Financial Fraud Detection using Data Mining: A Survey

Sudhansu Ranjan Lenka[1]
Research Scholar, Computer Science & Applications,
Utkal University,
Bhubaneswar, Odisha, India

Dr. Bikram Kesari Ratha[2]
Reader, Computer Science & Applications,
Utkal University,
Bhubaneswar, Odisha, India

Abstract
Due to levitate and rapid growth of E-Commerce, most of the organizations are moving towards cashless transaction Unfortunately, the cashless transactions are not only used by legitimate users but also it is used by illegitimate users and which results in trouncing of billions of dollars each year worldwide. Fraud prevention and Fraud Detection are two methods used by the financial institutions to protect against these frauds. Fraud prevention systems (FPSs) are not sufficient enough to provide fully security to the E-Commerce systems. However, with the combined effect of Fraud Detection Systems (FDS) and FPS might protect the frauds. However, there still exist so many issues and challenges that degrade the performances of FDSs, such as overlapping of data, noisy data, misclassification of data, etc. This paper presents a comprehensive survey on financial fraud detection system using such data mining techniques. Over seventy research papers have been reviewed, mainly within the period 2002–2015, were analyzed in this study. The data mining approaches employed in this research includes Neural Network, Logistic Regression, Bayesian Belief Network, Support Vector Machine (SVM), Self Organizing Map(SOM), K-Nearest Neighbor(K-NN), Random Forest and Genetic Algorithm. The algorithms that have achieved high success rate in detecting credit card fraud are Logistic Regression (99.2%), SVM (99.6%) and Random Forests (99.6%). But, the most suitable approach is SOM because it has achieved perfect accuracy of 100%. But the algorithms implemented for financial statement fraud have shown a large difference in accuracy from CDA at 71.4% to a probabilistic neural network with 98.1%. In this paper, we have identified the research gap and specified the performance achieved by different algorithms based on parameters like, accuracy, sensitivity and specificity. Some of the key issues and challenges associated with the FDS have also been identified.

Keywords
Fraud detection systems; Data mining; Concept drift; Overlapping of data; Noisy data; Misclassification of data

I. INTRODUCTION

With the invention of the modern technologies and the booming of the internet has given rise to fraudulent activities associated with all E-Commerce business and financial transactions. Recently, most of the organizations, companies and financial institutions have implemented E-Commerce systems in order to increase the efficiency of their businesses and services. Some of the areas where most of the fraudulent activity takes place are credit card fraud, financial statement fraud, insurance fraud, money laundering, telecommunications fraud and computer intrusion fraud [1][4][16][50].

Fraud means obtaining services/goods and/or money by unethical means, and is a major problem all over the world. According to Association of Certified Fraud Examiners (ACFE) fraud means “deliberately misuse or misapplication of the employee organization's resources for personal benefit” [8]. Financial fraud is a broad term which can also be defined as the intentional use of illegal methods or activities for the purpose of getting financial gain [77]. Financial losses due to fraud not only affect the banks and organizations but also affect in our daily life. Fraud destabilizes economies, create a negative reputation of the industries and largely affect the people’s life. As stated in Symantec Internet Security Threat Report, more than three billion malware attacks were reported in 2010 and the number of denial of service attacks increased drastically by 2015. According to Symantec Internet Security Threat Report 2016[63], there were more than one million web attacks each day in 2015. The Global State of Information Security Survey 2016 [65] had found an increase in fraudulent activities. Figure 1 shows there is a constant rise in the security issues from 2009 to 2015.

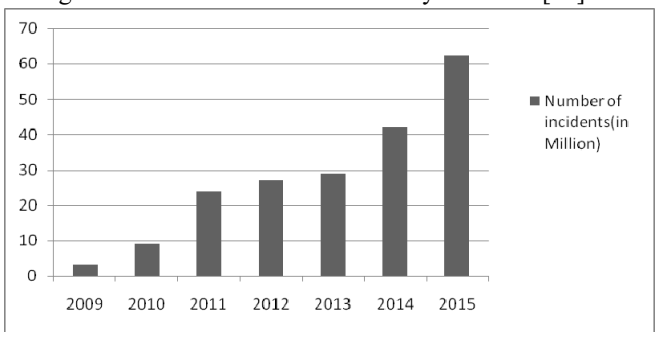
There are two mechanisms to prevent the fraudulent activities i.e. fraud prevention and fraud detection. The purpose of fraud prevention mechanism is to protect the system against fraud by halting the occurrence of fraud at the first level. However, this mechanism alone is not sufficient enough to completely stop fraudulent activities. So the second level of defense, fraud detection mechanism should be implemented to improve the security systems. FDS detects and recognizes fraudulent activities as they enter the system and reports about it to a system administrator [11].

Similar to Intrusion Detection System (IDS), FDS also follows misuse and anomaly based methods to discover the fraud in the system. Both these approaches implements data mining techniques to detect fraud from large data stream [46]. Data mining techniques are implemented to detect the fraud from huge voluminous of data and it has the ability to detect and extract the hidden truths form huge

Manuscript received September 5, 2024
Manuscript revised September 20, 2024
https://doi.org/10.22937/IJCSNS.2024.24.9.20

voluminous data [46]. Data mining techniques have already successfully implemented in areas like credit card fraud detection, bankruptcy prediction and for analysis in share markets [49]. It can be implemented as a tool to classify the transactions as fraudulent transactions and such transactions can be used for further consideration. It follows two types of classification approaches, such as statistical and computational approaches. Statistical approach implements traditional mathematics techniques, such as logistic regression and Bayesian theory. Computational approach implements intelligence techniques, such as neural networks and support vector machines. Both these approaches have some similarities, but the main difference is that computational approaches are capable of learning from the problem domain, whereas statistical approaches have no learning capacity. Both these approaches of data mining will be surveyed in this paper.

Fig. 1. Growth of information security incidents [65].



However, there exist so many issues and challenges that degrade the performance of FDS for E-Commerce system, such as concept drift, detecting fraud at real time, skewed distribution, dealing with huge voluminous data, misclassification cost, etc. This gives rise to high false rates, low accuracy and slows down detection rate. In this paper, we will focus the implementation of FDS in five different areas, like credit card, telecommunication, automobile insurance, health insurance and online auction. The remainder of the paper is outlined as follows. Section 2 contains the related review and survey papers in the fraud detection system. Section 3 addresses the different types of financial fraud. Section 4 list out the approaches and techniques used to protect against fraud. Section 5 presents an overview of the various data mining techniques used. Section 6 presents the outcomes of previous researchers in identifying the financial fraud. Section 7 presents the challenges and hurdles faced by fraud detection systems. Section 8 provides a conclusion to our research and a discussion of our research.

II. RELATED WORK

Table1 shows the timeline and the research methods undertaken to detect financial fraud. Initially lot of researchers implemented statistical methods such as logistic regression, as well as neural networks in financial fraud detection [54][74]. Zhang et al. had implemented neural network for forecasting the fraud in financial applications [76]. Sohl and Venkatachalam was the first person to predict financial statement fraud using Back-Propagation network [61].

Bolton and Hand [15][16] have implemented statistical techniques for fraud detection in area such as credit card, money laundering, telecommunications etc , similarly Rezaee [57] had done detail investigation on financial fraud statement . In 2002, Maes et al. have done a comparative study between the Bayesian Belief Network (BBN) and Artificial Neural Network algorithm taking real credit card data and the result of their study shows that BBNs were more accurate and much faster as compared to ANN [44].This shows that NN might not be a best approach for credit card fraud detection but the results can be improved by integrating NN with some other alternative algorithms.

Recently, the researchers are implementing totally different approaches for fraud detection, but still the former techniques have importance. In 2004 Kou et al. have detail surveyed on fraud detection system implementing analytic approaches and neural networks [40]. Zhang and Zhou [23] have investigated data mining techniques in the context of financial applications and also they have compared different data mining techniques. In the next year Vatsa et al. modeled fraudsters and detection methods by using game theory approach and their approach shows large financial advantages [67].

In 2006 Yang and Hwang detect fraud and abuse in health care by using process mining approach [73]. In 2007 both Pinquet et al. and Viane et al. have investigated on logistic regression to detect fraud in automobile insurance industry using Spanish automobile insurance database[53][68]. In the same year Kirkos et al. have done comparisons between statistical methods with neural network to detect fraudulent activities in Greek manufacturing companies [39]. Bose and Wang implement classification and regression trees to detect fraud in selected Chinese companies [17]. In 2007 Hoogs et al implemented genetic algorithm on Accounting and Auditing Enforcement Releases to detect fraudulent activities in US based companies [34]. Yeu et al. in 2007 claimed that the classification based algorithms are very successful and commonly researched techniques for fraud detection [74].

Bai et al. implement one new statistical technique, known as Classification and Regression Tree (CART) to identify the fraud financial statement in 2008 [49]. Bermudez et al. in 2008 have implemented a statistical

method to detect insurance fraud using the same data samples as Pinquet et al. and Viaene et al. used before. [13]. Quah et al. took Self -Organizing Maps to detect credit card fraud taking real-world data samples from an international bank of Singapore [54]. Wu and Banzhaf combined the artificial immune system method with a co-evolutionary technique to solve financial fraud with automatic teller and point-of-sale data for a financial organization [72]. In 2009 Holton utilized the same hybrid technique by combining text mining approach with Bayesian belief networks to detect the fraudulent activities likely to be committed by the disgruntled employees of the organization [33]. In the same year Panigrahi et al. combined three different approaches such as rule-based filtering, Dempster–Shafer theory and Bayesian learning to solve the credit card fraudulent activities using synthesised data [49]. Whitrow et al. solved credit card fraud detection problem and also they have compared support vector machines methods with decision trees by aggregating common transactional variables to generate new data samples [70]. Delamaire et al. have studied different types of credit card fraudulent activities, namely bankruptcy fraud, counterfeit fraud, theft fraud, application fraud and behavioral fraud, additionally they have also focused on different methods to detect them, such as pair-wise matching, decision trees, clustering techniques, neural networks, and genetic algorithms [25].

Similarly, in 2010 Cecchini et al. studied Accounting and Auditing Enforcement Releases (AAER) and tried to predict financial statement fraud in US companies by combining two different approaches i.e. text mining and support vector machine [18]. In 2011 Bhattacharyya et al. evaluated the performance of random forests, support vector machines and logistic regression on a large real credit card transactions data to detect fraudulent behavior [14].

Duman and Ozcelik integrate both genetic algorithms and scatter search method to detect the occurrence of credit card fraud [26]. Similarly, Humphrys et al. implement the same hybrid technology by combining different classifier like support vector machine, decision tree, and Bayesian belief network and able to successfully identify the fraud form company's financial statements [36]. In 2011 Ngai et al. have done detailed review of the different data mining techniques, like classification, regression and clustering, additionally they focus on their implementations to different financial fraud, like bank fraud, insurance fraud, securities and commodities fraud, etc [46]. In the same year Raj et al. survey on different techniques to identify the fraudulent activities in credit card transactions and evaluates the performance of each methods [55].

Behdad et al. reviewed the electronic fraud detection systems along with the nature inspired detection techniques. They have done detailed study on Nature inspired techniques i.e. artificial intelligence techniques along with the obstacles and challenges faced by the computational intelligence systems [11]. In 2012 Wong et al. implementing artificial immune system tried to solve the credit card fraud related problems for a major Australian bank [71].

In 2013 Huang implements logistic regression and a support vector machine to identify the financial statement fraud in a group of Taiwanese companies [35]. Lookman Sithic and Balasubramanian in 2013 have done detailed survey on different categories of fraud in medical and motor insurance sectors and the different data mining techniques that are used to identify fraud in these insurance sectors [43].

In 2014 Olszewski detect credit card fraud implementing self organizing maps [48]; Soltani Halvaiee and Akbari identify the credit card fraud for a Brazilian bank using an artificial immune system [62].

III. TYPES OF FINANCIAL FRAUD

Any system where money and services involved they definitely fraudulent activities can be expected, example of such systems are credit card , telecommunication , health care insurance system, etc [6]. Fig. 2 depicts the most common areas of fraud, Such as credit card, telecommunication, health care insurance, automobile insurance, online auction, etc and this section includes the brief description of such areas. Fig. 3 shows the statistics report of the fraud areas and from the figure we conclude that maximum research work is based on bank fraud followed by insurance fraud. Tele communication and Internet marketing are the areas where less research work have been done [2].

TABLE 1. Timeline and the techniques applied to FDS

References	**Techniques applied/ Area of Research**	**Fraud Area**	Year of Publi-cation
[40]	Analytic approaches and neural networks	credit card, telecommunication and computer intrusion detection	2004
[67]	Game theory approach	Credit card	2005
[73]	Process mining	health care sector	2006
[53]	Logistic regression	Insurance fraud	2007
[68]	Logistic regression	Automobile insurance	2007
[17]	Statistical methods and neural network	Different Financial Organizations	2007
[39]	Statistical methods and neural network	Different Financial Organizations	2007
[34]	Genetic algorithm	Financial statement	2007
[74]	Survey on fraud detection approach	Financial statement	2007

[10]	Classification and Regression Tree (CART)	Financial statement	2008
[13]	Statistical approach	Insurance fraud	2008
[54]	Self -Organising Maps	Credit card	2008
[72]	Artificial immune system	Financial statement	2008
[33]	Bayesian Belief Network	Corporate sector	2009
[70]	Comparison of methods	Credit card	2009
[49]	A hybrid approach	Credit card	2009
[33]	Bayesian Belief Network	Corporate fraud	2009
[59]	Self- Organising maps	Credit card	2009
[18]	Text mining hybrid	Financial statement	2010
[46]	Review of fraud defined research	Financial statement	2011
[37]	Process mining	Corporate fraud	2011
[14]	Comparison of data mining techniques	Credit card	2011
[36] [31]	Text mining hybrids	Financial statement	2011
[56]	Detailed comparison of detection methods	Credit card	2011
[71]	Artificial immune system	Credit card	2012
[58]	Decision Trees	Financial statement	2013
[35]	Support vector machine	Financial statement	2013
[48]	Self- Organising maps	Credit card	2014

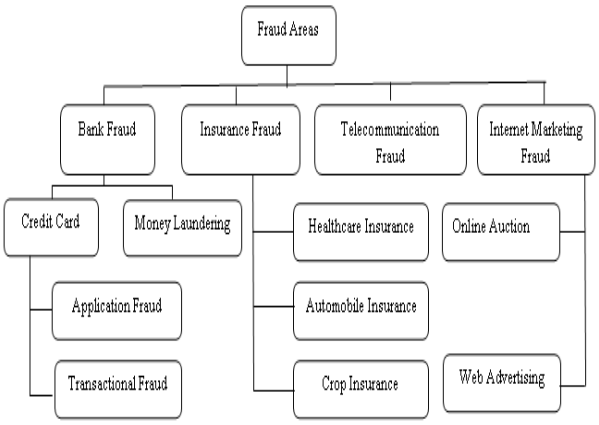


Fig.2 Classification of the most common areas of fraud

A. *Credit Card Fraud*

Credit card fraud means illegal use of someone's credit card without the owner's knowledge to perform fraudulent transactions [46]. The transactions can be either in offline mode (Card Present) or in online mode (Card Not Present) [55]. The former transaction takes place using physical card is lost or stolen by the fraudster, using it as actual owner. This type of fraud very rarely occurs because when the card holder complains about the card lost, then the bank personnel immediately lock the lost card. The online or CNP transaction occurs remotely and in this case the credit card information's are acquired by the fraudsters by using any of these following methods. Phishing is the method, in which the fraudster acts as bank personnel and tried to convince the user to divulge their details, skimmers is the technology which provides an interface to an ATM or POS device to retrieve the card information directly, or fraudster breach the bank security system and able to retrieve entire databases of user's information [55] [49].

Credit card fraud can also be classified into two categories, namely application fraud and behavioral fraud [26]. Application fraud occurs when fraudster opens a new credit card account by filling wrong information in the application form. Fraudster may provide another person's information to obtain the new credit card with the intension to purchase the product or services and never to repay the amount [57]. On the other hand, behavioral fraud occurs when the fraudsters cleverly obtain the card holder's information and using the information the fraudster can buy products either through the E-commerce channel or through mobile apps where only the card information required [57].

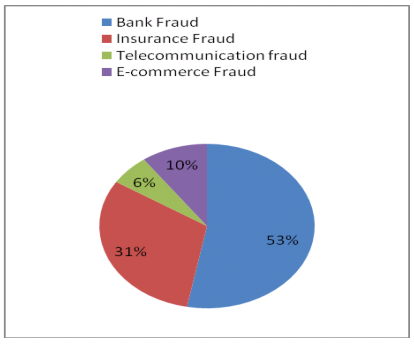


Fig.3. Summary for the quantity of most researched area of fraud

B. *Insurance Fraud*

Insurance fraud occurs at any point during the insurance process, and can be committed any person in the chain, like consumers, agents, insurance company employees, healthcare providers or others. The fraudster claims insurance by providing totally false information. The insurance frauds are classified into three categories, namely health insurance, automobile insurance, and crop insurance. The most common type of claims fraud is automobile insurance fraud, in which the fraudsters create false or intentionally commit accidents, which results in excessive maintenance, and faked injuries. On the other hand, fraud health care insurance takes place using various

methods, like false billing reimbursement, billing for services not rendered, durable medical equipment fraud, Medically Unnecessary Services, Duplicate claims and Kickbacks [28]. Similarly, in crop insurance fraudsters commit fraud by either overstating the loss of their crops due to natural disasters or loss of money due to downfall in the price of agricultural commodities.

C. Securities and Commodities Fraud

Securities and commodities fraud are the various ways by which a person is misguided to invest into an organization based on wrong information. According to FBI [28][46], such types of fraud are Pyramid Schemes, market manipulation, Ponzi Schemes, Advance Fee Fraud, Foreign Exchange Fraud and Embezzlement". According to [22], a securities fraud means theft and fraudulent activities on the market, theft from securities accounts, and wire fraud.

D. Financial Statement Fraud

Financial statements are the company's document that includes company expenses, loans, incomes, profits and losses [56]. It also includes remarks or explanations from management on the business performance and expected problems may arise in the near future [31]. The company releases different types of financial statements which include the overall company's status and it includes the company's success, prices in the stock market, and also it indicates the company is eligible for loans [56]. Financial statement fraud, also known as corporate fraud where the fraudster intention is to show that the company's are making more profit, as a result to improve the stock performance, reduce tax obligations, or to show exaggerate company's performance in order to reduce the managerial pressure [56]. Such types of fraud are very difficult to detect because it was generally committed by managerial people within the industry who are quite capable of hiding their mistakes and another factor is such types of fraud very rarely occurs [75]. According to FBI [28], corporate fraud investigates on the following activities: falsification of financial statements, fraudulent activities committed by corporate insiders and such criminal activities are protected by insiders.

E. Mortgage Fraud

Mortgage fraud is a special type of financial fraud which includes manipulation of a property or mortgage documents. Such fraudulent activity creates an impression on the lender to fund loan for it [46]. According to FBI, Mortgage fraud is defined as material misstatement, misrepresentation, or false statement related to property or mortgage which makes the lender to fund or purchase the property based on the false statement [28].

F. Money Laundering

Money laundering is the process by which the criminals convert their proceeds obtained from illegal business into the stream legal businesses. This hides the origin of the money and creates an impression that the moneys are income through legal way. These types of crimes are very difficult to track. According to Gao and Ye [75], money laundering is defined as the method by which criminals wash dirty money or black money and make it clean and legal.

IV. METHOD TO PROTECT AGAINST FRAUD

With the increase in modern technology and the usage of E-commerce in most of the global businesses, make the fraudsters more attract towards the system. Day–to-day the financial institutions are implementing new methods to protect the system against such fraud, but the fraudsters are using new tactics to perpetuate through the system. As a result, to protect the system against such fraud has become a major issue to be explored [40][45]. As shown in Fig. 4 two mechanisms are used i.e. detection and prevention to fight against such fraud.

A. Fraud Prevention System

Fraud prevention system is the front line of defense to secure the systems against fraud. The purpose of FPS is to halt the fraudulent activities from entering in to system. The mechanism of this layer is to stop, suppress, control, remove, or prevent the occurrence of such attacks. One of such mechanism is encryption algorithm that is implemented to scramble data. Another mechanism is firewall that act as a barrier between the internal network and the external world. This mechanism only protects the outsiders form entering into the system. However, this first phase layer is not always efficient and strong enough to protect the entire system from fraud. [12]. There are so many cases where the fraudsters could able breach the prevention layer.

B. Fraud Detection System

Fraud detection is the next line of defense, if the prevention fails then immediately fraud detection techniques starts its role to identify the fraud as soon as possible. In FDS the system will detect and identify the fraud as they enter the system, then immediately report about them to the system administrator [11]. In the previous years, frauds are detected manually by using discovery sampling techniques [64]. But these techniques are inefficient as it takes more time to detect the fraud in areas like, finance, business, economics and law. Therefore, to increase the effectiveness of this manual approach, a computerized automated FDS technique was invented.

However, automated FDS have limited functionality because its working principle depends on predefined rules as stated by professionals [41]. Recently, more complex and effective FDS are being developed by integrating various data mining techniques for an effective fraud detection system. Data mining techniques are appropriate for fraud detection because to detect fraudulent activities the most important criteria's are the algorithm must be very fast and must be accurate.

Data mining techniques includes statistical, computational intelligence and machines learning techniques which are used to design and implement systems that can identify and extract valuable information and retrieve knowledge from large databases, examples of such systems are decision support systems and intelligent systems. The advantages of such systems are fraud patterns are automatically generated from the data, for each cases the fraud behavior are planned and new fraud behavior are also defined [41]. Data mining techniques are broadly classified into six main approaches, namely Classification, Prediction, Clustering, Regression, Outlier detection and Visualization [54, 55]. Each of these approaches includes specific techniques, such as neural network and support vector machine techniques follows classification approach and K-means technique follow clustering approach. Nowadays, researchers are designing new fraud detection systems by integrating both anomaly based approach and misuse based approach by using different data mining methods [5] [60]. So, FDS are classified into two categories: anomaly based detection and misuse based detection.

Anomaly based FDS perform based on the behavioral approach, the system stores the behavior of each individual unit and if any deviation occurs from the normal behavior, then it indicates a fraud occurs. Anomaly based FDS are very popular, it has been implemented by various authors in different fraud areas [30]. This method can be further classified into three categories i.e. unsupervised, Semi-supervised and supervised [3].

In the supervised learning the data set are associated with output class labeled as "fraud" and "no fraud" and also the model trains the classifier. One of the most important advantage of the supervised learning is that the outputs produced by the algorithm of this approach can be used for pattern classification and data regression. However, supervised learning has got several disadvantages. First, the most important is difficulty in collecting supervision or labels. Mainly when there is a huge volume of input data, it is very difficult to label all of them. Second, it is extremely difficult to find distinct label, there are uncertainties and ambiguities in the labels. Therefore, to overcome these disadvantages in some cases unsupervised learning and Semi-supervised learning approaches are also used [42].
The supervised learning techniques are classified into two categories, namely classification and regression. The algorithms following classification approach are neural network, K-nearest neighbors, decision trees, Naïve–Bayes and support vector machine (SVM).
In unsupervised learning, the instances of the data set are unlabelled. In this learning approach the system determines how much the test data sets are deviates from the normal behavior. Unsupervised learning, the model identify the training data having similar behavior to form a group or cluster. Therefore unsupervised learning is also known as "cluster analysis" and aims to form a cluster which helps to create a model that can classify data instances labels automatically. The main advantage of unsupervised learning technique is that it does not depend on exact identification of the label data [15]. The unsupervised learning techniques are classified into two categories, namely clustering and dimensionality reduction approach. The algorithm following dimensionality reduction approach is Principal Component Analysis (PCA).
Semi-supervised technique follows both supervised and unsupervised approach because it includes both labeled and unlabelled data instances. The working principle of Semi-supervised model is that it can train a classifier using both labeled and unlabeled data samples [30]. Since semi-supervised learning works on both labeled and unlabeled data samples, it shows better performance than supervised learning but this is only applicable for fewer labeled data samples.
In misuse detection approach, a data base was created on fraudulent behaviors and normal behaviors. Misuse fraud detection approach implements rule-based, statistics, or heuristic approaches to detect the happenings of fraudulent transaction [32]. Misuse detection system can identify only known fraudulent behavior present in the database but unable any new kinds of fraud.

V. DATA MINING TECHNIQUES FOR FINANCIAL FRAUD DETECTION

Analyzing and retrieving useful information and converting the information into actionable form is a huge challenge faced by most of the organizations. Data mining is the process of extracting meaningful patterns and rules from huge voluminous data. It plays an indispensable in the field of fraud detection, market segmentation, credit and behavior scoring and benchmarking. This section contains a brief description of different data mining techniques used in the reviewed literature, and Table 2 shows strengths and weakness of each techniques.

A. Bayesian Belief Networks

Bayesian belief networks implements statistical classification approach, it makes use of Bayes theorem to determine the probability that a given hypothesis is true. The Bayes theorem states that for a hypothesis H i.e. a

random variable X can be classified to a particular class, the probability P of the hypothesis is defined as:

$$P(H \mid X) = \frac{P(X \mid H) \cdot P(H)}{P(X)} \quad (1)$$

A Bayesian belief network act as a classifier to evaluate $P(C_i \mid X)$ for all class labels *Ci* and classify the data sample X into the class with highest $P(C_i \mid X)$ .In this way the network can classify each sample X into the class that it is most likely to belongs to [39]. A Bayesian Network is a graphical model, with each node represents a random variable, and the directed edges represent the conditional dependence assumptions.

B. Logistic Rergression or Logistic Model

Logistic Regression(LR) implements statistical method to classify binary data which uses a linear model, also known as a logistic model, to evaluate regression on a set of variables [46][56]. It is a usual method to predict patterns for the data with unambiguous or numeric attributes [14][46]. It is a usual method to predict patterns for the data with unambiguous or numeric attributes [14][46].

Logistic regression by taking input from a series of vectors and a dependent response variable, using the logarithm to determine the probability that the result lies within a particular class. Let $x_i \in R^d$ denote a vector of d dimensions representing the ith data sample, and $y_i \in \{0, 1\}$ represents its corresponding output class. For binary classification, the response variable takes a binary value and using the following formula we can calculate that X_i belongs to class 1:

$$P(Y_i = 1 \mid X_i) = \frac{\exp(w_0 + w^T X_i)}{1 + \exp(w_0 + w^T X_i)} \quad (2)$$

Where, w_0 and w are the regression tuning parameters representing the intercept and coefficient vector respectively [56]. LR is mainly used to handle the fraudulent activities in automobile insurance and corporate fraud.

C. Neural Network

A neural network is a computational model of the human brain which can be represented as a graph where vertices are represented as neurons and edges as synapses [46]. This model consists of interconnected neurons (nodes) which are connected to the next layer of neurons through synaptic weights. Each node takes its input from the previous layer nodes. Each neuron j receives input signal as given below:

$$U_J = \sum W_{IJ} * X_J \quad (3)$$

Where, W_{ij} is the weight associated between neurons i and j and X_i is the input signal to neuron i. If the result is greater than a threshold value, then the neuron j fires and the output of j becomes input for the next layer. Neural networks are widely applied to credit card, automobile insurance and to identify the fraud in corporate.

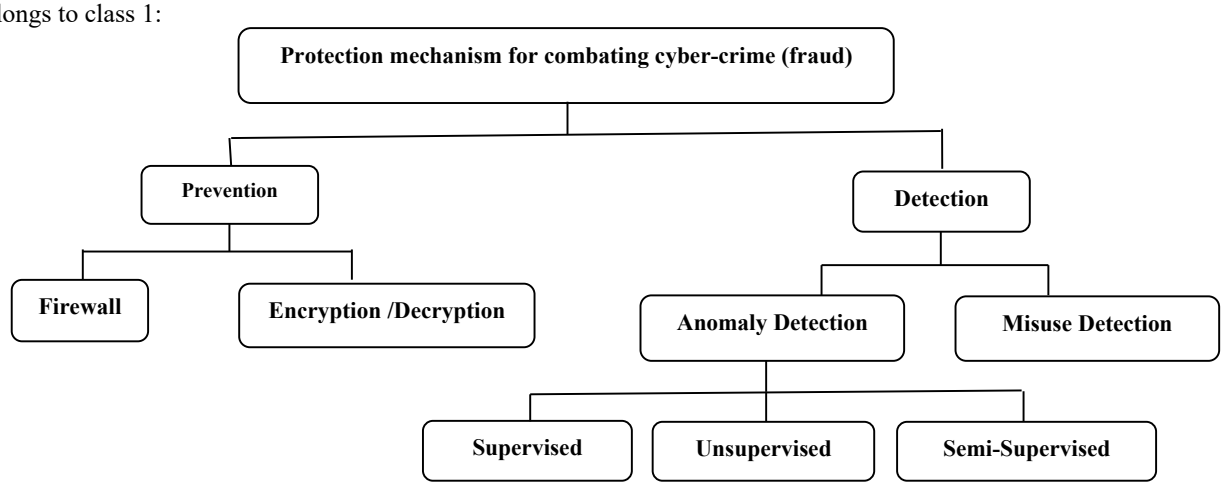


Fig. 4. Protection systems against fraud

Table 2. Strength and weakness of data mining techniques

Technique	Strengths	Weakness
Neural Network	• Performs better results for fraud detection. • Suitable for binary classification problems.	• Requires high computational power for training and testing. • Potential for overfitting if training set is not suitable for the problem domain, so requires constant retraining to adapt to new methods of fraud. • Not suitable real-time data
Logistic model	• Simple to implement. • Performs better results for fraud detection.	• Lower classification performance than other data mining techniques
Bayesian belief network	• Proven suitability with other non-algorithmic, binary classification problems. • High computational efficiency gives potential for real-time data.	• Requires strong understanding of typical and abnormal behavior for the investigated fraud type.
Support vector machine	• Capable of solving non-linear classification problems like fraud detection. • Training and testing requires low computational power and performs well for real-time operation.	• Difficult for auditors to process results due to transformation of input set
Decision trees and Random Forest	• Simple to implement and understand. • Training and operation requires low computational power, which gives potential for real-time operation.	• Potential for overfitting if training set is not a good representation of the problem domain, so requires constant retraining to adapt to new methods of fraud. • Optimisation during initial setup requires high computational power.
Self-Organising Map	• Simple to implement and very easy for auditors to understand given visual nature of results.	• Visualisation requires auditor observation • Cannot be fully automated easily.
Naïve Bayes	• It classifies better than conventional classifier. • It shows better performance in predicting financial fraud • computational time of bayesian network is lower than decision tree	• Predictive accuracy is highly correlated with the assumption of class conditional independence
K-Nearest Neighbor	• It is not necessary to create a predictive model before classification	• Does not generate a simple classification formula • Its predictive accuracy totally depends on the distance and the cardinality K of the neighborhood.
Genetic algorithm	• Simple to implement using classification accuracy as the fitness solution. • Proven suitability with other non-algorithmic, binary classification problems.	• Requires high computational power for training and operation • Unsuitable for real-time function. • Difficulty adapting to new fraud methods due to local maxima/minima problem.

.

D. Support Vector Machine

Support vector machines (SVMs) are statistical learning methods [66] have been successfully implemented in various classification problems. SVMs are very suitable for binary classification problems like fraud detection. SVMs are linear classifiers that coverts a linear problem into in a high-dimensional feature space. The main advantage of SVMs working principle is that the non-linear classification task in the original input space converts to a linear classification task in the high-dimensional feature space. So highly complicated, non-linear problems, like financial fraud detection could be solved by linear classification approach without increasing computational complexity.

Through the kernel function the dataset gets transformed from non-linear input space to higher dimensional linear space. The kernel function is defined as:

$$k(x_1, x_2) = (\phi(x_1), \phi(x_2)) \quad (4)$$

Where, ϕ: $X \rightarrow H$ maps the non-linear input space X to higher dimensional space H and $k(x_1, x_2)$ is the kernel function. After implementing the kernel function to the data sample, a hyper plane is constructed which separate the classes is defined by the following equation:

$$\langle w*x \rangle + b = 0 \quad (5)$$

Where ⟨w*x⟩ represents the dot product of the coefficient vector w and the vector variable x.

SVM uses a hyperplane as a decision boundary to classify the instances of different classes from one another. The main aim of SVM approach is to select the hyperplane that separates the classes in a best way. So the principle of SVM to solve this issue is to select the maximum marginal hyperplane (MMH) i.e. the one with largest margin is considered as the most accurate classifier because it reduces the errors caused due to overtraining. So, the classification equation for a support vector machine can be defined as:

$$\sum \quad \propto_i y_i k(x_i, x) + b = 0 \qquad (6)$$

For classification different kernel functions can be used, but the depending on the classification requirements kernel function is used. But the most commonly used kernel functions are Gaussian radial basis function and polynomial function [14].

E. Decision Trees and Random Forests

Decision trees are a technique that classifies the data set using a tree structure, where the internal nodes are representing binary choices on features and branches representing the result of that feature. Decision trees are also known as Classification and Regression Trees (CART) [39].

Random forest is a collection of decision trees used to overcome the overtraining and instability issues that occurs with a single decision tree [14]. Random forests use different training dataset between trees and randomly select the pool of features while constructing the internal nodes [14]. Another way to reduce the overfitting in decision trees is pruning, which increases the overall accuracy of the tree by removing the irrelevant nodes [39]. These approaches make random forests more robust to overtraining and noise because each tree grows independently, but on the other side there is slight increase in computational complexity [14].

F. Self Organizing Maps

Self-Organising maps are unsupervised neural network which consists of a single matrix of neurons. SOMs are the non-linear algorithms used to map input vectors from a high-dimensional space to the two-dimensional array of neurons. The mapping is done to model similar types input vectors as neurons that are closer to each other in the resulting matrix. The distance among the neurons is measured based on Euclidean distance formula or Gaussian formula [54]. The clustering function applied to each neutron is defined as:

$$\mathbf{Y_{i+1} = Y_i + (X_i - Y_{i-1})} \qquad \mathbf{(7)}$$

where, *Yi* is the current weight of a specific node, *Xi* is the current input vector, and α is the distance function. The clustering done number of times before the algorithm terminates [48].

G. Naïve Bayes

Naïve Bayes classifier uses simple probabilistic rule based on Bayes conditional probability approach. Naïve Bayes implements strong statistical independence assumptions for the predictor variables. Such classifier is easy to implement and mainly applicable if the dimension of the dataset is high. It performs better classification than conventional classifier. It shows better performance in predicting financial fraud with results in no false positives and relatively low false negatives. Naïve Bayes classifiers are widely used in banking and financial institutes to detect fraudulent activities. As compared
to decision tree, though the accuracy of decision tree is far better than naïve bayes but computational time of Bayesian

H. K-Nearest Neighbor (KNN)

K-Nearest Neighbor (KNN) classifiers the dataset based on principle of learning by analogy. When an unknown data sample is tested, KNN classifier searches the data samples that are closest to the unknown sample. Closeness is measured in terms of Euclidean distance. The unknown sample is classified to the most common class among its K-Nearest Neighbor. The main advantage of this technique is that it is not necessary to create a predictive model before classification. The major disadvantages are that KNN does not generate a simple classification formula and its predictive accuracy totally depends on the distance and the cardinality K of the neighborhood.

I. Genetic Algorithm

Genetic algorithms use the concept of evolutionary technique in an iterative way to solve a particular problem. They generate solutions for optimization problems using evolutionary techniques, like selection, crossover and mutation. It randomly generates a population, then repeatedly reproduce each population using various methods and selecting the chromosomes based on their fitness value. Reproduction takes place by selecting pairs of chromosomes from the current generation and then implementing crossover followed by mutation on the single point of the resulting child. The best chromosomes are selected best on the fitness value and the same processes are repeated. The algorithm terminates only after acquiring the required fitness, otherwise the algorithm must terminate after a particular iteration. Genetic algorithms are similar to neural network because they need no prior knowledge about the problem domain and are able to identify the underlying relationships between the data instances [56].

VI. CLASSIFICATION OF EXISTING FINANCIAL FRAUD DETECTION

In this section we will classify the financial fraud detection systems based on success rate, methods implemented, and fraud types to be detected. This section will cover the recent research methods which have been successfully implemented and any factors that have not been included.

A. Classification based on Performance

A number of parameters have been used to measure the performance of the algorithm, but the three most widely used parameters are accuracy, sensitivity, and specificity. Accuracy of the algorithm can be measured based on the ratio of all successfully classified data instances to unsuccessful ones. Sensitivity compares the number of instances correctly treated as fraud and the number of instances incorrectly identified as fraud. It can also be defined as ratio of true positives to false positives. Specificity define ratio of true negatives to false negatives [14][56].

Tables 3 shows the performance of financial fraud detection research based on these three parameters. In addition to these measures several other performance measures have also been discussed in the survey. For example, Duman et al. have shown the results of sensitivity in the form of graph, instead of numerical values [26]. Similarly, some research used software-determined success levels or case-based procedures to measure the efficiency of the FDS [59].

From the performance analysis report we can find that CI methods typically achieved better success rate than statistical methods. Bhattacharyya et al [14] have achieved better results for sensitivity as compared to specificity and accuracy using random forests and support vector machines than logistic regression A combined approaches consisting of genetic programming, support vector machines, probabilistic neural networks, and group method of data handling have achieved better performance than regression in all the three cases [56]. Bayesian belief networks have shown better accuracy than neural network and decision trees [39]. Bermúdez et al., 2008 [13], have combined the Bayesian logic with regression which performed better than logistic regression alone. Additionally, a neural network with exhaustive pruning technology had achieved more specific and accuracy as compared to CDA [44]. A major problem with fraud detection system is misclassification costs and there lies a large difference in misclassification costs i.e. it is much more expensive to classify a fraudulent transaction as legitimate than to classify a legitimate transaction as fraudulent. The CDA and CART methods, as well as the neural networks, Bayesian belief networks and decision trees performed better in this regard, where all the algorithms and methods shows a somewhat higher ability to classify fraudulent transactions than legitimate ones [39][44].

B. Classification based on Detection Algorithm

Classifying fraud detection techniques based on the detection algorithms is a method to identify the suitable and effective techniques for this problem domain. Furthermore, we can also identify the research gaps in the particular algorithms which can be further improved. Table 4 shows different classification of financial fraud detection systems and the techniques used.

Previously most of the researchers were mainly implementing statistical methods and neural networks. However, these methods are still very popular. Neural network are widely used [39][44][56], some are implementing logistic regression [53][14][56][35], while others are adopting Bayesian belief networks and Bayesian algorithms [13][33][24. The techniques like neural network and Bayesian algorithms are widely used as compared to support vector machines and genetic programming. Initially researchers used to implements single algorithm for fraud detection, such as self-organizing maps [54][48], logistic regression [53][68], and fuzzy logic [59]. Nowadays, to achieve better performance they are integrating multiple techniques like classification and regression trees [49], Bayesian belief networks [33], and statistical techniques like Dempster–Shafer theory and Bayesian learning [49].

Since each algorithms has some drawbacks which decreases the efficiency of the detection system, so to overcome these pitfalls the researchers implements the hybrid approaches through which they can combine the strengths of multiple algorithms to classify the data samples. Duman and Ozcelik [26] have implemented the combination of scatter search and genetic algorithm to identify the wrongly classified number of transactions. But Panigrahi et al. [49] have taken different approach, Depster–Schaefer method to generate rules and Bayesian learner to identify the fraudulent activities. Similarly, some researchers have integrated traditional computational intelligence approaches with text mining to identify the fraud in the financial statements [18][36].

C. Classification based on Fraud Type

Each type of frauds has different nature, so different techniques required to identify a particular fraud. By classifying the existing techniques we can identify the techniques more appropriate to handle a particular type of fraud. Table 5 shows the different types of fraud and the technique used to detect the fraud. From the literature survey we can see that majority of the papers have focused on two types of financial fraud i.e. credit card fraud and

financial fraud. Very few researchers have studied on securities and commodities fraud, and insurance fraud.
Different measures are used to evaluate the performance of the FDS but accuracy was the most common used parameter. Table 6 shows the accuracy of each fraud category, as well as comparisons with other methods. Self organizing map had achieved 100% accuracy for credit card fraud detection and other methods like logistic regression, support vector machine, random forests and artificial immune systems had achieved high performance.

For credit card fraud detection system SOM is highly recommended due to its perfect accuracy over 10,000 data samples [48]. The methods implemented to detect financial statement fraud have shown large variance in results, starting from CDA where the accuracy level is 71.4% [17] to neural network where the accuracy level is 98.1% [56]. Several other methods, like Bayesian belief networks, support vector machine, genetic programming, group method of data handling, and some hybrid methods based on text mining had achieved accuracy of more than 90%. As mentioned earlier very few researchers have studied on insurance fraud, they have achieved 60.7% accuracy using logistic regression, while the hybrid method have achieved significant result of 99.5% accuracy[13].

Table 4 – Qualitative analysis of methods researched in existing fraud detection literature

Technique used	**Fraud investigated area**	**Research**
Neural Network	Financial statement fraud	Bose and Wang [17], Kirkos et al. [39], Ravisankar et al. [56]
Logistic model	Credit card fraud Insurance fraud Financial statement fraud	Bhattacharyya et al. [14] Pinquet et al. [53], Viaene et al. [68], Bermúdez et al. [13] Ravisankar et al. [56], Huang [35]
Support vector machine	Credit card fraud Financial statement fraud	Bhattacharyya et al. [14], Whitrow et al. [70] Ravisankar et al. [56], Huang [35]
Decision trees, forests and CART	Credit card fraud	Bhattacharyya et al. [14], Whitrow et al. [70], Sahin et al. [58]
Genetic algorithm/ programming	Financial statement fraud	Ravisankar et al. [56], Hoogs et al. [34]
Self-organizing map	Credit card fraud	Quah and Sriganesh [54], Sánchez et al. [59], Olszewski[48]
Bayesian belief network	Corporate fraud	Holton [33]

Table 3. Accuracy, Sensitivity and Specificity results

Research	Fraud Investigated	Method Implemented	Accuracy	Sensitivity	Specificity
Bhattacharyya et al. [14]	Credit card transaction fraud from a real world example	Logistic model (regression) Support vector machines Random forests	96.6–99.4% 95.5–99.6% 97.8–99.6%	24.6–74.0% 43.0–68.7% 42.3–81.2%	96.7–99.8% 95.7–99.8% 97.9–99.8%
Olszewski [48]	Credit card transaction fraud from a bank in Warsaw	Self-organizing map	100%	100%	NA
Soltani Halvaiee and Akbari [62]	Credit card fraud from a Brazilian bank	Artificial immune system	94.6–96.4%	33.6%–52.6%	97.8–98.1%
Bermúdez et al. [13]	Insurance fraud from automobile insurance claims for a Spanish	Logistic regression	60.68%	85.149%	60.430%
Kirkos et al. [39]	Financial statement fraud from a selection of Greek manufacturing firms	Decision trees Bayesian skewed regression	73.6% 99.538%	75.0% 85.149%	72.5% 99.677%
Ravisankar et al. [56]	Financial statement fraud with financial items from a selection of public Chinese companies	Neural networks Bayesian belief networks Support vector machine Genetic programming Neural network (feed -forward) Group method of data handling Logistic model regression) Neural network (probabilistic)	80% 90.3% 70.41– 73.41% 89.27– 94.14% 75.32– 78.77% 88.14–	82.5% 91.7% 55.43– 73.60% 85.64– 95.09% 67.24– 80.21% 87.44–	77.5% 88.9% 70.41–73.41% 89.27–94.14% 75.32–78.77% 88.34–95.18% 70.66–78.88% 94.07–98.09%
Bose and Wang[17]	Financial statement fraud with financial items from a selection of public Chinese companies	CDA CART Neural network (exhaustive pruning) Support vector machine	71.37% 72.38% 77.14% 71%–92%	61.96% 72.40% 80.83% 76–98%	80.77% 72.36% 73.45% 11–85%

Manuscript received September 5, 2024
Manuscript revised September 20, 2024
https://doi.org/10.22937/IJCSNS.2024.24.9.20

VII. ISSUES AND CHALLENGES

Fraud detection system efficiency can be improved by either increasing accuracy rate or by reducing the false alarms. It is extremely difficult to handle all types of fraud in E-commerce system. FDS deals with multiple issues and challenges, as a result the efficiency of the FDS reduces. In this section, we will focus on the challenges associated with financial fraud detection system.

A. Concept Drift

In credit card transaction, the card holder behavior depends upon the transaction amount and the frequency of transaction. These two parameters changes with time due to person life style and availability of resource. In addition, day to day fraudsters are implementing new tactics to commit fraud, so the detection system must adapt new techniques to deal with these fraudsters [24].

Concept drift mainly deals with online supervised learning scheme, where the relation between the input vectors and the targeted output constantly changes over time. In supervised learning scheme, the model is designed based on the training instances and in this learning the input instances and the corresponding output have been specified. But according to concept drift phenomenon, the relation between the input data and the targeted output may change [29]. Therefore, when the new instance arrives, the model may not predict the output, so this will leads to high false alarms. So, for non-stationary behaviour, the adaptive learning algorithms must be used to handle concept drift issues. So in FDS the concept drift is the major challenge.

B. Skewed Class Distribution

One of the most important issues faced by FDS is skewed distribution or imbalanced class distribution. This type of issue arises when the fraudulent instances are very less as compared to normal instance [44]. In credit card transaction the number of fraudulent transactions are very less as compared to legitimate transaction, so that might reduce the performance of the credit card FDS. So to overcome such issues a balancing mechanism is needed to bring the ratio between normal and fraudulent transaction to 1:1, and this will increase the performance of the FDS. There are two approaches to balance the data set. The data set get balanced either at data level or at algorithm level.

Data level balancing techniques are implemented at pre-processing step to rebalance the instances or remove the noisy data-instances before applying classification algorithm. In FDS environment most of the researchers are implementing data level balancing by using under sampling or over sampling techniques. Sahin and Duman [58], Bhattacharya et al. [14], Phua et al. [52], and Duman et al. [26] implements under sampling techniques to deal with skewed class distribution for credit card transaction. Under sampling technique follows data reduction approach by removing data instances that belongs to majority class [19]. In over sampling technique the data set belongs to minority classes are replicated.

Balancing at algorithm level uses two techniques: cost -sensitive learning and learning by itself to handle skewed distribution. In cost-sensitive learning a cost matrix is used for different types of misclassification error. The cost matrix helps the FDS model to reduce the misclassification cost and maximize the benefit. Sahin et al. [58] implements cost sensitive classifier to handle class imbalance problem.

C. Reduction of Large Data Set

The performance of the FDS can be improved by reducing the data set. The smaller data set will greatly reduce the transaction processing time as well as reduces the complexity. Therefore, the existing fraud detection systems are using different data reduction approaches to reduce the data set [69]. In detection system different data reduction techniques like data compression, feature selection and feature construction are the most commonly and widely used techniques.

Broadly, the existing system follows two reduction approaches, namely, dimensionality reduction and numerosity reduction. Numerosity approach uses aggregate technique, which is a non-parametric approach to capture the consumer buying habits before each transaction and using these attributes it can identify the fraudulent transactions. Whitrow et a. [70], and Dal pozzolo et al. [24] have implemented the numerosity reduction approach.

Balancing at algorithm level uses two techniques: cost - sensitive learning and learning by itself to handle skewed distribution. In cost-sensitive learning a cost matrix is used for different types of misclassification error. The cost matrix helps the FDS model to reduce the misclassification cost and maximize the benefit. Sahin et al. [58] implements cost sensitive classifier to handle class imbalance problem.

D. Reduction of Large Data Set

The performance of the FDS can be improved by reducing the data set. The smaller data set will greatly reduce the transaction processing time as well as reduces the complexity. Therefore, the existing fraud detection systems are using different data reduction approaches to reduce the data set [69]. In detection system different data reduction techniques like data compression, feature selection and feature construction are the most commonly and widely used techniques.

Broadly, the existing system follows two reduction approaches, namely, dimensionality reduction and numerosity reduction. Numerosity approach uses aggregate technique, which is a non-parametric approach to capture the consumer buying habits before each transaction and

Manuscript received September 5, 2024
Manuscript revised September 20, 2024
https://doi.org/10.22937/IJCSNS.2024.24.9.20

using these attributes it can identify the fraudulent transactions. Whitrow et a. [70], and Dal pozzolo et al. [24] have implemented the numerosity reduction approach.

E. Overlapping Data

Overlapping data occurs when the fraudulent transaction is treated as legitimate transaction and vice versa. This is because the fraudsters always trying to make their transactions exactly same as normal transaction, so the system unable to detect the fraud and this leads low detection rate [38]. Therefore, the FDS model should be designed in such a way that it should able to handle overlapping data problem and that might be resolved by implementing a suitable classifier and proper method as [19] in credit card fraud detection.

Table 5-Types of fraud and thetechnique used

Fraud Type	Method used	Research on the type of fraud
Credit card	Support Vector Machine Decision Tree Self Organizing Maps Fuzzy Logic Artificial Immune system Hybrid methods	Bhattacharyya et al. [14] investigated credit card fraud from an international operation. Quah and Sriganesh [54] investigated a banking database from the Singapore branch of a well-known international bank. Sánchez et al. [59] investigated fraud in multinational department stores. Duman and Ozcelik [26] investigated typical consumer spending to determine fraud in a major bank in Turkey. Panigrahi et al. [49] investigated variation in legitimate customer transaction behavior with synthesized credit card data. Wu and Banzhaf [72] investigated automated bank machines and point of sale from an anonymous financial institution. Whitrow et al. [70] investigated credit card transactions from two separate banks. Wong et al. [71] investigated transactions from a major Australian bank. Sahin et al.[58] investigated six months of transactions from an anonymous bank. Olszewski [48] investigated credit card accounts of residents inWarsaw. Soltani Halvaiee and Akbari [62] investigated credit card transactions from a Brazilian bank.
Financial statement	Neural networks Decision trees Bayesian belief networks Support vector machine Genetic algorithms Group method of data handling Logistic model (regression) Text mining Hybrid methods	Kirkos et al. [39] investigated a selection of Greek manufacturing firms. Ravisankar et al. [56], Bose and Wang, [17], and Bai et al., [10] investigated a series of public Chinese companies. Glancy and Yadav [31] and Humpherys et al., [36] investigated managerial statements from official company documents. Hoogs et al. [34] and Cecchini et al. [18] investigated Accounting and Auditing Enforcement Releases authored by a selection of US companies, Huang [35] investigated Taiwanese companies that had been accused of fraud.
Securities and commodities and other corporate	Bayesian belief network Process mining	Jans et al. [37] investigated internal transactional fraud from a successful, anonymous European financial institution. Holton, [33] Investigated emails and discussion group messages to detect corporate fraud.
Insurance fraud	Logistic model Hybrid methods	Pinquet et al. [53], Viaene et al. [68], and Bermúdez et al., [13] all investigated motor insurance claims from Spanish insurance companies.

Table 6-Most successful methods for each fraud type based on the accuracy measure

Fraud Area	Best Performer		Comparative Performers	
	Algorithms	Results	Algorithms	Results
Credit Card fraud	Self-organising map	100%	Logistic regression Support vector machine Random forest Artificial immune system	99.4% 99.6% 99.6% 96.4%
Insurance fraud	Hybrid method	99.5%	Logistic regression	60.7%
Financial statement fraud	Neural network	98.1%	Decision tree Bayesian belief network Support vector machine Genetic programming Group method of data handling Logistic regression Hybrid method Text mining CDA	73.6% 90.3% 92.0% 94.1% 93.0% 79.0% 95.7% 75.4% 71.4%

F. Noisy Data

In Fraudulent transactions the data set tends to be noisy, incomplete and inconsistent. Therefore, most researchers in their model are implementing data cleansing routines to remove noisy data, filling up the missing values and correcting inconsistencies in the data. Many researchers resolve the noisy data issues in the preprocessing stage by filling up the missing values. Such noisy data should be filtered out because it negatively affect the effectiveness and efficiency of the classifier as a result it reduces the predictive accuracy [51]. Baharim et al. [9] have implemented leveraging missing values method in telecommunication area to detect the possibility of corrupted and missing values in CDR using the Naïve–Bayes approach, which led to the identification of usable data instances that lies in rejected CDR.

G. Misclassification Cost

Misclassification cost is the error costs due to false positive and false negative. False positive means authenticate transactions are identified as fraudulent transaction and false negative means fraudulent transactions are identified as genuine transaction. False negative error creates more loss than false positive error [52]. The Misclassification costs are not fixed, they may vary from instance to instance and can change over time. From the FDS review, the misclassification cost issues are resolved by increasing the number of correct classification of fraudulent transactions while reducing the false ones and this reduces the costs of losses. In fraud detection System the primary objective is to minimize the percentage of misclassified transactions. Sahin et al. [58] in credit card area have designed a new cost-sensitive decision tree algorithm which reduces the misclassification costs and the algorithm performs better than the existing well-known methods on the specified problem domain with respect to the well-known performance metrics such as accuracy and true positive rate. Similarly, Duman and Ozcelik [26] have suggested a new approach to handle the misclassification issues by combining meta-heuristic techniques, such as genetic algorithm and scatter.

VIII. CONCLUSION

Fraudulent activities have increased significantly in recent decades, particularly in financial areas. Hence, there is a huge requirement to detect and stop the fraudulent activities. There are two mechanisms to protect the fraud, such as Fraud prevention and detection mechanism. Fraud prevention is not sufficient enough to protect the system against fraud. So Fraud detection is another layer of protection used to protect vital information's in financial systems. In this survey paper we have explored the fraud in different areas. Additionally, we have covered different data mining techniques, like Artificial neural networks (ANN), support vector machines (SVM), Bayesian belief network, decision trees, logistic regression, and genetic algorithms. Either these techniques can be used alone or integrate with other meta-learning techniques to develop a strong fraud detection classifier.

In this article we have classified different financial fraud detection techniques based on performance, detection algorithm and fraud types. Next, we highlight the set of challenges that reduces the performance and efficiency of fraud detection system. Mainly we focus on the issues that hinder the performance, namely, concept drift, skewed data distribution, noisy data and misclassification cost. As a result, the FDS's model becomes extremely complex with weak predictive accuracy. The cost sensitive is the major challenge faced by FDS. The misclassification cost can be due to false positive or false negative. But the false negative error is usually more costly than false positive error. In very few articles the researchers have explicitly focused on cost in their FDS model. In the future research on the FDS using data mining techniques should taken into account cost sensitive misclassification.

References

[1] A Abdallah, M. Aizaini, Anazida Zainal, Farud Detection System, Journal of Network and Computer Applications,ACM Digital Library,Vol-68,Issue C 2016,pp 90-113.

[2] Aisha Abdallah , Mohd Aizaini Maarof, Anazida Zaina, Fraud detection system: A survey, Journal of Network and Computer Applications:68,2016, 90–113.

[3] Akhilomen, John, 2013.Data mining application for cyber credit-card fraud detection system. In Lecture Notes in Engineering and Computer Science. pp.1537– 1542.

[4] Allan, Tareq, Zhan, Justin, Dako, South, 2010. Towards Fraud Detection Methodologies.

[5] Allen, Julia, 2000. State of the Practice of Intrusion Detection Technologies. January.

[6] Almeida, Pedro, Jorge, Marco, Cortesão, Luís, Martins, Filipe, Vieira, Marco, Gomes, Paulo,2008.Supporting Fraud Analysis in Mobile Telecommunications Using Case-Based Reasoning.pp.562–572.

[7] Amor N. B., Benferhat S., Elouedi Z., Naive Bayes vs decision trees in intrusion detection systems, In Proceedings of the ACM symposium on Applied computing; 2004; p. 420-424.

[8] Association of Certified Fraud Examiners, 2002. Report to the Nations on Occupational Fraud and Abuse.

[9] Baharim, Khairul Nizam, Kamaruddin, Mohd. Shafri, Jusof, Faeizah,2008. Leveraging missing values in call detail record using naïve bayes for fraud analysis. In: Proceedings of the 2008 International Conference on Information Networking. January,pp.1–5. http://dx.doi.org/10.1109/ICOIN.2008.4472791.

[10] Bai B, Yen J, Yang X. False financial statements: characteristics of China's listed companies and CART detecting approach. International Journal of Information Technology & Decision Making 2008;7:339–59.

[11] Behdad Mohammad, Barone Luigi, Bennamoun Mohammed, French Tim, Nature-inspired techniques in the context of fraud detection. IEEE Transactions on Systems, Man and Cybernetics-Part-C (Appl.Rev.) 42(6), 1273–1290, 2012.

[12] Belo, Orlando, Vieira, Carlos,2011.Applying User Signatures on Fraud Detection in Telecommunications Networks. pp.286–299.

[13] Bermúdez L, Pérez J, Ayuso M, Gómez E, Vázquez F. A Bayesian dichotomous model with asymmetric link for fraud in insurance. Insur Math Econ 2008;42:779–86.

[14] Bhattacharyya S, Jha S, Tharakunnel K,Westland JC. Data mining for credit card fraud: a comparative study. Decision Support Systems 2011;50:602–13.

[15] Bolton RJ, Hand DJ. Unsupervised profiling methods for fraud detection. Credit Scoring and Credit Control VII 2001;5–7.

[16] Bolton, RichardJ. Hand, David J. 2002. Statistical fraud detection: a review.Stat. Sci. 17(3), 235–255.

[17] Bose I,Wang J. Data mining for detection of financial statement fraud in Chinese Companies. Paper presented at the International conference on electronic commerce, administration, society and education, Hong Kong, 15–17 August 2007.

[18] Cecchini M, Aytug H, Koehler GJ, Pathak P. Making words work: Using financial text as a predictor of financial events. Decision Support Systems 2010;50:164–75.

[19] Chen, Rong-Chang, Luol, Shu-Ting, Lee, VincentC.S., 2005.Personalized approach based on SVM and ANN for detecting credit card fraud. In: Proceedings of the IEEE InternationalConferenceonNeuralNetworksandBrain.Beijing,China.pp. 810–815.

[20] Chen, Rong-Chang,2006. A new binary support vector system for increasing detection rate of credit card fraud. Int.J.PatternRecognit.Artif.Intell.20(2), 227–239.

[21] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995; 20(3):273–97.

[22] CULS, Cornell University Law School, White-Collar Crime: an overview, http://topics.law.cornell.edu/wex/White-collar_crime, 2009.

[23] D. Zhang, L. Zhou, Discovering golden nuggets: data mining in financial application, IEEE Transactions on Systems, Man and Cybernetics 34 (4) (2004) Nov.

[24] DalPozzolo,Andrea,Caelen,Olivier,Borgne,Yann-A.ëlLe,Waterschoot,Serge, Bontempi, Gianluca,2014.Learned lessons in credit card fraud detection from a practitioner perspective. Expert Syst. Appl.41(10),4915–4928. http://dx.doi.org/10.1016/j.eswa.2014.02.026.

[25] Delamaire, Linda, Abdou, Hussein, Pointon, John, 2009. Credit card fraud and detection techniques : a review. Banks Bank Syst. 4 (2).

[26] Duman E, Ozcelik MH. Detecting credit card fraud by genetic algorithm and scatter search. Expert Systems with Applications 2011;38:13057–63.

[27] Edelstein, Herb, 1997. Data Mining: Exploiting the Hidden Trends in Your Data. DB2 Online Mag. 2, 1.

[28] FBI, Federal Bureau of Investigation, Financial Crimes Report to the Public Fiscal Year, Department of Justice, United States, 2007, http://www.fbi.gov/publications/financial/fcs_report2007/financial_crime_2007.htm.

[29] Gama, Joao, Bifet, Albert, Pechenizkiy, Mykola, Bouchachia, Abdelhamid, 2013. A survey on concept drift adaptation. ACM Comput. Surv. 1 (1).

[30] Ghosh, S., Reilly, D.L., 1994.Credit card fraud detection with a neural-network. In: Proceedings of the Twenty-Seventh Hawaii International Conference on System Science. 3, pp.621–630.

[31] Glancy FH, Yadav SB, A computational model for financial reporting fraud detection. Decision Support Systems 2011;50:595–601.

[32] Hand, David J. Crowder,MartinJ.,2012.Overcoming selectivity bias in evaluating new fraud detection systems for revolving credit operations. Int. J.Forecast.28 (1), 216–223. http://dx.doi.org/10.1016/j.ijforecast.2010.10.005.

[33] Holton C. Identifying disgruntled employee systems fraud risk through text mining: a simple solution for a multi-billion dollar problem. Decision Support Systems 2009;46:853–64.

[34] Hoogs B, Kiehl T, Lacomb C, Senturk D. A genetic algorithm approach to detecting temporal patterns indicative of financial statement fraud. Intelligent Systems in Accounting, Finance and Management 2007;15:41–56.

[35] Huang SY. Fraud detection model by using support vector machine techniques. International Journal of Digital Content Technology and its Applications 2013;7:32–42.

[36] Humpherys SL, Moffitt KC, Burns MB, Burgoon JK, Felix WF. Identification of fraudulent financial statements using linguistic credibility analysis. Decision Support Systems 2011;50:585–94.

[37] Jans M, van derWerf JM, Lybaert N, Vanhoof K. A business process mining application for internal transaction fraud mitigation. Expert Systems with Applications 2011;38:13351–9.

[38] Kim,Min-jung,Kim,Taek-soo,2002.A Neural Classifier with Fraud Density Map for Effective Credit Card Fraud Detection.pp.378–383.

[39] Kirkos E, Spathis C, Manolopoulos Y. Data mining techniques for the detection of fraudulent financial statements. Expert Systems with Applications 2007;32:995–1003.

[40] Kou Y, Lu C-T, Sirwongwattana S, Huang Y-P. Survey of fraud detection techniques. In: 2004 IEEE international conference on networking, sensing and control, vol. 2. NewYork: IEEE; 2004. p. 749–54.

[41] Li, Jing, Huang, Kuei-Ying, Jin, Jionghua, Shi, Jianjun, 2008.A survey on statistical methods for healthcare fraud detection. HealthCare Manag.Sci.11(3), 275–287. http://dx.doi.org/10.1007/s10729-007-9045-4.

[42] Liu, Q., Wu, Y., 2012. Supervised learning. Encycl. Sci. Learn.

[43] Lookman Sithic, H.,Balasubramanian T.,2013.Survey of insurance fraud detection using data mining techniques.Int.J.Innov.Technol.Explor.Eng.3(2013), 62–65.

[44] Maes, S., K. Tuyls, B. Vanschoenwinkel, and B. Manderick. "Credit Card Fraud Detection Using Bayesian and Neural Networks." Proceedings of the 1st International NAISO Congresson Neuro Fuzzy Technologies. Havana, Cuba, 2002.

[45] Magalla, Asherry, 2013. Security, Prevention and Detection of Cyber Crimes. Tu- maini University Iringa University College. Cyber Crime. Prepared by Asherry Magalla (LL . M-ICT LAW-10919), Supervised by Dr . Puluru (2013).

[46] Ngai E, Hu Y,Wong Y, Chen Y, Sun X. The application of data mining techniques in financial fraud detection: a classification framework and an academic review of literature. Decision Support Systems 2011;50:559–69.

[47] Noor, N.M.M., Hamid,S.Ha,Mohemad,R.,Jalil,Ma,Hitam,M.S.,2015.A review on a classification framework for supporting decision making in crime prevention. J. Artif. Intell. http://dx.doi.org/10.3923/jai.2015.17.34.

[48] Olszewski D. Fraud detection using self-organizing map visualizing the user profiles. Knowledge-Based Systems 2014;70:324–34.
[49] Panigrahi S, Kundu A, Sural S, Majumdar AK. Credit card fraud detection: a fusion approach using Dempster–Shafer theory and Bayesian learning. Information Fusion 2009;10:354–63.
[50] Pejic-Bach, Mirjana,2010. Invited paper: profiling intelligent systems applications in fraud detection and prevention: survey of research articles. In: Proceedings of the 2010 International Conference on Intelligent Systems, Modelling and Simulation. IEEE, pp. 80–85. http://dx.doi.org/10.1109/ISMS.2010.26.
[51] Philip, Nimisha, Sherly, K.K., 2012.Credit card fraud detection based on behavior mining. TISTInt.J.Sci.Technol.,7–12.
[52] Phua, Clifton, Smith-Miles, Kate, Lee, Vincent, Gayler, Ross, 2012.Resilient identity crime detection. IEEE Trans. Knowl. Data Eng.24(3),533–546. http://dx.doi. org/10.1109/TKDE.2010.262.
[53] Pinquet J, Ayuso M, Guillen M. Selection bias and auditing policies for insurance claims. Journal of Risk and Insurance 2007;74:425–40.
[54] Quah JT, Sriganesh M. Real-time credit card fraud detection using computational intelligence. Expert Systems with Applications 2008;35:1721–32.
[55] Raj S. Benson Edwin, Portia A. Annie, Analysis on Credit Card Fraud Detection Methods. International Conference on Computer, Communication and Electrical Technology – ICCCET 2011, 18th&19thMarch, 2011.
[56] Ravisankar P, Ravi V, Raghava Rao G, Bose I. Detection of financial statement fraud and feature selection using data mining techniques. Decision Support Systems 2011;50:491–500.
[57] Rezaee Z. In Financial statement fraud: prevention and detection. Hoboken, NJ: John Wiley & Sons; 2002.
[58] Sahin, Yusuf, Bulkan, Serol, Duman, Ekrem, 2013.A cost-sensitive decision tree approach for fraud detection. Expert Syst. Appl. 40(15),5916–5923. http://dx. doi.org/10.1016/j.eswa.2013.05.021.
[59] Sánchez D, Vila M, Cerda L, Serrano J-M. Association rules applied to credit card fraud detection. Expert Systems with Applications 2009;36:3630–40.
[60] Sasirekha, M., 2012. A defense mechanism for credit card fraud detection.Int.J. Cryptogr. Inf. Secur.2(3),89–100. http://dx.doi.org/10.5121/ijcis.2012.2308.
[61] Sohl JE, Venkatachalam A. A neural network approach to forecasting model selection. Information & Management 1995;29:297–303.
[62] Soltani Halvaiee N, Akbari MK. A novel model for credit card fraud detection using artificial immune systems. Applied Soft Computing 2014;24:40–9.
[63] Symantec internet security threat report 2016, accessed: 2016-12-29.URL: http://www.symantec.com/.
[64] Tennyson, Sharon, Forn, Pau, 2002.Claims auditing in automobile insurance:fraud detection and deterrence objectives. J.RiskInsur.69 (3),289–308.
[65] The Global State of Information Security Survey 2016, accessed: 2016-12-29.URL ⟨http://www.pwc.com⟩.
[66] V. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[67] Vatsa V, Sural S, Majumdar AK. A game-theoretic approach to credit card fraud detection. In: Information systems security. Berlin, Germany: Springer; 2005. p. 263–76.
[68] Viaene S, Ayuso M, Guillen M, Van Gheel D, Dedene G. Strategies for detecting fraudulent claims in the automobile insurance industry. Eur J Oper Res 2007;176:565–83.
[69] Viaene, S., Derrig, R.A., Dedene, G., 2004. A case study of applying boosting naive bayes to claim fraud diagnosis. IEEETrans. Knowl. Data Eng.16(5),612–620. http://dx.doi.org/10.1109/TKDE.2004.1277822.
[70] Whitrow C, Hand DJ, Juszczak P,Weston D, Adams NM. Transaction aggregation as a strategy for credit card fraud detection. Data Mining and Knowledge Discovery 2009;18:30–55.
[71] Wong N, Ray P, Stephens G, Lewis L. Artificial immune systems for the detection of credit card fraud: an architecture, prototype and preliminary results. Information Systems Journal 2012;22:53–76.
[72] Wu SX, Banzhaf W. Combatting financial fraud: a co-evolutionary anomaly detection approach. In proceedings of the 10th annual conference on genetic and evolutionary computation. pp. 1673-80, ACM. 2008.
[73] Yang W-S, Hwang S-Y. A process-mining framework for the detection of healthcare fraud and abuse. Expert Systems with Applications 2006;31:56–68.
[74] Yue D,Wu X,Wang Y, Li Y, Chu C-H. A review of data mining based financial fraud detection research. In WiCom 2007 international conference on Wireless Communications, Networking and Mobile Computing, 2007. pp. 5519-22, IEEE.2007.
[75] Z. Gao, M. Ye, A framework for data mining-based anti-money laundering research, Journal of Money Laundering Control 10 (2) (2007) 170–179.
[76] Zhang G, Eddy Patuwo B, Hu MV. Forecasting with artificial neural networks: the state of the art. Int J Forecast 1998;14:35–62.
[77] Zhou W, Kapoor G. Detecting evolutionary financial statement fraud. Decision Support Systems 2011; 50:570–575.