed that combinatorial utilize of fuzzy clustering improved detection rate with 93.90% TP and less than 6.10% FP.

HaratiNik et al. [45] introduced a Hybrid model called FUZZGY based on the fuzzy and Fogg behavioral model to detect suspicious behavior in credit card transactions based on different payment systems especially merchant frauds. A fuzzy expert system was used to observe merchant historical activities, while Fogg behavioral model was applied to describe the merchant's behavior in two different but related dimensions: motivation and ability to make a fraud. Afterward, the FUZZGY model calculated the suspicious degree of the incoming transactions. The result of this study showed better performance and acceptable results by adopting these two kinds of methods. Nezhad and Shahriari [46]

proposed a suitable solution to detect fraud in the bank system using fuzzy logic followed by the Neural-fuzzy Takagi–Sugeno training method. This study aimed to increase the accuracy of detecting fraudulent and non-fraudulent patterns of the banking transactions and to overcome the fraud targeting cash card services specifically on cash-dispenser bases. The authors defined some rules for a fuzzy logic in accordance with the experience of the experts and trained Neural-fuzzy over 350 input data to improve the detection accuracy. The result suggested applying Neural-fuzzy during the training phase will increase the detection of fraud in financial systems.

Numerous data mining techniques were proposed for predicting fraud as well as analyze the data, but some of them are not efficient for many data with a few frauds. To overcome such a problem, Supraja and Saritha [43] proposed a model by adopting Fuzzy Logic to construct fuzzy rules aiming to increase fraudulence identification. The result of the study found out that the Fuzzy Logic method is suitable for big datasets with a high performance of accurate detection rate.

4.3. Hidden Markov model (HMM)

The HMM is a dual embedded random process used to provide more complex random processes compared to a classic Markov model. The platform is assumed to be a Markov process with no noticeable states. The HMM states are unobserved, but state-dependent outputs are visible [47]. Agrawal et al. [48] suggested a model identify credit card fraud based on a case study by a combination of HMM, Behavior-based and Genetic Algorithm (GA). The proposed model consists of three steps, firstly, the authors used the HMM to maintain a log for the previous transaction. Secondly, a Behavior-based technique is used to cluster incoming transactions into several clusters based on the profiles that are categorized into lower, medium, and higher profiles. Lastly, GA is utilized to calculate the threshold value. According to the authors, the proposed model claimed to be quite beneficial for discovering fraud in credit cards. Khan et al. [49] implemented HMM to discover the behavior of the card owner via observed the incoming transactions. The authors divided the number of transactions value into low, medium, and high. However, the author applied a clustering algorithm to differentiate between legal and fraudulent patterns by utilizing data conglomeration of regions of the parameter. The experimental result of this study revealed that HMM can remove the complexity of the system.

Mhamane and Lobo [50] worked on a technique to improve Internet bank fraud detection using HMM to disclose genuine users and trace their suspicious behaviors by proposing a framework. The result of this study removes the drawback of the previous study and improves the detection accuracy. Wang et al. [51] conducted simulation experiments to increase the effectiveness in bank fraud problems. This study aims to detect online payment frauds using HMM along with K-means in real-world bank transaction data. K-means technique was applied for clustering the given dataset into several clusters and HMM was used to train with the regular behavior of the account. The result of the simulation experiments demonstrated that the study was able to discover fraud of bank transactions to a specific extent. Bhusari and Patil [52] proposed a solution to address the problem of the existing systems that suffer from detecting fraudulent activities after the transaction is done in credit card operations. This study aims to solve this problem by using HMM. The findings of this study showed that HMM improved fraud detection while reducing the false alert rate. As credit card transactions are increasing constantly, a robust and accurate system for identifying fraud in credit cards is required. Iyer et al. [53] presented a study to increase the accuracy and efficiency in revealing credit card frauds utilizing HMM. The authors also used the K-means technique for clustering purposes to find the closest centroids of the clusters and combine them into one group. The result of this study obtained high false alerts and low performance.

4.4. Artificial neural network (ANN)

ANN is a set of non-linear works similar to human thought that provides a good performance in testing large datasets. The primary component of the neural network is the neuron that is structured into layers of computing units. [54]. Srivastava et al. [55] investigated the identification of fraud in the credit card on the trader side and proposed a model relying on the neural network (NN). The proposed model connects the payment gateway with the merchant. The payment gateway is nothing less than a medium connection between the fraud detection model and merchant, holds the details of the customer's credit card. However, NN is self-trained to discover whether transactions were passed by the payment gateway are fraudulent or no. According to the authors, the system is expected to be more effective for stopping fraudulent transactions once it is implemented. Ghobadi and Rohani [56] proposed a hybrid model to detect fraud in credit cards based on Cost-Sensitive Neural Network and evaluated in real transactional data. The proposed model does not deal with imbalanced data since the given data contains a small portion of fraudulent transactions. Therefore, the authors used the Meta Cost procedure to tackle such a problem. To prove the efficiency of the suggested model, the authors compared the proposed model with the Neural Network and AIS-based Fraud Detection Model. As a result, the suggested method increased detection rate & minimized the cost of false negative.

Randhawa et al. [57] conducted a study identifying fraud in credit cards using machine learning algorithms considering that confidentiality issue on analyzing real-world credit card information is a big concern for most of the studies. The study applied a total of 12 ML algorithms ranging from standard NNs to deep learning (DL) models. Several standard algorithms were first used comprising of SVM, NB, and DL and then applied hybrid methods by employing AdaBoost and majority voting techniques. Sahin and Duman [54] applied data mining to detect fraud in credit cards using LR and ANN for enhancing the security and accuracy of credit card transactions in an automatic and efficient manner. The authors employed 13 classification techniques to construct fraud identification methods. However, the findings of this study revealed that the suggested ANN method outperformed LR method in addressing issues of credit card fraud detection. El Bouchti, Chakroun et al. [58] introduced a detailed explanation theory of Deep Reinforcement Learning (DRL) and its possible implementations of fraud detection/risk management in the banking environment. In this paper, several interesting facts about DRL are discussed and it reveals the competitive performance by DRL method. The paper is somewhat theoretical, however, and a new approach to fraud detection has emerged in front of the research community.

Lately, notable cases of fraud in financial statements have dominated the news. Ravisankar et al. [59] used data mining techniques to predict financial statement fraud on a dataset involving 202 Chinese companies of which 101 were illegally and 101 were legally companies using Genetic Programming (GP), Logistic Regression (LR), Multilayer Feed Forward Neural Network (MLFF), Group Method of Data Handling (GMDH), Support Vector Machines (SVM), and Probabilistic Neural Network (PNN). The authors applied the feature selection stage utilizing the simple t-statistic method. Among all the techniques used in this study, findings showed that PNN outperformed other techniques by showing higher accuracy while GP achieved lesser accuracies. This proves that choosing the right features can improve the performance of the detection rate.

4.5. Genetic algorithm (GA)

GA is inspired by natural evolution that searches for the best solution with a set of suggested solutions that are traditionally represented in the form of binary strings known as chromosomes [47]. Özçelik et al. [60] offered a solution to solve issues of discovering suspicious transactions in credit cards utilizing the Genetic Algorithm. The study was implemented on a real application project using real-world transactional data. The authors found all misclassification errors have a similar rate. However, the findings obtained by this study proved that the existing solution increased performance by approximately 200%. Benchaji et al. [61] suggested a new model for detecting fraud in credit cards to overcome the problem of traditional classifiers in detecting minority class objects in the imbalanced dataset using K-means and GA. The authors firstly applied K-means for grouping and classifying minority instances then used the genetic algorithm in each cluster to produce the new instances to gain a new training dataset. The result of this study showed that applying k-means and genetic algorithms enhanced credit card fraud discovery and decrease the number of false alerts. Liang and Lv [62] designed a technique to improve the effectiveness and accuracy of detecting fraud in financial statements relying on the Fuzzy Genetic Algorithm BPN (FGABPN) using a dataset proposed by GTA Research Service Center. The empirical findings of the study showed that the FGABPN method has improved accuracy with a high detection rate.

4.6. K-Nearest Neighbor algorithm (KNN)

KNN technique is a type of non-parametric method used for classification and regression. KNN technique is a type of non-parametric method used for classification and regression. Its purpose is to find the closest neighborhoods based on similarity from a given dataset and generate a new sample point based on the distance measure between two data samples. However, one of the most well-known methods to compute the distance is the Euclidean distance [63,64]. Malini and Pushpa [65] analyzed credit card fraud utilizing two techniques: the outlier detection method and KNN algorithm. The authors studied the implementation of the performance result of both techniques through the credit card approval system. The empirical findings proved that the KNN technique is accurate and effective for detecting credit card fraud. Heryadi et al. [66] presented a study on debit card fraud transaction recognition to address the problem of detecting anomaly fraudulent transactions of the previous study that used the HMM model. The authors used a combination of Chi-Square Automatic Interaction Detection (CHAID) and k-Nearest Neighbor to improve the detection accuracy of anomaly transactions. The finding of the study showed that the CHAID method improved the performance for debit card fraud transaction recognition. An approach to detecting fraud in Auto Insurance is proposed by Badriyah et al. [67] using Nearest Neighbor-based Method that includes three methods density-based, distance-based, and interquartile range in car insurance data. The study focuses on looking at the influence of feature selection processes on accuracy scores. However, the results obtained from this study showed that there is a noticeable impact of improving the detection accuracy when using feature selection methods. Awoyemi et al. [68] investigated the performance of 3 machine learning models implemented on credit card transactions to identify fraudulent behavior. The authors performed a comparative analysis using LR, NB, and K-NN trained on a credit card dataset proposed by European cardholders comprising 284,807 transactions. The finding of the comparative analysis demonstrates that K-Nearest Neighbor outperformed LR and NB techniques.

4.7. Bayesian network (BN)

BN is a form of graphical model that works in conditional dependencies and independent relationships between different variables. The essential concept of BN is in the form of nodes and edges in a directed graph. BN is helpful for searching anonymous probabilities computations specific known probabilities in the existence of uncertainty [47]. Deng [69] studied published financial data and design a model to facilitate the task of identifying fraudulent financial statements (FFS) by adopting NB classifier. The aim of this study was to improve classification accuracies. The experiment was performed in a 44 FFS dataset containing normal and abnormal financial statements. The result of the study discovered that released financial statement information includes falsification indicators. Detecting the anomalous behavior of physicians in the healthcare industry is a difficult task especially if they are behaving outside of their specialty. A machine learning model has been proposed by Bauder, Khoshgoftaar, Richter and Herland [70] to address such a problem based on the physicians' medical procedure history using a multinomial Naïve Bayes algorithm. The study applied in a dataset containing the number of physicians and procedures provided by Medicaid Services The goal of the study is to anticipate supplier kind. If the anticipated supplier kind is not met the expected behavior, thus the supplier can be classified as anomalous.

Furthermore, in their subsequent study, Herland et al. [71] expanded upon their previous work into medical specialty anomaly detection to increase fraud detection capabilities using Multinomial Naïve Bayes. The authors tested and validated the efficiency of their method using real-world fraud cases including List of Excluded Individuals dataset. The authors proposed three strategies namely specialist medical grouping, feature selection, and removal of given overlapping specialties to improve model performance. The authors also confirmed that their model was able to detect 67% of physicians contained in the dataset as fraudulent. Hajek and Henriques [72] suggested an intelligent detection focusing on fraudulent financial statements in non-fraudulent companies by collecting particular features obtained from financial reports to perform fraud classification using 14 machine learning techniques including decision trees (JRip, Simple CART, logistic model trees, and C4.5), support vector machines, Bayesian classifiers (Naïve Bayes hybrid classifier or Decision Table, BBN, and Naïve Bayes (NB)), logistic regression, AdaboostM1, ensemble classifiers (AdaboostM1, Random Forests. and, Bagging), and neural networks (MLP and voted perceptron). In the experiment, the authors considered several variables of both the financial and linguistic of financial statements in the management discussion. The result showed that BBN outperformed other machine learning methods.

4.8. Decision tree (DT)

DT is a classification or regression tree utilized to create a decision support tool in the trees of an inner node that represents binary options over the features and dependencies that represent the result of this selection. The decision tree model is divided into subgroups until ultimately separated into an exclusive mutual subset [16]. Kho and Vea [73] studied the transaction behavior of credit cardholders to distinguish between anomalous and legitimate transactions. They used various classification techniques including NB, Bayes-Net, J48, RT, libSVM, and MOLEM to determine which technique is likely accurate. The study was implemented using two synthetic datasets. In the evaluation process, it was found that out of the five classification techniques Random Tree (RT) and J48 were at the two top classifiers. However, Random Tree achieved the best accuracy than J48 with an

accuracy rate of 93.50% and 94.32%. Devi and Kavitha [74] proposed a supervised approach to categorize credit card transaction data as normal or suspicious by comparing various classification methods such as SVM, RF, and DT. The three classification methods were evaluated using different accuracy metrics rates and measures. The result showed that DT outperformed other classification methods with high accuracy. In the traditional approach of insurance fraud, when a person submits false insurance claims to gain benefits, the system cannot classify them as insurance fraud. Hence, Roy and George [75] focused on detecting auto or vehicle fraud by using a machine learning technique. The authors compare three different classification algorithms which are DT, RF, and NB to improve the accuracy and precision of insurance fraud using raw data containing a set of attributes. The result proved that decision tree and RF methods outperformed Naïve Bayes (NB).

The problem of imbalanced data has affected the detection accuracy for a long time. Subudhi and Panigrahi [76] introduced a novel method to identify fraud in the auto insurance sector using an adaptive oversampling technique to delete the imbalance classes from the insurance dataset. The authors used three supervised classifiers such as Multi-Layer Perceptron, SVM, and DT on the training dataset to create their relevant trained classifier models to separate the anomalous classes from the normal ones. The authors also employed the 10-fold cross-validation method to get a superior performing classifier model. The overall result obtained a high fraud detection rate using SVM and Decision Tree classifiers.

4.9. Logistic Regression (LR)

LR is a method utilized to analyze datasets based on a linear model statistically. It works by performing regression on a collection of variables. It is generally an applicable method to describe patterns and to clarify relationships between single or multiple dependent binary variables [16]. Yue et al. [77] examined published data and proposed an appropriate method identifying factors related to Fraudulent Financial Statements using Logistic Regression in a Chinese dataset involving 174 companies. The authors selected 21 ratios as potential predictors for financial fraud statements to identify the best parameters of the logistic regression method by performing a set of testing. The authors compared the predictive ability of the suggested method against other detecting methods. The results of the developed method demonstrate that the predictive ability is higher than other models at about 10%.

4.10. Outliers detection

The term outlier is known as an anomaly that is used to express the existence of strange behavior. Outlier detection is considered as a branch of data mining which is the most popular method utilized to identify data objects that do not behave as expected in accordance with general behavior or data model [78]. Anomaly detection is the problem of uncovering data patterns in datasets that differently behave from the norm. The anomaly detection is a binary classification problem, this includes two types of labels: abnormal and normal. The importance of anomaly detection relies on the fact that unexpected behaviors in data translate to important information in a wide diversity of application domains [79].

The problems of detecting anomalous patients lie in the inefficiency of the traditional fraud detection techniques in identifying suspicious behavior. Thus, Peng et al. [80] developed an outlier-based fraudulent detection method by analyzing the correlation of patients to detect fraud in large medical insurance data. The authors divided the proposed method into 4 steps. The first step includes data preprocessing that used to clean and cluster the data. In the second step, extracting informative features from the dataset, this includes medicines, diseases, and patients to build a heterogeneous information network. After this, the calculation of the correlation scores of different patients by analyzing the information network. In the last step, a discrimination rule is created to differentiate between genuine and suspicious patients. The result of this study showed that the suggested technique was efficient and accurate from various angles.

The problem of an efficient and accurate technique for automobile insurance fraud detection has not been solved where most of the techniques suffer from time complexity and fraud detection rate. Yan and Li [81] developed a technique based on the NN to detect fraud on auto insurance claims applied the rules of association for mining the auto insurance fraud law that improved the identification model to detect fraud in auto insurance. The findings proved that the developed technique reduced the time complexity, high accuracy, increased the detection rate, and decreased the impact on the K value. Zhang and He [82] analyzed the problems in the medical insurance system and suggested an outlier detection framework for the detection of fraud in Medicare insurance. The authors applied the suggested method in 3 real medical insurance datasets. However, the proposed method comprises two parts. The first part was an improved algorithm of the simple Local Outlier Factor (LOF) known as (imLOF) that relying on a spatial density-based algorithm, which is more efficient when using medical insurance dataset. The second part was a robust regression to describe the relationship between the dependencies and independencies of the variables. The finding of the experiments revealed that the suggested method is effective.

Monamo et al. [96] explored the adoption of Trimmed K-Means clustering for discovering suspicious activities in the Bitcoin network based on pattern recognition. The authors adopted K-means algorithm to group the data sample into several classes with the purpose to raise the capability of detecting. The authors implemented the purposed methods on a real bitcoin dataset including all transactions up to April 2013. Therefore, the result of this study has proved the ability to detect 5 suspicious transactions out of the 30 reported theft transactions.

In their subsequent study, Monamo et al. [97] proposed a multifaceted approach by adopting an unsupervised learning algorithm using Trimmed K-Means and kd-trees. Local Outliers are those instances based on distances from the neighborhood established by kd-trees while Global Outliers are those instances far from the nearest centroids using Trimmed K-Means Clustering. The two approaches were further investigated using 3 classification algorithms as boosted binary regression models, random forests, and maximum likelihood-based. The two approaches were tested based on 30 known fraudulent cases. Due to the computational resource requirements, both approaches were implemented differently in terms of the number of instances. In the first approach, the dataset was only restricted to the first 1 million instances implemented by trimmed K-Means while in the second approach, a full dataset was implemented by kd-trees. The authors used a classification-based approach to evaluate the superiority of the techniques regarding performance according to few known criminal incidents. Out of the three classification models, random forest performed better than other models with 8 features regardless of class imbalances.

4.11. Other classification techniques

Regarding the problem of imbalanced data, Mishra and Ghorpade [40] implemented various classification and ensemble methods to identify suspicious transactions in credit cards. The techniques include Gradient Boosted Tree, RF, LR, Stacked Classifier,

Table 2
Summarized classification for Support Vector Machine in the financial fraud domain (2009–2019).

No.	Ref.	Fraud type	Technique used	Dataset used	Validation methods	Results
1	[42]	Financial statement fraud	Support Vector Machine	73 FFS, 44 FFS, and 99 non-FFS testing dataset during 1999–2002	N/A	The results of the suggested model correspond with previous study findings pointing that issued financial statement information is intentionally falsified.
2	[33]	Credit card fraud	Support Vector Machine	Commercial bank's business database	N/A	The proposed model was compared with ID3+BP hybrid model. As a result, SVM outperformed ID3+BP hybrid model thereby verifying its feasibility.
3	[36]	Credit card fraud	Support Vector Machines with Spark (SVM-S)	German & Australian dataset of credit cards	linear regression, and logic regression	The result shows an improvement of SVM method by 80% which is considered a better performance comparing with similar works.
4	[41]	Credit card fraud	One-class support vectors machine (OSVM) and an autoencoder	Kaggle Credit Card dataset	Measures and metrics rate Geometric mean (GMean)	The result of the two combined models achieved promising results, especially when evaluated using metrics performance.
5	[38]	Auto insurance fraud	One Class support vector machine and Decision Tree	Angoss KnowledgeSeeker software	Accuracy rate	The result revealed that the Decision Tree outperformed other classification methods and decreased the complexity of the system.
6	[39]	Health insurance fraud	Support Vector Machine	N/A	N/A	The findings improved the accuracy detection rate for detecting fraud in insurance claims.
7	[37]	Credit card fraud	Hybrid support vector machine (HSVM)	BL Database	N/A	The authors demonstrated that the findings of the proposed model outperformed the findings of the other two existing methods and achieved good performance.
8	[83]	Financial statement fraud	Support vector machine and Logistic Regression	Dataset of 112 Chinese listed companies	N/A	The results of our study proved to be beneficial in terms of the prediction accuracy where SVM achieved an accuracy rate of 86.612% while Logistic Regression achieved an accuracy rate of 83.036%.
9	[76]	Auto insurance fraud	An adaptive oversampling technique (ADASYN) SVM, and DT	Insurance dataset known as "carclaims.txt"	Accuracy, Sensitivity and Specificity	The result obtained from this study showed that SVM achieved Sensitivity = 94.74%, and DT achieved Sensitivity = 94.52%.
10	[84]	Health insurance fraud	Support vector machine	Database of a Turkish insurance company	N/A	The study was able to identify 6595 claims estimated between 50.0% to 67.3%.
11	[34]	Credit card fraud	SVM and the fusion of Danger theory	A dataset generated by UCI repository containing credit card transactions	F-Measures and Time Complexity	The study performed better than previous work using the SVM model along with Danger Theory in the two aspects with respect to F-Measures and Time Complexity.

and SVM. The problem with standard algorithms is the detection of majority classes that contain only legitimate transactions which ignore the minority classes having fraudulent transactions. The authors tackle this problem by oversampling and undersampling the data which in fact replicating the samples in the minority class, so the data gets well-balanced. Furthermore, Li et al. [106] suggested a model for addressing the problem of unbalanced data and large data in automobile insurance data by using Random Forest (RF) Algorithm. In this study, RF algorithm decreased the filtering of variables and automatically identified the most important input variables when implemented on a large dataset containing a large number of related variables. The result concluded from this study proved that random forest has better accuracy and robustness and can be suitable for large datasets and unbalanced data.

Saia and Carta [111] investigated the benefits of two novel proactive strategies using Fourier transform and Wavelet transform in order to move the data into a new domain. The authors compared their proactive approach to Random Forests method with k-fold cross-validation and then conducted an experiment to test the performance with ten data mining algorithms. The results confirmed that Random Forests is the best performing approach.

Table 3
Summarized classification for Fuzzy Logic in the financial fraud domain (2009–2019).

No.	Ref.	Fraud type	Technique used	Dataset used	Validation methods	Results
1	[44]	Credit card fraud	Fuzzy Clustering and Neural Network	Synthetic dataset	TP/Sensitivity, TN/Specificity & FP	The result showed that combinatorial utilize of fuzzy clustering improved detection rate with 93.90% TP and less than 6.10% FP.
2	[46]	Internet bank fraud	Fuzzy logic and Neural-fuzzy	N/A	N/A	The authors suggested applying Neural-fuzzy training for detecting fraud in financial systems.
3	[45]	Credit card fraud	Fuzzy model with the Fogg behavioral model	Synthetic dataset	False Positive, Negative Positive, & False Negative	The result of this study showed better performance and acceptable results by adopting these two kinds of methods.
4	[62]	Financial statement fraud	Fuzzy Genetic Algorithm BPN.	A dataset provided by GTA Research Service Center	N/A	The findings of this study showed that the FGABPN method scored high detection accuracy.
5	[43]	Auto insurance fraud	Fuzzy Logic	Synthetic datasets	N/A	The result found that the FL method achieved good accuracy in discovering insurance claims and decreased time-consuming for large datasets with high dimensional.

Table 4
Summarized classification for neural network and genetic algorithm in the financial fraud domain (2009–2019).

No.	Ref.	Fraud type	Technique used	Dataset used	Validation methods	Results
1	[85]	Financial statement fraud	SVM & artificial neural network (ANN)	Synthetic dataset	Feature Selection	The finding proved that feature selection helped in increasing the accuracy of SVM method with 88.37% while ANN gave the most effective accuracy with 90.97%.
2	[56]	Credit card fraud	Cost Sensitive Neural Network	The real dataset provided by a big Brazilian company	Metrics rate TP, FP, NP, and FN	The suggested method increased the detection rate & minimized the cost of FN.
3	[55]	Credit card fraud	Neural Networks	N/A	N/A	According to the authors, the system is expected to be more effective for stopping fraudulent transactions once it is implemented.
4	[54]	Credit card fraud	Artificial Neural Network and LR	Four different datasets	N/A	The findings showed that ANN outperformed LR in addressing the issues of detecting credit card fraud.
5	[86]	Health insurance fraud	Neural Network and Pharmacopoeia Spec-trum Tree	Medical records database of China.	N/A	The experiment results proved neural network improved the accuracy in healthcare fraud detection up 86%.
6	[60]	Credit card fraud	Genetic Algorithm	Real life dataset	N/A	The result proved that the existing solution increased performance by approximately 200%.
7	[87]	Credit card fraud	Genetic Algorithm & scatter search	Real dataset contained 1050 fraud transactions	N/A	The findings have improved the performance of the existing solution by 200%

Bauder and Khoshgoftaar [112] explored several machine learning methods including supervised (Deep Neural Network, GBM, RF, and NB), unsupervised (LOF, Mahalanobis-distance, KNN, and autoencoder), and hybrid (Bayesian probability and multivariate-regression) to detect more fraudulent events in Medicare fraud and compared performance results and statistical significance. The authors also conducted a comparative study among different machine learning approaches by adopting several performance metrics. The study applied on the 2015 Medicare data containing suspicious records provided by List of Excluded Individuals dataset. The findings of the experiments showed that supervised learner's performance outperformed unsupervised and hybrid learners.

Kowshalya and Nandhini [107] presented a study using data mining techniques to identify abnormal claims in Automobile Insurance by employing three classification techniques like RF, NB, and J48. The authors executed their study on Weka 3.8 tool using a generated synthetic dataset due to the lack of real insurance claims datasets. The study was evaluated based on several performance metrics and the result obtained from the experiment revealed that RF outperformed the other two methods on the Insurance claim dataset while NB achieved the best performance on the three test options. On the other hand, Carta, Fenu et al. [113] suggested a novel data intelligence technique to maximize the effectiveness of fraud detection and addressed the data imbalance problem that occurred while analyzing the credit card transaction dataset. The study was based on a Prudential Multiple Consensus model that combines the effectiveness of several modern classification algorithms such as complete agreement, the majority voting, and weighted voting. The results of their study proved its effectiveness in terms of sensitivity and Area Under the ROC Curve (AUC). Singh and Jain [114] studied and compared the performance of 5 machine learning techniques used for the detection of credit card frauds. Results obtained from this study revealed that J48 and Projective Adaptive Resonance Theory (PART) techniques have scored the best accuracy while

Table 5
Summarized classification for hidden Markov model and decision tree in the financial fraud domain (2009–2019)..

No.	Ref.	Fraud type	Technique used	Dataset used	Validation methods	Results
1	[88]	Credit card fraud	Hidden Markov Model, Behavior based and Genetic Algorithm	N/A	N/A	The authors recommended that the suggested method can be quite beneficial in identifying credit card fraud.
2	[49]	Credit card fraud	Hidden Markov Model	Synthetic dataset	N/A	The experimental result of this study revealed that HMM can remove the complexity of system.
3	[50]	Internet bank fraud	Hidden Markov Model	Synthetic dataset	N/A	The result of this study removes the drawback of the previous study and improves the detection accuracy.
4	[48]	Credit card fraud	HMM, Behavior based and GA	N/A	N/A	According to the authors, the proposed model claimed to be quite beneficial for discovering fraud in credit cards.
5	[51]	Internet bank fraud	Hidden Markov Model and K-means	Real-world bank transaction data	TP, FP, Recall, and Precision	The result of the simulation experiments demonstrated that this study could reveal fraudulent transactions of a bank to a specific extent.
6	[52]	Credit card fraud	Hidden Markov model (HMM)	N/A	N/A	The findings of this study showed that HMM scored a high fraud coverage where 84% transactions are a legitimate and very low false alarm which is about 7.
7	[53]	Credit card fraud	Hidden Markov Model	3 Synthetic datasets	N/A	The finding of this study was not as excepted which produced high false alarms with low performance.
8	[89]	Credit card fraud	Hidden Markov Models	Real transaction data	N/A	The proposed method proved to be effective and contained good F-score.
9	[74]	Credit card fraud	Various classification methods like SVM, RF, and Decision Tree.	Synthetic dataset	Metric rate, Sensitivity & Specificity, F1-score,	The result of this study revealed that decision tree classification algorithm is appeared the best accuracy compared to other classification algorithms.
10	[90]	Credit card fraud	Cost-sensitive decision tree	Credit card dataset for 12 months	Saved Loss Rate (SLR)	The findings proved that cost-sensitive decision tree outperformed DT ANN and SVM models.

precision and sensitivity of J48, AdaBoost, and the random forest have increased.

Therefore, due to its high importance, much attention has been given to financial fraud detection techniques in previous research. In this paper, we collect different ways to discover suspicious transactions using data mining techniques. Tables 2–9 show all the examined and analyzed papers with a total of 75 articles. Table 2 summarizes the research work on Support Vector Machine (SVM) for financial fraud detection with a total of 11 articles. In Table 3, research work on fuzzy logic for financial fraud detection with a total number of 5 articles has been thoroughly discussed. Table 4 highlights the research work done on neural networks and genetic algorithms for financial fraud detection with a total of 7 articles. Table 5 gives a summary of the research work on hidden Markov model and decision tree for financial fraud detection with a total of 10 articles. Table 6 presents the research work outlier detection, clustering, and unsupervised learning techniques for financial fraud detection with a total of 20 articles. Table 7 provides a detailed description of the research work on Naïve Bayes for financial fraud detection with a total of 8 articles. In Table 8, a discussion of the research work on the random forest has been done and Table 9 explains the research on logistic regression for financial fraud detection with a total of 8 and 6 articles, respectively. Each article was analyzed and examined based on several criteria including fraud type, techniques used, datasets used, validation methods, and results. However, based on the analysis criteria, we can define the methods that are regularly applied to detect financial fraud and disclose the best technique well-performed across types of fraud.

5. Results

This section summarizes and outlines the findings of our review to draw a conclusion including the latest data mining methods applied in detecting financial fraud found in our review. It also provides a classification associated with frequent use, description, pros, cons, and business application. Section 5.1 provides a distribution of papers by journal and conference title. Classification of financial fraud types has been discussed in Section 5.2. A detailed description of the pros and cons of the data mining techniques found in our review has been presented in Section 5.3. Distribution of paper by data mining techniques and a classification based on their fraud types is provided in Section 5.4. Finally, Section 5.5 gives a distribution of examined papers by publication year.

5.1. Distribution of papers by journal and conference title

Table 10 lists the distribution of sixty-three papers found in our analysis for both conferences and journals. This includes the journal or conference title, frequency (paper count), and percentage (the percentage number of the overall papers related to a journal or a conference). The Journal of Expert Systems with Applications contained the most relevant articles (5.3%, or 4 of the 75 papers), followed by the International Conference on

Table 6
Summarized classification for outlier detection, clustering, and an unsupervised learning techniques in the financial fraud domain (2009–2019).

No.	Ref.	Fraud type	Technique used	Dataset used	Validation methods	Results
1	[65]	Credit card fraud	Outlier detection method and K-Nearest Neighbor	N/A	N/A	The empirical findings proved that KNN technique is accurate and effective for detecting credit card fraud.
2	[80]	Health insurance fraud	Outlier Detection	Real health insurance dataset	N/A	The result of this study showed that the suggested technique was efficient and accurate from various angles.
3	[91]	Health insurance fraud	Isolation Forest, Unsupervised Random Forest, Local Outlier Factor, autoencoders, and k-Nearest Neighbor	Medicare Part B big dataset	Accuracy, Sensitivity, and Specificity TPR, FPR	The findings revealed that LOF method achieved the best result and Autoencoders, KNN, and 5 neighbors were the worst at detecting Medicare.
4	[92]	Health insurance fraud	Pairwise comparison Module, Policy Verification Module, and Outlier detection Module	Health insurance dataset	Confusion matrix for accuracy	The result of this study proved that using an effective method helped to identify suspicious behavior and improved the accuracy detection rate.
5	[93]	Health insurance fraud	Outlier Detection	Dental claims data	N/A	The result of the case study concluded that new suspicious transactions can be detected through outlier detection and might be employed in future automated detection mechanisms.
6	[81]	Auto insurance fraud	Outlier detection method based on the nearest neighbor	Synthetic dataset	Experimental analysis	The findings proved that the improved algorithm reduced the time complexity, high accuracy, increased the detection rate, and decreased the impact on the K value.
7	[82]	Health insurance fraud	improved local outlier factor (imLOF) and robust regression	3 datasets provided by the medical insurance company	Performance metrics	The result of the experiments revealed that the suggested method is effective.
8	[94]	Health insurance fraud	PageRank-based algorithm	Medicare-B dataset	N/A	The result obtained from the proposed algorithm showed that tens of anomalies in the dataset have been successfully identified.
9	[67]	Auto insurance fraud	Nearest Neighbor based Method	German car insurance data	Accuracy F-Measure	The result of this study proved that using feature selection methods increased the detection accuracy.
10	[95]	Mobile healthcare fraud	SSIsomap activity clustering method, SimLOF	Real-world dataset provided by Dareway Medical Insurance	Performance metrics	The findings showed that the suggested method was approximately 50% more effective than the existing method, it also reduced the time complexity.
11	[96]	Cryptocurrency fraud	Trimmed K-Means and K-Means	A publicly available data of Bitcoin up to 7 April 2013	N/A	The proposed method achieved a detection rate up to 16.67%.
12	[97]	Cryptocurrency fraud	Trimmed k-means Kd-trees, Random forest	A publicly available data of Bitcoin up to 7 April 2013	N/A	The result of the proposed method detected 20% known thefts using Trimmed K-Means while 23.33% known thefts using kd-trees.
13	[98]	Credit card fraud	Artificial Neural Networks & Cluster Analysis	Real credit card data	Mean Squared Error (MSE) & (FPR)	The result of this study showed that the application of Cluster Analysis achieved a good result at qualitative data normalization.
14	[99]	Health insurance fraud	Evolving Clustering Method (ECM) and SVM	Insurance claims dataset	N/A	The result of this study revealed a better performance in detecting unknown health insurance claims.
15	[61]	Credit card fraud	K-means Clustering and GA	N/A	N/A	The result of this study showed that applying k-means and genetic algorithms enhanced credit card fraud discovery and decrease the number of false alerts.

(*continued on next page*)

Table 6 (*continued*).

No.	Ref.	Fraud type	Technique used	Dataset used	Validation methods	Results
16	[100]	Auto insurance fraud	Statistical Decision rules and k-means clustering	Real world medical dataset	N/A	The findings indicate that rule-based mining is effective in discovering insurance claims fraud.
17	[101]	Financial statement fraud	Growing hierarchical self-organizing map (GHSOM)	116 firms 1992–2006	Test sample containing fraud and non-fraud	The findings proved to be promising with respect to other classifications techniques which are NN, BPNN, SVM, SOM+LDA, and GHSOM+LDA method.
18	[102]	Financial statement fraud	Self-organizing map (SOM), K-means clustering, and SVM	A dataset of 100 organizations in China during 1999–2006.	Accuracy rate	The result obtained from this experiment showed that the combination of the V-KSOM achieved a better result than using SOM alone.
19	[66]	Credit card fraud	Chi-Square Automatic Interaction Detection (CHAID) and k-Nearest Neighbor	Dataset of a private bank in Indonesia	Feature Selection	The findings of this study showed an improvement in detection accuracy of anomaly transactions where CHAID model scored accuracy = 8.0 while K-NN model accuracy is = 0.7
20	[103]	Health insurance fraud	Resolution Based (RB) algorithm and density factor	Insurance company dataset	N/A	The result of the experiments proved that using a combination of RB and DF algorithms is more accurate and achieved higher precision.

Table 7
Summarized classification for Naïve Bayes in the financial fraud domain (2009–2019).

No.	Ref.	Fraud type	Technique used	Dataset used	Validation methods	Results
1	[104]	Credit card fraud	Bayesian Network Classifier (BNC)	A dataset provided by PagSeguro, in Brazil	F1 metric	The results obtained from this study proved that Fraud-BNC developed the effectiveness of the existing company's economy is up to 72.64%.
2	[72]	Financial statement fraud	DT, SVM, Bayesian classifiers, LR, ensemble classifiers, and neural networks	A dataset provided by Securities and Exchange Commission (SEC)	Performance Metrics including Accuracy, TP rate, TN rate, FP and FN, F-measure	The result showed that Bayesian belief networks outperformed other machine learning methods.
3	[71]	Health insurance fraud	Multinomial Naïve Bayes	Dataset provided by office clinics in Florida	Confusion matrices including Recall, Precision and F-Score	The result of this study confirmed that the proposed method detected 67% of physicians included in LEIE dataset as fraudulent.
4	[105]	Internet bank fraud	Bayesian classification and association rule	Real data of a bank in Taiwan containing fraudulent & normal accounts	N/A	The proposed method achieved a good result in identifying can fraudulent accounts.
5	[70]	Health insurance fraud	multinomial Naïve Bayes algorithm	A Medicare dataset proposed by Medicaid Services	Performance Metrics with 5-fold cross-validation	The result showed that this study was good for several specialties and could classify physicians who are likely misusing insurance systems.
6	[57]	Credit card fraud	Naïve Bayes, Support Vector Machine, and Deep Learning	Credit card dataset produced by Malaysian financial institute	AdaBoost & majority voting methods	The performance result of this study proves a good accuracy in discovering credit card fraud by adopting the majority voting technique.
7	[35]	Credit card fraud	K-Nearest Neighbor algorithms, Naïve Bayes, and Support Vector Machines	Real world credit card dataset generated by UCSD-FICO	Fraud Catching Rate, False Alarm Rate, Balanced Classification Rate and Matthews Correlation Coefficient	The result revealed that Bagging classifier based on three methods yielded a good evaluation performance and it is appropriate to detect credit card fraud.
8	[69]	Financial statement fraud	Naïve Bayes Classifier	73 FFS testing dataset during 1999-2002 – 44 FFS and 99 non-FFS testing dataset during 1999–2002	N/A	The result of this study obtained a good performance and found out that released financial statement information includes falsification indicators.

Table 8
Summarized classification for Random Forest in the financial fraud domain (2009–2019).

No.	Ref.	Fraud type	Technique used	Dataset used	Validation methods	Results
1	[40]	Credit card fraud	Gradient Boosted Tree, Random Forest, Logistic Regression, Stacked Classifier, and Support Vector Machine	A dataset provided by European credit cards holders	Metrics rate which includes TPR, FPR based on Precision and Recall	The result of the proposed study showed Random Forest outperformed at classification than other classification techniques. With a recall of 96% and precision of 6%.
2	[73]	Credit card fraud	Naive Bayes, Bayes Net, J48, Random Tree, libSVM, and MOLEM	Synthetic dataset	Measurements such as Precision (average) and Recall (average)	The result showed that out of the five classification techniques, Random Tree (RT) and J48 were the two top classifiers. It was found that RT scored the best accuracy rate over J48 with an accuracy rate that reached 93.50% and 94.32%.
3	[106]	Auto insurance fraud	Random Forest Model	Real data of an automobile insurance company	N/A	The findings of this study showed that the random forest has better accuracy and robustness and can be appropriate for large datasets as well as imbalanced data.
4	[32]	Credit card fraud	Random forests, support vector machines, and logistic regression	Real-life dataset on credit card	TPR, Accuracy, Sensitivity, Recall, Specificity, F-measure, Precision, and G-mean.	The results proved that the RF showed outperformed other methods.
5	[107]	Auto insurance fraud	Classification algorithms such as Naïve Bayes, J48, and Random forest	Synthetic dataset	Accuracy, Precision, and Recall	The result obtained from the study showed that RF outperformed the other two methods on the Insurance claim dataset.
6	[108]	Cryptocurrency fraud	Bayes Network, Random Forest, and Repeated incremental pruning to produce error reduction	Synthetic dataset	True Positive Rate (TPR), Accuracy, Recall, Specificity, F-measure, Precision, G-mean, and AUC	The result shows that Random Forest achieved higher accuracy than other classifiers.
7	[109]	Credit card fraud	Logistic regression, Decision tree, random forest.	German credit card fraud dataset.	Precession, recall, accuracy, and F1 score.	Out of the three models, RF and DT achieved great results with high accuracy, recall, and precision
8	[110]	Health fraud	Random Forest models	A dataset of Physician and a few Supplier Data calendar	k-fold cross-validation	The findings proved that the best fraud detection performance group distribution is 90:10 while the poorest group distribution is 99.9:0.01.

Computing for Sustainable Global Development (INDIACom) (4%, or 3 papers).

To determine which country is exposed to financial crime, Table 11 presents a list of the most exposed countries to financial fraud, respectively. However, most of the studies were conducted in India followed by China, the United States, and Turkey.

Fig. 3 classifies and provides an analysis of financial fraud types used by data mining techniques. It is evident that the papers found in credit card fraud constitute the largest portion with a total of 32 papers followed by health insurance fraud with 16 papers. This number corresponds to 48 papers out of the 75 reviewed papers. However, some fraud types are not listed in the figure such as mortgage, securities and commodities, and money-laundering fraud because most of the methods examined in this review did not apply to these issues.

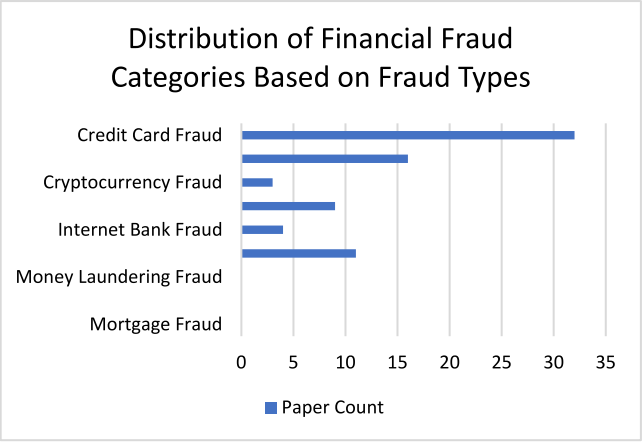

Fig. 3. Classification of fraud types used by data mining.

5.2. Classification of financial fraud types

This subsection demonstrates the analysis of the papers reviewed in financial fraud areas. Based on the results of our analysis, we categorized the types of financial fraud into four groups which are bank, financial statement, insurance, and cryptocurrency fraud (Table 12). Table 12 explains the type of fraud and the number of reviewed papers found. Referring to Table 12, it is apparent that most papers are extensively applied to bank fraud and insurance fraud with a total of 61 papers out of 75 reviewed papers which constitute the largest portion equal to 81.33% of the overall papers.

Table 9
Summarized classification for Logistic Regression in the financial fraud domain (2009–2019).

No.	Ref.	Fraud type	Technique used	Dataset used	Validation methods	Results
1	[68]	Credit card fraud	Naïve Bayes, K-Nearest Neighbor Logistic Regression	A real-life credit card dataset proposed by European cardholders	Matthews Correlation Coefficient, specificity, accuracy, sensitivity, Metric rate	The finding of the comparative analysis demonstrates that K-Nearest Neighbor outperformed LR and NB techniques.
2	[77]	Financial statement fraud	Logistic Regression	A dataset of 174 listed companies in China	F-values and p-values	The results of the suggested method proved that the predictive ability is higher than other models at about 10%.
3	[75]	Auto insurance fraud	Machine learning techniques including DT, RF, and NB	Raw dataset containing a set of attributes	N/A	The findings of the comparison of classification algorithm proved that DT and RF algorithms performed better than NB.
4	[59]	Financial statement fraud	Logistic Regression, Group Method of Data Handling, SVM, and Probabilistic Neural Network	A dataset involving 202 Chinese companies	Ten-fold cross-validation	Findings based on AUC showed that PNN outperformed other techniques followed by GP that achieved mostly fewer accuracies.
5	[112]	Health fraud	supervised, unsupervised, and hybrid algorithms	2015 Medicare PUF dataset provided by Medicaid Services	Performance metrics including F-measure, and G-measure	The results of the experiments showed that supervised learners performance outperformed the unsupervised and hybrid learners.
6	[115]	Financial fraud	Logistic Regression, Decision Tree, and many more	credit cards, insurance, and financial statement.	True Positive Rate (TPR), Accuracy, and Sensitivity.	The result proved that hybrid fraud detection techniques outperformed several traditional detection methods.

5.3. Pros and cons of data mining techniques

As stated earlier, the primary aim of this study is to comprehensively review the latest data mining used in detecting fraud in financial domains. However, this subsection highlights the content of our review of most DM techniques utilized in the financial field. We classified our findings based on their frequency usage. Out of 34 data mining techniques used in this review, we summarized the most applied ones from 2009 to 2019. As a result, we discovered that the SVM technique is the most widely financial fraud detection technique used in data mining with a 23%, followed by both of Naïve Bayes and Random Forest, with a 15%. While the neural network is represented by a 13.5 and logistic regression, with a 12.15. This indicates that the SVM technique is the leading DM technique used in fraud detection of the financial domains. Table 13 gives a summary of data mining techniques and their pros, cons, and business application. Based on the table below, the supervised learning techniques (e.g., SVM, NB, RF, NN, and DT) have been widely used more than unsupervised techniques. Therefore, supervised learning approaches were found to show better performance methods than unsupervised approaches due to the existence of labeled data whereas unsupervised techniques suffer from a lack of labeled data in identifying fraud in financial areas.

5.4. Distribution of papers by data mining techniques

To define the most widely used algorithms in detecting financial fraud, we classify and provide detailed analysis by showing the frequency of subcategories of types of financial fraud in Table 14. Table 14 consists of four main groups which are the bank, insurance, financial, and Cryptocurrency fraud. Bank fraud is divided into groups namely credit card and Internet bank fraud whereas insurance fraud is divided into two groups namely auto and healthcare fraud. Our results reveal that thirty-four data mining techniques have been used for identifying fraud in financial domains. However, as shown in Table 14, SVM method found to be the most frequent technique used in financial fraud detection followed by outlier fraud detection techniques. This indicates that the two most common methods used to discover fraud in financial areas, while fraud in credit card and insurance appear to be the most prominent types of fraud. On the other hand, we can clearly see that money laundering, stock, commodity, and mortgage fraud were not included in our review for some reason. one of the reasons is the difficulties of gathering this data or maybe revealing the results that are related to sensitive subjects may not be permitted.

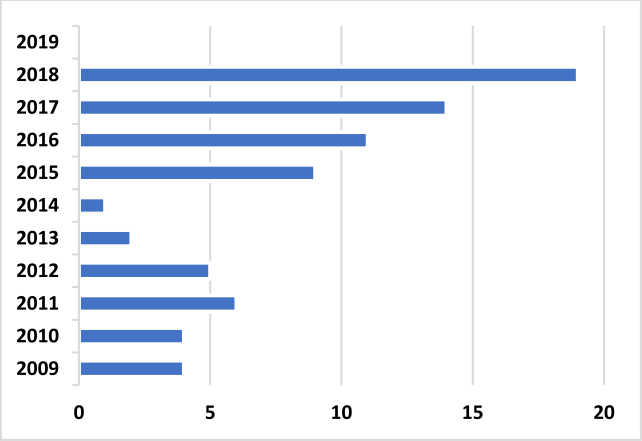

Fig. 4. Distribution of examined papers by publication year.

5.5. Distribution of examined papers by publication year

To determine the most frequent financial fraud and publication year, Table 15 and Fig. 4 present the distribution of examined papers by publication year. Referring to the table and figure below, a yearly distribution of 75 papers reviewed from 2009 to 2019 is described. As shown in the table, the gray highlighted years (2015, 2016, 2017, and 2018) represent more than 70%

Table 10
Distribution of papers related to financial crime by journals and conferences title (2009–2019).

Journal/Conference title	Frequency	Percentage (%)
Expert Systems with Applications	4	5.33
International Conference on Computing for Sustainable Global Development (INDIACom)	3	4.00
Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)	2	2.67
Decision Support Systems	2	2.67
IEEE International Conference on Information Reuse and Integration for Data Science	2	2.67
IEEE International Conference on Machine Learning and Applications	2	2.67
IEEE publisher	2	2.67
International Conference on Computer Science & Education (ICCSE)	2	2.67
International conference on information, communication & embedded systems	2	2.67
Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)	1	1.33
Computers in Human Behavior	1	1.33
Crypto Valley Conference on Blockchain Technology (CVCBT)	1	1.33
Cyber Security in Networking Conference (CSNet)	1	1.33
Engineering Applications of Artificial Intelligence	1	1.33
IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)	1	1.33
IEEE International Conference on Granular Computing	1	1.33
IEEE International Conference on Image, Vision and Computing	1	1.33
IEEE International Conference on Information Management and Engineering	1	1.33
IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)	1	1.33
IEEE Region 10 International Conference	1	1.33
IEEE Trustcom/BigDataSE/ISPA	1	1.33
Information Security for South Africa (ISSA)	1	1.33
International Conference of Signal Processing and Intelligent Systems (ICSPIS)	1	1.33
International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEEICB17)	1	1.33
International Conference on Applied Engineering (ICAE)	1	1.33
International Conference on Artificial Intelligence and Computational Intelligence	1	1.33
International Conference on circuits Power and Computing Technologies [ICCPCT]	1	1.33
International Conference on Communication, Information & Computing Technology (ICCICT)	1	1.33
International Conference on Computer Supported Cooperative Work in Design ((CSCWD))	1	1.33
International Conference on Computer, Communication and Control	1	1.33
International Conference on Computing Networking and Informatics (ICCNI)	1	1.33
International Conference on Computing, Communication and Networking Technologies	1	1.33
International Conference on Contemporary Computing (IC3)	1	1.33
International Conference on Current Trends in Computer, Electrical, Electronics and Communication	1	1.33
international Conference on Data Science and Business Analytics	1	1.33
International Conference on e-Commerce in Developing Countries:with focus on e-Security	1	1.33
International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)	1	1.33
International Conference on E-Product E-Service and E-Entertainment	1	1.33
International Conference on Information Reuse and Integration (IRI)	1	1.33
International Conference on Information Technology — New Generations	1	1.33
International Conference on Instrumentation and Measurement, Computer, Communication & Control	1	1.33
International Conference on Intelligent Computing in Data Sciences (ICDS)	1	1.33
International Conference on Intelligent Computing, Communication & Convergence	1	1.33
International Conference on Inventive Communication and Computational Technologies (ICICCT)	1	1.33
International Conference on Knowledge, Information and Creativity Support Systems	1	1.33
International Conference on Management of e-Commerce and e-Government (ICMECG)	1	1.33
International Conference on Management Science & Engineering	1	1.33
International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery	1	1.33
International Conference on Networking and Information Technology	1	1.33
International Conference on Science in Information Technology (ICSITech)	1	1.33
International Conference on Tools with Artificial Intelligence	1	1.33
International Journal of Accounting Information Systems	1	1.33
International Symposium on Innovations in Intelligent Systems and Applications	1	1.33
International Symposium on Telecommunications (IST)	1	1.33
Knowledge-Based Systems	1	1.33
Nirma University International Conference on Engineering (NUiCONE)	1	1.33
Procedia Computer Science	1	1.33
Procedia -Social and Behavioral Sciences	1	1.33
Second International Conference on Advances in Computing and Communication Engineering	1	1.33
Second World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)	1	1.33
The IEEE International Conference on Big Knowledge (ICBK)	1	1.33
World Conference on Futuristic Trends in Research and Innovation for Social Welfare	1	1.33
World Congress on Information and Communication Technologies	1	1.33
Total	**75**	**100**

percent of the reviewed papers with a total of 53 articles out of 75 articles in financial fraud detection, which indicates a high growth rate of publications for financial fraud during these years. Specifically, there was a massive rise in the published articles in the year 2018, with 19 articles followed by the year 2017, with 14 articles. While 11 articles were published in 2016 and 9 articles in 2015. However, Table 15 demonstrates that the research studies are not included in any publication of the year 2019. This might be due to the papers found in 2019 are not meeting our search criteria.

6. Conclusion and future work

Fraud is a serious problem that does not pay attention to the law or procedures of the organizations that grants an unlawful financial benefit to malicious people without any right.

Table 11
Distribution of papers categorized by countries for detecting financial fraud.

Country	Frequency	Percentage (%)
India	24	32.00
China	14	18.67
USA	9	12.00
Turkey	4	5.33
South Africa	3	4.00
Iran	3	4.00
Indonesia	3	4.00
Brazil	2	2.67
Morocco	2	2.67
Taiwan	2	2.67
Kuwait	1	1.33
Philippines	1	1.33
Malaysia	1	1.33
Nigeria	1	1.33
Czech Republic	1	1.33
Netherlands	1	1.33
Italy	1	1.33
Ghana	1	1.33
Canada	1	1.33
Total	**75**	**100**

Fraud detection is an essential part of modern financial institutions, especially in significant and sensitive technical areas. There has been a noticeable rise in the number of financial fraud and fraudulent activities in recent years. In contrast, several studies and surveys in financial fraud detection have been proposed to address these problems. However, this was not comprehensive enough as they left out some financial fraud types, validation methods, pros, and cons of most data mining techniques. In this paper, we presented a comprehensive revision of the latest data mining (DM) techniques used in detecting fraud in financial areas from the year 2009 to 2019 and classified them based on their types of fraud and techniques. This review also creates a roadmap for researchers and practitioners to make it more understandable. We observed that a few numbers of papers have been published on financial fraud detection in the past decade and it was noted that most of the researchers categorized fraud types into three main groups such as Bank, Insurance, Financial statement fraud. Hence, a new fraud type called Cryptocurrency fraud has been introduced in this paper. Considering the classification of financial fraud types table (Table 12), it is possible to conclude that cryptocurrency fraud has received less attention, even though there is notable growth in cryptocurrency in both value and use in recent years. This will help researchers to conduct more studies on identifying new methods for cryptocurrency fraud detection.

The work of Albashrawi and Lowell [11] has been extended and our thorough review provides an up-to-date comprehensive analysis. We have expanded the sample of the reviewed articles that were not included by previous research and presented a summary of the prominent works done by various researchers in the field of financial fraud that helps to disclose the context in which financial fraud types are highly exposed for such fraudulent activities. The pros, cons, and business applications of data mining techniques on financial fraud detection have been identified. The paper also provides a summary regarding various datasets and validation metrics used to estimate the performance of the data mining techniques and presents primary information with the help of tables and figures to better comprehend this field. In sum, this review can work as a reference source for both academic and practical applications by providing them with useful information and fast access to a variety of kinds of financial fraud and up-to-date data mining techniques. Besides that, researchers can take advantage of the available information and select proper techniques, datasets, and validation methods for their research project depending on the most frequently used methods.

Our result reveals that most of the papers found are related to bank fraud and insurance fraud with a total of 61 papers out of 75 reviewed papers that constitute the largest portion equal to 81.33% of the overall papers. On the other hand, we observed that the SVM is the most widely financial fraud detection technique used in data mining with a 23%, followed by both of Naïve Bayes and Random Forest, with a 15%. While the neural network is represented by a 13.5 and logistic regression,

Table 12
Classification of financial fraud types reviewed by data mining methods from 2009 to 2019.

Fraud type	Paper found	Description	Percentage (%)
Bank fraud	36	It is defined as when someone intentionally obtains money, assets, or any type of other properties possessed by, or under the control of financial institutions without any right, which is considered as a criminal offense. Fraudsters utilize a credit card to perform illegitimate transactions that are resulting in huge losses to cardholders and banks [28]. Bank fraud is classified into sub-categories which are Internet banking and credit card fraud.	48%
Insurance fraud	25	This fraud referred to the act of committing an unethical crime with the unlawful intent or misuse of an insurance policy to gain illegal profit from an insurance company. Generally, insurance is made to protect the assets and businesses of individuals or organizations from financial loss. It may occur at any stage within the insurance procedure by anyone like clients, or agents [4]. This type of fraud is sorted into three sub-categories which are healthcare insurance, auto insurance, and corp insurance fraud.	33.33%
Financial statement fraud	11	It is one of the most prevalent fraud types, which is recognized as corporate fraud, which includes correcting these statements to enable the organization to look more beneficial. The primary purpose of committing fraud is to raise the share price, getting personal loans, reducing tax liabilities or attract as many investors as possible [16].	14.67%
Cryptocurrency fraud	3	This type of fraud is one of the newest frauds targeting the underlying construction of novel technology. It is defined as a form of financial fraud that deliberately deceives naive users by providing them fake investments or services e.g. Ponzi schemes [29]. Cryptocurrency fraud includes Bitcoin fraud.	4%

Table 13
List of the most data mining technique used, pros, cons, description, and business application.

No.	Method	Description	Pros	Cons	Business application
1	Support Vector Machine (SVM)	It is essentially a classification method used in linear classification by constructing a hyperplane as the decision plane [32].	Efficient in solving non-linear classification problems and the training is relatively easy-to-use for many highly accurate financial fraud detection systems in discovering credit card fraud	Difficult to process datasets with high dimensions.	Bankruptcy prediction
2	Naïve Bayes	This tool has the ability to predict group membership [116].	Easy to implement and capable to scale with datasets.	The prediction accuracy is low compared with other classification algorithms.	Sentiment analysis
3	Random Forest (RF)	RF is one of the classification methods that operate by combining a multitude of decision trees during the time of training [32].	Fast during runtime and can work with unbalance data.	Inefficient to predict new patterns beyond the range of the training data.	Credit risk prediction
4	Neural Network	It is a multi-layer network that works similar to human thought that provides a good performance in testing a large dataset [117].	Works efficiently for detecting fraud in credit card and financial statement that has a problem binary classification and non-algorithmic	Requires high computational power for data training.	Credit Rating
5	Logistic Regression	It is an ideal classification technique used to produce the dependent variable [59].	It proves high accuracy in detecting fraud in the financial domain especially credit cards due to its ease of implementation.	Low performance with high computational.	Predict the probabilities of Credit Card failure in marketing certain product
6	K-Nearest Neighbor (KNN)	It is a type of a non-standard method of classification and regression which classifies data points according to their similar and closest classes [118].	Easy to implement and works effectively in large datasets.	The result might be inaccurate if the value of k is not determined and requires high computational power.	Money laundering analysis
7	Outliers Detection	It is a tool that used to identify data objects that do not behave as expected in accordance with general behavior or data model [78]	Suitable for detecting unlabeled data and help for model constructing involving non-standard data types.	Inefficient in detecting collective outlier effectively and causes a high false-positive rate.	Marketing analysis, money laundering, and credit card fraud
8	Decision Tree (DT)	It is an ideal classification and regression tree technique that is utilized to create a decision support tool in a tree with an internal node representing binary [119].	Simple and easy to implement, capable to process real-time data with very low computational power, and it is widely used in detecting financial statement fraud.	Requires high computational power during the initial setup.	Stock market prediction
8	Hidden Markov Model (HHM)	A dual embedded random process is used to provide more complex random processes compared to a classic Markov model [47].	Efficient learning algorithms-learning and strong statistical foundation that may happen immediately from raw sequencing data. It is extremely utilized in detecting credit card fraud.	Highly expensive.	Predict credit card frauds.
9	Genetic Algorithm (GA)	It is inspired by natural evolution that searches for the best way for solving the problem with respect to the suggested solutions that are known as chromosomes [47]	Efficient method algorithm non-algorithmic and binary classification; performs well in financial statement fraud detection.	Training set needs high computational cost	Credit card transactions fraud and Marketing mix strategizing.
10	K-Means Clustering	It is considered one of the easiest and common unsupervised ML techniques that partition a dataset into clusters with similar forms [120].	Simple to apply and computationally quicker than other clustering techniques; mainly used in healthcare insurance fraud and money laundering.	Sensitive to initial condition when predicting k value which may produce different results.	Market price and cost modeling

(*continued on next page*)

Table 13 (*continued*).

No.	Method	Description	Pros	Cons	Business application
11	Fuzzy logic (FL)	FL is logic that indicates that methods of thinking are estimated and not accurate. It is ideal in human thinking patterns, especially in almost logical reasoning [43].	Capable to deal with uncertainty and nonlinearity, simple, and does not require high computing power with respect to other techniques.	It is not capable to fix every problem.	Decision making, pattern recognition, and models for project risk assessment.
12	Bayesian Network (BN)	The basic concept of the graphical model is in the form of nodes and edges in a directed graph. BNs are useful for researching odd probability computations given the odd probabilities in the existence of uncertainty [47].	Appropriate for non-algorithmic and binary classification problems	Requires full knowledge about the behavior of the fraud type.	Financial statement fraud.

Table 14
Classification of data mining techniques based on their fraud types.

No	Method	Frequency	Fraud types					
			Bank fraud		Insurance fraud		Cryptocurrency fraud	
			Credit card fraud	Internet bank fraud	Auto	HealthCare	Financial fraud statement	Bitcoin fraud
1	SVM	17	9		1	3	4	
2	Decision Tree	6	3		2		1	
3	Logistic Regression	9	6				3	
4	Outliers Detection	7	1		3	3		
5	Bayesian Network	2	1				1	
6	K-Nearest Neighbor	8	4		2	2		
7	Neural Network	10	6		1	2	1	
8	Genetic Algorithm	5	4				1	
9	Hidden Markov Model	6	3	3				
10	Fuzzy Logic	4	2	1	1			
11	Danger Theory	1	1					
12	Naïve Bayes	11	5		2	3	1	
13	GMDH	2	1				1	
14	Gradient Boosting Tree	4	3			1		
15	Random Forest	11	4		3	3	1	
16	Autoencoder	3	1			2		
17	Evolving Clustering Method (ECM):	1				1		
18	Neural-fuzzy	1		1				
19	Behavior-based	1	1					
20	K-Means Clustering	5	2	2	1			1
21	Hierarchical Clustering	2	1		1			
22	Decision Stump	1	1					
23	Random Tree	2	2					
24	Multilayer Perceptron (MLP)	4	2		2			
25	Linear Regression (LIR)	1	1					
26	Cluster Analysis	1	1					
27	Multilayer Feed Forward Neural Network (MLFF)	1					1	
28	Probabilistic Neural Network (PNN)	1					1	
29	Ensemble classifiers	1					1	
30	PageRank-based	1				1		
31	Isolation Forest	1				1		
32	Local Outlier Factor	3				2	1	1
33	Trimmed K-Means clustering	1						1
34	kd-trees	1						1

with a 12.15. This indicates that the support vector machine technique is the leading DM technique utilized in fraud detection of the financial domains. We emphasized that our analysis proved that supervised learning approaches were found to be better performance approaches than unsupervised approaches. This is mainly due to the limitation of studies that adopt financial fraud detection techniques utilizing outlier fraud detection because of the difficulty of detecting outliers. According to [121], outlier detection is not an easy task of finding abnormal patterns linked with very few data points from huge data repositories. Outlier detection is strongly appropriate to differentiate abnormal points from normal points in the financial fraud detection area and thus deserves more attention. This paper confirms that there has been a significant increase in research on the area of financial fraud in the years 2015, 2016, 2017, and 2018. This indicates a high growth rate of publications for financial fraud during these years. Specifically, in the year 2018, with a total of 19 articles.

Table 15
Distribution of examined papers by publication year.

Year	Article count
2009	4
2010	4
2011	6
2012	5
2013	2
2014	1
2015	**9**
2016	**11**
2017	**14**
2018	**19**
2019	0
Total	**75**

This research has two limitations, first, our research scope was from 2009 to 2019 based on 4 online databases resulting in difficulty to cover all fraud type areas that are related to financial fraud detection. It apparently can see that money laundering, stock, commodity, and mortgage fraud were not included in our review due to the difficulties of gathering this data or maybe revealing the results that are related to sensitive subjects may not be permitted. Second, the 75 examined papers may not disclose the full story of utilizing data mining in financial fraud. Thus, future work could be expanded to cover several online databases seeking to include other financial fraud areas.

CRediT authorship contribution statement

Khaled Gubran Al-Hashedi: Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Visualization, Writing - original draft. **Pritheega Magalingam:** Supervision.

Declaration of competing interest

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.

Acknowledgment

The authors would like to thank the Ministry of Higher Education (MOHE), Government of Malaysia and Research Management Centre, Universiti Teknologi Malaysia for supporting this work through the Tier-2 Grant, vote number Q.K130000.2656.16J48.

References

[1] S. Alimolaei, An intelligent system for user behavior detection in Internet Banking, in: Fuzzy and Intelligent Systems (CFIS), 2015 4th Iranian Joint Congress on, IEEE, 2015, pp. 1–5.
[2] A. Abdallah, M.A. Maarof, A. Zainal, Fraud detection system: A survey, J. Netw. Comput. Appl. 68 (2016) 90–113.
[3] D. Choi, K. Lee, An artificial intelligence approach to financial fraud detection under IoT environment: A survey and implementation, Secur. Commun. Netw. 2018 (2018).
[4] S. Barman, U. Pal, M.A. Sarfaraj, B. Biswas, A. Mahata, P. Mandal, A complete literature review on financial fraud detection applying data mining techniques, Int. J. Trust Manag. Comput. Commun. 3 (2016) 336–359.
[5] M.P.S.-B. Almeida, Classification for Fraud Detection with Social Network Analysis (Masters Degree Dissertation), Engenharia Informática e de Computadores, 2009.
[6] P. Travaille, Electronic Fraud Detection in the US Medicaid Health Care Program, University of Twente, 2011.
[7] E.W. Ngai, Y. Hu, Y. Wong, Y. Chen, X. Sun, The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature, Decis. Support Syst. 50 (2011) 559–569.
[8] P. Richhariya, P.K. Singh, A survey on financial fraud detection methodologies, Int. J. Comput. Appl. 45 (2012) 15–22.
[9] Q. Liu, M. Vasarhelyi, Healthcare fraud detection: A survey and a clustering model incorporating Geo-location information, in: 29th World Continuous Auditing and Reporting Symposium (29WCARS), Brisbane, Australia, 2013.
[10] H.L. Sithic, T. Balasubramanian, Survey of insurance fraud detection using data mining techniques, 2013, arXiv preprint arXiv:1309.0806.
[11] M. Albashrawi, M. Lowell, Detecting financial fraud using data mining techniques: A decade review from 2004 to 2015, J. Data Sci. 14 (2016) 553–569.
[12] N.F. Ryman-Tubb, P. Krause, W. Garn, How Artificial Intelligence and machine learning research impacts payment card fraud detection: A survey and industry benchmark, Eng. Appl. Artif. Intell. 76 (2018) 130–157.
[13] R.R. Popat, J. Chaudhary, A survey on credit card fraud detection using machine learning, in: 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI), IEEE, 2018, pp. 1120–1125.
[14] E.A. Lopez-Rojas, S. Axelsson, A review of computer simulation for fraud detection research in financial datasets, in: Future Technologies Conference (FTC), IEEE, 2016, pp. 932–935.
[15] S. Makki, R. Haque, Y. Taher, Z. Assaghir, G. Ditzler, M.-S. Hacid, H. Zeineddine, Fraud analysis approaches in the age of big data-A review of state of the art, foundations and applications of self* systems (FAS* W), in: 2017 IEEE 2nd International Workshops on, IEEE, 2017, pp. 243–250.
[16] J. West, M. Bhattacharya, Intelligent financial fraud detection: a comprehensive review, Comput. Secur. 57 (2016) 47–66.
[17] R.J. Bolton, D.J. Hand, Statistical fraud detection: A review, Statist. Sci. (2002) 235–249.
[18] Y. Kou, C.-T. Lu, S. Sirwongwattana, Y.-P. Huang, Survey of fraud detection techniques, in: IEEE International Conference on Networking, Sensing and Control, IEEE, 2004, pp. 749–754.
[19] T. Allan, J. Zhan, Towards fraud detection methodologies, in: 2010 5th International Conference on Future Information Technology, IEEE, 2010, pp. 1–6.
[20] M. Pejic-Bach, Profiling intelligent systems applications in fraud detection and prevention: survey of research articles, in: 2010 International Conference on Intelligent Systems, Modelling and Simulation, IEEE, 2010, pp. 80–85.
[21] L. Delamaire, H. Abdou, J. Pointon, Credit card fraud and detection techniques: a review, Banks Bank Syst. 4 (2009) 57–68.
[22] D. Zhang, L. Zhou, Discovering golden nuggets: data mining in financial application, IEEE Trans. Syst. Man Cybern. C 34 (2004) 513–522.
[23] S.B.E. Raj, A.A. Portia, Analysis on credit card fraud detection methods, in: 2011 International Conference on Computer, Communication and Electrical Technology (ICCCET), IEEE, 2011, pp. 152–156.
[24] C. Phua, V. Lee, K. Smith, R. Gayler, A comprehensive survey of data mining-based fraud detection research, 2010, arXiv preprint arXiv:1009.6119.
[25] J. Li, K.-Y. Huang, J. Jin, J. Shi, A survey on statistical methods for health care fraud detection, Health Care Manag. Sci. 11 (2008) 275–287.
[26] M. Ahmed, A.N. Mahmood, M.R. Islam, A survey of anomaly detection techniques in financial domain, Future Gener. Comput. Syst. 55 (2016) 278–288.
[27] K. Chaudhary, J. Yadav, B. Mallick, A review of fraud detection techniques: Credit card, Int. J. Comput. Appl. 45 (2012) 39–44.
[28] G. Potamitis, Design and Implementation of a Fraud Detection Expert System using Ontology-Based Techniques, University of Manchester, 2013).
[29] I. DeMartino, The Bitcoin Guidebook: How to Obtain, Invest, and Spend the World's First Decentralized Cryptocurrency, Simon and Schuster, 2018.
[30] M.K.-M. Ly, Coining Bitcoin's legal-bits: Examining the regulatory framework for Bitcoin and virtual currencies, Harv. JL & Tech. 27 (2013) 587.
[31] M. Vasek, T. Moore, There's no free lunch, even using Bitcoin: Tracking the popularity and profits of virtual currency scams, in: International Conference on Financial Cryptography and Data Security, Springer, 2015, pp. 44–61.
[32] S. Bhattacharyya, S. Jha, K. Tharakunnel, J.C. Westland, Data mining for credit card fraud: A comparative study, Decis. Support Syst. 50 (2011) 602–613.

[33] W. Xu, Y. Liu, An optimized SVM model for detection of fraudulent online credit card transactions, management of e-Commerce and e-Government (ICMeCG), in: 2012 International Conference on, IEEE, 2012, pp. 14–17.
[34] I. Rajak, K.J. Mathai, Intelligent fraudulent detection system based SVM and optimized by danger theory, in: Computer, Communication and Control (IC4), 2015 International Conference on, IEEE, 2015, pp. 1–4.
[35] M. Zareapoor, P. Shamsolmoali, Application of credit card fraud detection: Based on bagging ensemble classifier, Procedia Comput. Sci. 48 (2015) 679–685.
[36] N.K. Gyamfi, J.-D. Abdulai, Bank fraud detection using support vector machine, in: 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), IEEE, 2018, pp. 37–41.
[37] V. Mareeswari, G. Gunasekaran, Prevention of credit card fraud detection based on HSVM, in: Information Communication and Embedded Systems (ICICES), 2016 International Conference on, IEEE, 2016, pp. 1–4.
[38] G.G. Sundarkumar, V. Ravi, V. Siddeshwar, One-class support vector machine based undersampling: Application to churn prediction and insurance fraud detection, in: Computational Intelligence and Computing Research (ICCIC), 2015 IEEE International Conference on, IEEE, 2015, pp. 1–7.
[39] C. Francis, N. Pepper, H. Strong, Using support vector machines to detect medical fraud and abuse, in: Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE, IEEE, 2011, pp. 8291–8294.
[40] A. Mishra, C. Ghorpade, Credit card fraud detection on the skewed data using various classification and ensemble techniques, in: 2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), IEEE, 2018, pp. 1–5.
[41] M. Jeragh, M. AlSulaimi, Combining auto encoders and one class support vectors machine for fraudulant credit card transactions detection, in: 2018 Second World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4), IEEE, 2018, pp. 178–184.
[42] Q. Deng, Application of support vector machine in the detection of fraudulent financial statements, in: Computer Science & Education, 2009. ICCSE'09. 4th International Conference on, IEEE, 2009, pp. 1056–1059.
[43] K. Supraja, S. Saritha, Robust fuzzy rule based technique to detect frauds in vehicle insurance, in: 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), IEEE, 2017, pp. 3734–3739.
[44] T.K. Behera, S. Panigrahi, Credit card fraud detection: a hybrid approach using fuzzy clustering & neural network, in: Advances in Computing and Communication Engineering (ICACCE), 2015 Second International Conference on, IEEE, 2015, pp. 494–499.
[45] M.R. HaratiNik, M. Akrami, S. Khadivi, M. Shajari, FUZZGY: A hybrid model for credit card fraud detection, in: Telecommunications (IST), 2012 Sixth International Symposium on, IEEE, 2012, pp. 1088–1093.
[46] F.S. Nezhad, H.R. Shahriari, Fuzzy logic and Takagi-Sugeno Neural-Fuzzy to Deutsche bank fraud transactions, in: E-Commerce in Developing Countries: With Focus on E-Security (ECDC), 2013 7th Intenational Conference on, IEEE, 2013, pp. 1–15.
[47] Z. Zojaji, R.E. Atani, A.H. Monadjemi, A survey of credit card fraud detection techniques: data and technique oriented perspective, 2016, arXiv preprint arXiv:1611.06439.
[48] A. Agrawal, S. Kumar, A.K. Mishra, Credit card fraud detection: A case study, in: Computing for Sustainable Global Development (INDIACom), 2015 2nd International Conference on, IEEE, 2015, pp. 5–7.
[49] A. Khan, T. Singh, A. Sinhal, Implement credit card fraudulent detection system using observation probabilistic in hidden markov model, in: Engineering (NUiCONE), 2012 Nirma University International Conference on, IEEE, 2012, pp. 1–6.
[50] S.S. Mhamane, L. Lobo, Internet banking fraud detection using HMM, in: Computing Communication & Networking Technologies (ICCCNT), 2012 Third International Conference on, IEEE, 2012, pp. 1–4.
[51] X. Wang, H. Wu, Z. Yi, Research on bank anti-fraud model based on K-means and hidden Markov model, in: 2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC), IEEE, 2018, pp. 780–784.
[52] V. Bhusari, S. Patil, Study of hidden Markov model in credit card fraudulent detection, in: Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave), World Conference on, IEEE, 2016, pp. 1–4.
[53] D. Iyer, A. Mohanpurkar, S. Janardhan, D. Rathod, A. Sardeshmukh, Credit card fraud detection using Hidden Markov model, in: 2011 World Congress on Information and Communication Technologies, 2011, pp. 1062–1066.
[54] Y. Sahin, E. Duman, Detecting credit card fraud by ANN and logistic regression, in: Innovations in Intelligent Systems and Applications (INISTA), 2011 International Symposium on, IEEE, 2011, pp. 315–319.
[55] A. Srivastava, M. Yadav, S. Basu, S. Salunkhe, M. Shabad, Credit card fraud detection at merchant side using neural networks, in: Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on, IEEE, 2016, pp. 667–670.
[56] F. Ghobadi, M. Rohani, Cost sensitive modeling of credit card fraud using neural network strategy, in: Signal Processing and Intelligent Systems (ICSPIS), International Conference of, IEEE, 2016, pp. 1–5.
[57] K. Randhawa, C.K. Loo, M. Seera, C.P. Lim, A.K. Nandi, Credit card fraud detection using adaboost and majority voting, IEEE ACCESS 6 (2018) 14277–14284.
[58] A. El Bouchti, A. Chakroun, H. Abbar, C. Okar, Fraud detection in banking using deep reinforcement learning, in: 2017 Seventh International Conference on Innovative Computing Technology (INTECH), IEEE, 2017, pp. 58–63.
[59] P. Ravisankar, V. Ravi, G.R. Rao, I. Bose, Detection of financial statement fraud and feature selection using data mining techniques, Decis. Support Syst. 50 (2011) 491–500.
[60] M.H. Özçelik, E. Duman, M. Işik, T. Çevik, Improving a credit card fraud detection system using genetic algorithm, in: Networking and Information Technology (ICNIT), 2010 International Conference on, IEEE, 2010, pp. 436–440.
[61] I. Benchaji, S. Douzi, B. ElOuahidi, Using genetic algorithm to improve classification of imbalanced datasets for credit card fraud detection, in: 2018 2nd Cyber Security in Networking Conference (CSNet), IEEE, 2018, pp. 1–5.
[62] J. Liang, W. Lv, Research on detecting technique of financial statement fraud based on fuzzy genetic algorithms BPN, in: 2009 International Conference on Management Science and Engineering, IEEE, 2009, pp. 1462–1468.
[63] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: A survey, ACM Comput. Surv. (CSUR) 41 (2009) 15.
[64] R. Hassanzadeh, Anomaly Detection in Online Social Networks: Using Data-Mining Techniques and Fuzzy Logic, Queensland University of Technology, 2014.
[65] N. Malini, M. Pushpa, Analysis on credit card fraud identification techniques based on KNN and outlier detection, in: Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), 2017 Third International Conference on, IEEE, 2017, pp. 255–258.
[66] Y. Heryadi, L.A. Wulandhari, B.S. Abbas, Recognizing debit card fraud transaction using CHAID and K-nearest neighbor: Indonesian Bank case, in: Knowledge, Information and Creativity Support Systems (KICSS), 2016 11th International Conference on, IEEE, 2016, pp. 1–5.
[67] T. Badriyah, L. Rahmaniah, I. Syarif, Nearest neighbour and statistics method based for detecting fraud in auto insurance, in: 2018 International Conference on Applied Engineering (ICAE), IEEE, 2018, pp. 1–5.
[68] J.O. Awoyemi, A.O. Adetunmbi, S.A. Oluwadare, Credit card fraud detection using machine learning techniques: A comparative analysis, in: Computing Networking and Informatics (ICCNI), 2017 International Conference on, IEEE, 2017, pp. 1–9.
[69] Q. Deng, Detection of fraudulent financial statements based on Naïve Bayes classifier, in: 2010 5th International Conference on Computer Science & Education, IEEE, 2010, pp. 1032–1035.
[70] R.A. Bauder, T.M. Khoshgoftaar, A. Richter, M. Herland, Predicting medical provider specialties to detect anomalous insurance claims, in: 2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI), IEEE, 2016, pp. 784–790.
[71] M. Herland, R.A. Bauder, T.M. Khoshgoftaar, Medical provider specialty predictions for the detection of anomalous medicare insurance claims, in: 2017 IEEE International Conference on Information Reuse and Integration (IRI), IEEE, 2017, pp. 579–588.
[72] P. Hajek, R. Henriques, Mining corporate annual reports for intelligent detection of financial statement fraud–A comparative study of machine learning methods, Knowl.-Based Syst. 128 (2017) 139–152.
[73] J.R.D. Kho, L.A. Vea, Credit card fraud detection based on transaction behavior, in: Region 10 Conference, TENCON 2017-2017 IEEE, IEEE, 2017, pp. 1880–1884.
[74] J.V. Devi, K. Kavitha, Fraud detection in credit card transactions by using classification algorithms, in: 2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC), IEEE, 2017, pp. 125–131.
[75] R. Roy, K.T. George, Detecting insurance claims fraud using machine learning techniques, in: 2017 International Conference on Circuit, Power and Computing Technologies (ICCPCT), IEEE, 2017, pp. 1–6.
[76] S. Subudhi, S. Panigrahi, Effect of class imbalanceness in detecting automobile insurance fraud, in: 2018 2nd International Conference on Data Science and Business Analytics (ICDSBA), IEEE, 2018, pp. 528–531.
[77] D. Yue, X. Wu, N. Shen, C.-H. Chu, Logistic regression for detecting fraudulent financial statement of listed companies in China, in: 2009 International Conference on Artificial Intelligence and Computational Intelligence, IEEE, 2009, pp. 104–108.

[78] M. Goldstein, S. Uchida, A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data, PLoS One 11 (2016) e0152173.
[79] S. Agrawal, J. Agrawal, Survey on anomaly detection using data mining techniques, Procedia Comput. Sci. 60 (2015) 708–713.
[80] J. Peng, Q. Li, H. Li, L. Liu, Z. Yan, S. Zhang, Fraud detection of medical insurance employing outlier analysis, in: 2018 IEEE 22nd International Conference on Computer Supported Cooperative Work in Design (CSCWD), IEEE, 2018, pp. 341–346.
[81] C. Yan, Y. Li, The identification algorithm and model construction of automobile insurance fraud based on data mining, in: 2015 Fifth International Conference on Instrumentation and Measurement, Computer, Communication and Control (IMCCC), IEEE, 2015, pp. 1922–1928.
[82] W. Zhang, X. He, An anomaly detection method for medicare fraud detection, in: 2017 IEEE International Conference on Big Knowledge (ICBK), IEEE, 2017, pp. 309–314.
[83] X. Li, S. Ying, Lib-SVMs detection model of regulating-profits financial statement fraud using data of chinese listed companies, in: 2010 International Conference on E-Product E-Service and E-Entertainment, IEEE, 2010, pp. 1–4.
[84] M. Kirlidog, C. Asuk, A fraud detection approach with data mining in health insurance, Proc.-Soc. Behav. Sci. 62 (2012) 989–994.
[85] A.A. Rizki, I. Surjandari, R.A. Wayasti, Data mining application to detect financial fraud in Indonesia's public companies, in: Science in Information Technology (ICSITech), 2017 3rd International Conference on, IEEE, 2017, pp. 206–211.
[86] H. Peng, M. You, The health care fraud detection using the pharmacopoeia spectrum tree and neural network analytic contribution hierarchy process, in: 2016 IEEE Trustcom/BigDataSE/ISPA, IEEE, 2016, pp. 2006–2011.
[87] E. Duman, M.H. Ozcelik, Detecting credit card fraud by genetic algorithm and scatter search, Expert Syst. Appl. 38 (2011) 13057–13063.
[88] A. Agrawal, S. Kumar, A.K. Mishra, Implementation of novel approach for credit card fraud detection, in: Computing for Sustainable Global Development (INDIACom), 2015 2nd International Conference on, IEEE, 2015, pp. 1–4.
[89] W.N. Robinson, A. Aria, Sequential fraud detection for prepaid cards using hidden Markov model divergence, Expert Syst. Appl. 91 (2018) 235–251.
[90] Y. Sahin, S. Bulkan, E. Duman, A cost-sensitive decision tree approach for fraud detection, Expert Syst. Appl. 40 (2013) 5916–5923.
[91] R. Bauder, R. da Rosa, T. Khoshgoftaar, Identifying medicare provider fraud with unsupervised machine learning, in: 2018 IEEE International Conference on Information Reuse and Integration (IRI), IEEE, 2018, pp. 285–292.
[92] M. Anbarasi, S. Dhivya, Fraud detection using outlier predictor in health insurance data, in: 2017 International Conference on Information Communication and Embedded Systems (ICICES), IEEE, 2017, pp. 1–6.
[93] G. van Capelleveen, M. Poel, R.M. Mueller, D. Thornton, J. van Hillegersberg, Outlier detection in healthcare fraud: A case study in the medicaid dental domain, Int. J. Account. Inf. Syst. 21 (2016) 18–31.
[94] J. Seo, O. Mendelevitch, Identifying frauds and anomalies in medicare-b dataset, in: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2017, pp. 3664–3667.
[95] Y. Gao, C. Sun, R. Li, Q. Li, L. Cui, B. Gong, An efficient fraud identification method combining manifold learning and outliers detection in mobile healthcare services, IEEE Access 6 (2018) 60059–60068.
[96] P. Monamo, V. Marivate, B. Twala, Unsupervised learning for robust bitcoin fraud detection, in: Information Security for South Africa (ISSA), IEEE, 2016, pp. 129–134.
[97] P.M. Monamo, V. Marivate, B. Twala, A multifaceted approach to bitcoin fraud detection: Global and local outliers, in: Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on, IEEE, 2016, pp. 188–194.
[98] E.M. Carneiro, L.A.V. Dias, A.M. da Cunha, L.F.S. Mialaret, Cluster analysis and artificial neural networks: A case study in credit card fraud detection, in: 2015 12th International Conference on Information Technology-New Generations (ITNG), IEEE, 2015, pp. 122–126.
[99] V. Rawte, G. Anuradha, Fraud detection in health insurance using data mining techniques, in: Communication, Information & Computing Technology (ICCICT), 2015 International Conference on, IEEE, 2015, pp. 1–5.
[100] N. Omar, Z.A. Johari, M. Smith, Predicting fraudulent financial reporting using artificial neural network, J. Financ. Crime 24 (2017) 362–387.
[101] S.-Y. Huang, R.-H. Tsaih, F. Yu, Topological pattern discovery and feature extraction for fraudulent financial reporting, Expert Syst. Appl. 41 (2014) 4360–4372.
[102] Q. Deng, G. Mei, Combining self-organizing map and K-means clustering for detecting fraudulent financial statements, in: 2009 IEEE International Conference on Granular Computing, 2009.
[103] W. Xiaoyun, L. Danyue, Hybrid outlier mining algorithm based evaluation of client moral risk in insurance company, in: 2010 2nd IEEE International Conference on Information Management and Engineering, IEEE, 2010, pp. 585–589.
[104] A.G. de Sá, A.C. Pereira, G.L. Pappa, A customized classification algorithm for credit card fraud detection, Eng. Appl. Artif. Intell. 72 (2018) 21–29.
[105] S.-H. Li, D.C. Yen, W.-H. Lu, C. Wang, Identifying the signs of fraudulent accounts using data mining techniques, Comput. Hum. Behav. 28 (2012) 1002–1013.
[106] Y. Li, C. Yan, W. Liu, M. Li, Research and application of random forest model in mining automobile insurance fraud, in: 2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), IEEE, 2016, pp. 1756–1761.
[107] G. Kowshalya, M. Nandhini, Predicting fraudulent claims in automobile insurance, in: 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), IEEE, 2018, pp. 1338–1343.
[108] M. Bartoletti, B. Pes, S. Serusi, Data mining for detecting Bitcoin Ponzi schemes, in: 2018 Crypto Valley Conference on Blockchain Technology (CVCBT), 2018, pp. 75–84.
[109] S. Patil, V. Nemade, P.K. Soni, Predictive modelling for credit card fraud detection using data analytics, Procedia Comput. Sci. 132 (2018) 385–395.
[110] R. Bauder, T. Khoshgoftaar, Medicare fraud detection using random forest with class imbalanced big data, in: 2018 IEEE International Conference on Information Reuse and Integration (IRI), IEEE, 2018, pp. 80–87.
[111] R. Saia, S. Carta, Evaluating the benefits of using proactive transformed-domain-based techniques in fraud detection tasks, Future Gener. Comput. Syst. 93 (2019) 18–32.
[112] R.A. Bauder, T.M. Khoshgoftaar, Medicare fraud detection using machine learning methods, in: 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE, 2017, pp. 858–865.
[113] S. Carta, G. Fenu, D.R. Recupero, R. Saia, Fraud detection for E-commerce transactions by employing a prudential multiple consensus model, J. Inf. Secur. Appl. 46 (2019) 13–22.
[114] A. Singh, A. Jain, Adaptive credit card fraud detection techniques based on feature selection method, in: Advances in Computer Communication and Computational Sciences, Springer, 2019, pp. 167–178.
[115] I. Sadgali, N. Sael, F. Benabbou, Performance of machine learning techniques in the detection of financial frauds, Procedia Comput. Sci. 148 (2019) 45–54.
[116] C. Holton, Identifying disgruntled employee systems fraud risk through text mining: A simple solution for a multi-billion dollar problem, Decis. Support Syst. 46 (2009) 853–864.
[117] S.-H. Liao, P.-H. Chu, P.-Y. Hsiao, Data mining techniques and applications–A decade review from 2000 to 2011, Expert Syst. Appl. 39 (2012) 11303–11311.
[118] O.R. Bidder, H.A. Campbell, A. Gómez-Laich, P. Urgé, J. Walker, Y. Cai, L. Gao, F. Quintana, R.P. Wilson, Love thy neighbour: automatic animal behavioural classification of acceleration data using the k-nearest neighbour algorithm, PLoS One 9 (2014).
[119] A. Gepp, J.H. Wilson, K. Kumar, S. Bhattacharya, A comparative analysis of decision trees vis-a-vis other computational data mining techniques in automotive insurance fraud detection, J. Data Sci. 10 (2012) 537–561.
[120] J. Wu, H. Xiong, P. Wu, J. Chen, Local decomposition for rare class analysis, in: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2007, pp. 814–823.
[121] M. Agyemang, K. Barker, R. Alhajj, A comprehensive survey of numeric and symbolic outlier mining techniques, Intell. Data Anal. 10 (2006) 521–538.