# A SURVEY OF ACCOUNTING FRAUD IN PUBLICLY TRADED FIRMS USING A MACHINE LEARNING APPROACH

**Siddharth Nanda[1], Dr. Vinod Moreshwar Vaze[2]**

[1]Research Scholar, Dept. of CSE, Shri JJT University, Jhunjhunu, Rajasthan, India
Email: siddnanda89@gmail.com
[2] Dr. Vinod Moreshwar Vaze, Dept. of CSE, Shri JJT University, Jhunjhunu, Rajasthan, India.
Email: vinod.vaze@gmail.com

**Article History: Received:** 16.04.2023 **Revised:** 03.05.2023 **Accepted:** 17.06.2023

**Abstract:** *Financial statement fraud has been a major trouble for "audit firms, investors, government regulators, and other capital market stakeholders". Moreover, this compromises the trustworthiness of financial markets, CEOs, and even the audit profession. Auditors are confronted with their obvious incapacity to identify large-scale fraud, which may be detected in a variety of ways. The bulk of the suggested techniques for identifying this issue is centered on current algorithms, and they have only tried to detect human or basic data mining techniques that have a large operating cost and are also expensive. So far, the data mining techniques described have either a significant processing cost or a poor level of accuracy. As a result, intelligent financial statement fraud detection technologies have been created to assist stakeholders in taking decisions. Recent studies have shown fraudulent distortion of financial information in management remarks. As a result, the goal of this research was to see whether characteristics obtained from financial data and management remarks in company annual reports might be combined to create a better financial fraud detection system. In this paper, the comparative analysis of various Machine Learning (ML) concerning accuracy, precision, and recall parameters is performed.*

**Keywords:** *Fraud detection, Classification, Machine Learning (ML), Stock market, Boosting learning*

**DOI: 10.48047/ecb/2023.12.Si8.700**

## INTRODUCTION

Financial statement fraud is described as significant misrepresentations or gaps in financial statements caused by a deliberate collapse to inform financial data in conformity with widely recognized accounting standards [1]. Individuals commit financial statement fraud for a range of purposes, according to the Center for Audit Quality, containing individual profit, the requirement to fulfill short-term financial expectations, and a need to cover bad news. Inadequate revenue identification, understatement/overstatement of income, expenditures, resources, or obligations, and misrepresentations in financial statement management analysis and examination are all examples of false financial statements [2]. For companies, governments, and investors, financial statement fraud is becoming a more significant issue. Auditors in certain circumstances seem to be incapable of detecting large-scale fraud. Numerous revelations of fraudulent activities by certain businesses have severely harmed the US financial markets in recent years. Financial statement fraud is jeopardizing the trustworthiness of financial markets, CEOs, and the audit profession. Auditors seem to be incapable of detecting large-scale fraud. Companies that use formal public accounting services are often hit with monetary judgments for hundreds of millions of dollars.

Fraud is a significant problem in auditing since it is frequently linked with attempting to hide, distort, or mislead consumers of audited data and reports. As a result of the collapse of big companies, it is generally recognized that attempts to falsify facts may also occur at the management level [3]. Wireless mobile terminals, which could authenticate the uniqueness of users using passwords, fingerprints, noises, and pictures, have become most Internet transactions. To commit fraud, the fraudster collects user information, for example, ID, password, age, profession, and other details and records for regular users to different trading platforms. This type of deception is extremely frequent nowadays, thanks to the fast advancement of "Information Technology", and it causes significant costs to consumers, companies, and society. Furthermore, conventional information security measures will not be able to prevent online transaction fraud after the identification of the data has been taken. Banks and other financial organizations offer a broad variety of services. However, fraud is common in the financial operations of many of these organizations. More services will produce more user data, increasing the opportunity for criminals to steal customers' data and commit fraud. Fraudulent transactions correctly and quickly have developed a critical financial security issue for the entire financial organizations, involving banks. To identify fraud, a range of ML and Deep Learning (DL) models are progressively utilized. ML has the benefit over conventional rule systems in that it can describe certain financial trends that are hard to find using traditional techniques using a huge amount of complicated data. Neural Networks (NNs), Deep Neural Networks (DNNs), Random Forests (RF), logistic regression, Support Vector Machine (SVM) [4–7], and other models are used to identify financial fraud. Most current models are used to identify credit card fraud and credit card transactions vary through online transactions; therefore, these models aren't completely appropriate for online purchases. The advantages of “NNs and DL” are that they could completely mimic any complicated nonlinear connection have great resilience, and fault tolerance, and can discover optimal solutions quickly. It excels in “image detection [8–10], video processing [11], natural language processing, and a variety of other areas”. However, “NN and DL models” perform poorly when trading with organized data, particularly online transaction data. Because transaction data dimensions are often restricted, certain large characteristics generated from most current models and previous industry knowledge do not contribute to learning [12].

This paper is established in 4 sections. The first section contains a brief introduction to the paper, the second section contains information about financial statement fraud, the third section contains the related work of various authors, the fourth section contains the comparative analysis part, and the final section contains the conclusion and future scope.

### 1.1. Definition of Fraud

Fraud is a broad term that indicates a variety of abnormalities and unlawful actions that are motivated by deceit. According to Webster's New World Dictionary deliberate deceit to encourage a people to give up property or other legal right. A general term for the myriad ways that one person uses deception to gain an unfair advantage over another. This involves every type of surprise, trick, cunning, dissembling, and unfair means by which another person is cheated. According to the legal definition of fraud. Deceit is intentionally performed to obtain an unfair or illegal advantage, according to the American Heritage Dictionary (second college edition). The deliberate use of deception, a trick, or other dishonest method to remove another of his/her/their money, property, or a legal right, according to Black's law dictionary. As a result, fraud may be described as deliberate conduct intended to persuade another individual to offer up a bit of value or relinquish a legal right [13]. Fraud is the intentional distortion or suppression of data with the intent to deceive or mislead. The following three types of fraud indicators may be used to classify fraud indications:

**a) Personal Shortcomings:**

1. People existing outside their means.
2. Elevated revenue of employees
3. Unusual behavior through co-workers or employees

**b) Financial Shortcomings:**

1. Unsolved items in records
2. Abnormally huge numbers of cash transactions
3. Modified insufficient documents.
4. Non-serial number transactions

**c) Operational Shortcomings:**

1. Shortage of interior controls
2. There is just one person in charge, and no tasks are delegated.
3. Supplies and financial records not resolved.
4. Illegal transactions

## 1.2. Fraud Classification

The term “fraud” refers to any intentional attempt to deceive or mislead another person to cause damage or hurt. Depending on the types of offenders, this deliberate, unlawful conduct may be distinguished and described in a variety of ways. Individual frauds, for example, misappropriation or theft, are differentiated from frauds committed by companies or top-level management, “financial statement fraud”. The former is referred to as staff fraud, whereas the latter is referred to as management fraud. Corporate fraud comes in a variety of forms for example Internal vs external fraud is the most obvious difference in fraud categorization. Fraud is outside if the victim is from outside the company; else, it is inside. For example, Internal fraud is defined as fraud perpetrated by workers, internal auditors, executives, the board of directors, and management who might incur a financial or reputational loss. External fraud is defined as deception perpetrated by third parties such as investors, creditors, suppliers, consumers, and external fraud [14]. Figure 1 shows the classification of Fraud.

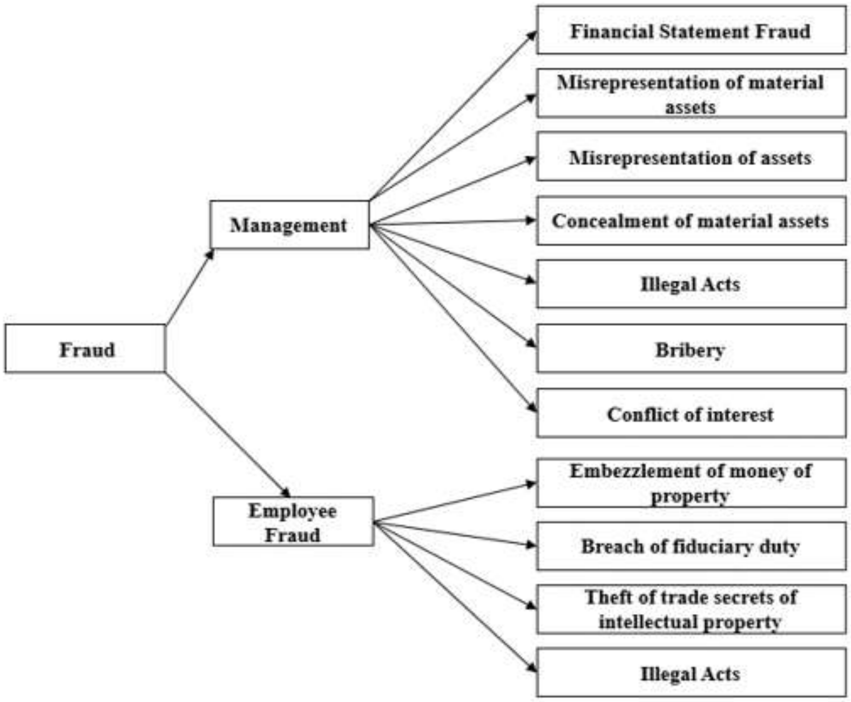
Fraud
Management
Employee Fraud
Financial Statement Fraud
Misrepresentation of material assets
Misrepresentation of assets
Concealment of material assets
Illegal Acts
Bribery
Conflict of interest
Embezzlement of money of property
Breach of fiduciary duty
Theft of trade secrets of intellectual property
Illegal Acts

**Figure 1: Fraud Classification**

### 1.2.1. Management Fraud

Management fraud is defined as intended and fraudulent activities undertaken by a company's senior executives or management team to deceive stakeholders and obtain personal advantage. This kind of fraud often includes the manipulation of financial records, the misrepresentation of firm performance, and hiding of liabilities to provide a misleading picture of the organization's financial health. Management fraud may result in inflated stock prices, investor losses, and reputational harm to the organization. To detect and prevent management fraud, strong internal controls, rigorous audits, and ethical leadership are required.
Example: Enron's executives manipulated financial statements to hide debt and inflate profits, leading to the company's collapse in 2001.
Example: Volkswagen deliberately manipulated emissions tests to falsely represent their vehicles as meeting environmental standards, damaging their reputation and leading to legal consequences.

### 1.2.2. Employee Fraud

The term "employee fraud" refers to any fraudulent activity committed by a firm employee. Workers who commit fraud do so with the intent of enriching themselves financially at the expense of their employer. The perpetrator of employee fraud is often not a high-up in the company's hierarchy. When an employee commits this form of fraud, they do it by bypassing the company's normal procedures and checks.
Example: Company workers steal the supplies by fudging the company's inventory documents.
Motivation: Fraud is committed by workers for financial gain.

## 2. FINANCIAL STATEMENT FRAUD

An accurate definition of "financial statement fraud" is necessary for an in-depth understanding of the nature, significance, and effects of false financial reporting. According to the ACFE, financial statement fraud is "the knowing and willful misstatement of a material fact or the omission of accounting data that would cause readers to change or alter their judgment or decision when all relevant information is known." When the management of a corporation knowingly releases false or misleading financial information, this is called financial statement fraud. A financial statement audit is often performed to check for significant errors or fraud in a company's financial statements. In today's competitive economy, businesses cannot afford to be unprepared to fight fraud [14]. Many people in the corporate sector would want to believe that fraud never occurs.

Management fraud in financial reporting occurs when business leaders intentionally misrepresent the company's financial position. Companies that are publicly traded might be accused of financial statement fraud if they provide materially incorrect financial information to their investors and creditors. When upper-level management falsifies financial figures to make the firm seem better financially, they commit what is known as "management fraud." Financial statement fraud is committed by a cunning and well-informed ring using complex strategies and extensive deception. The following methods can be used to forge financial reports [14]:

1. Fabrication, alteration, or utilization of material business transactions, financial records, or supporting documents.
2. Fraud in financial reporting occurs when there are material misstatements, omissions, or misrepresentations of events, transactions, accounts, or other essential data utilized in the preparation of accounting records.
3. Intentional, ignorant, or improper use of accounting principles, theories, methods, and tools used to ascertain, recognize, and document monetary and business-related activities.
4. Accounting standards, principles, methods, and related financial data that have been intentionally omitted or disclosed, or for which inadequate admissions have been made.
5. The execution of questionable profit management strategies and aggressive accounting techniques.
6. The existing rules-based accounting standards are overly complicated and easy to circumvent, and they also have loopholes that allow corporations to disguise the economic substance of their implementation, which makes them vulnerable to manipulation.

### 2.1. Causes of Financial Statement Fraud

Despite corporate governance or environmental pressure, an organization's management may deploy financial statement fraud as a strategic weapon due to the "Financial Statement Fraud" intrinsic features of loyalty, aggressiveness, control ineptitude, and lack of suitable standards. Below, the fraud triangle is used to explain these traits.

### 2.1.1. Fraud Triangle

Financial statement fraud is a purposeful wrongdoing by senior management of publicly listed businesses. Opportunity, attitude or rationalization, and motivation or pressure are the three most common elements of fraud. The Fraud Triangle is shown in Figure 2 which is made up of these three elements, which may be seen in different ways in the types of a company that engages in false financial reporting. The following are the elements [15]:

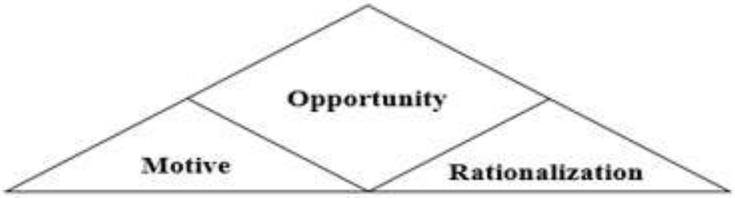

**Figure 2: Fraud Triangle.**

a) An opportunity is a set of circumstances that allows management to make a significant financial statement misrepresentation. Internal control that is poor or non-existent may be a risk factor for financial statement fraud. Inadequate supervision by the board of directors, as well as a complicated organizational structure, all contributed to the lack of a competent audit committee.

b) Rationalization refers to the capacity to behave by one's own moral and ethical beliefs. Criminals that engage in fraud often rationalize their actions to themselves. Management could think about engaging in fraudulent activities related to financial statements as a strategy to get an advantage in the marketplace or accomplish internal goals. Executives at the top of a company may lie to investors and say they are just trying to protect them by manipulating financial reports and the stock price.

c) Motive (incentive) refers to the pressures that management faces when it comes to substantially misrepresenting the financial statement. These pressures may be categorized as “psychotic, egocentric, ideological (belief in the moral superiority of the cause), or economic”. Because of a bad financial situation, a loss of consumers, a decreasing market, and other factors, the management of a company is often pushed to join in fraudulent activities.

### 2.1.2. Flowchart of Financial Statement Fraud

The main goal of companies that engage in false financial coverage is to exaggerate their sales and profits to conceal losses and escape regulatory penalties. The income of such businesses is boosted through fake transactions. Overstating revenue might increase account receivables and/or cash (or vice versa), and since income is the main cause of shareholders' equity, together income and equity might be overstated. When a company chooses to understate costs rather than overstate sales, the impact is almost the same since understating expenses raises cash and/or inventory. Considering the above, Figure 3 depicts a flowchart of financial statement fraud [15].

### 2.1.3. Balance Sheet

A balance sheet is a financial statement that displays a company’s properties, obligations, and shareholder fairness at a convincing point in time. It may be applied to determine rates of return and assess a company’s capital structure. It’s a financial statement that specifies how much money shareholders have endowed what a company owns and owes.

### 2.1.4. Income Statement

The income statement is the most important financial statement used to describe the results of an organization's financial operations over a certain time. The income statement, which is also known as the profit and loss statement or the statement of revenue and expenditure, is a kind of financial statement that details the income and expenses of a business over a certain time frame. Financial Statement of Fraud Process Flow Diagram (Figure 3) [15].

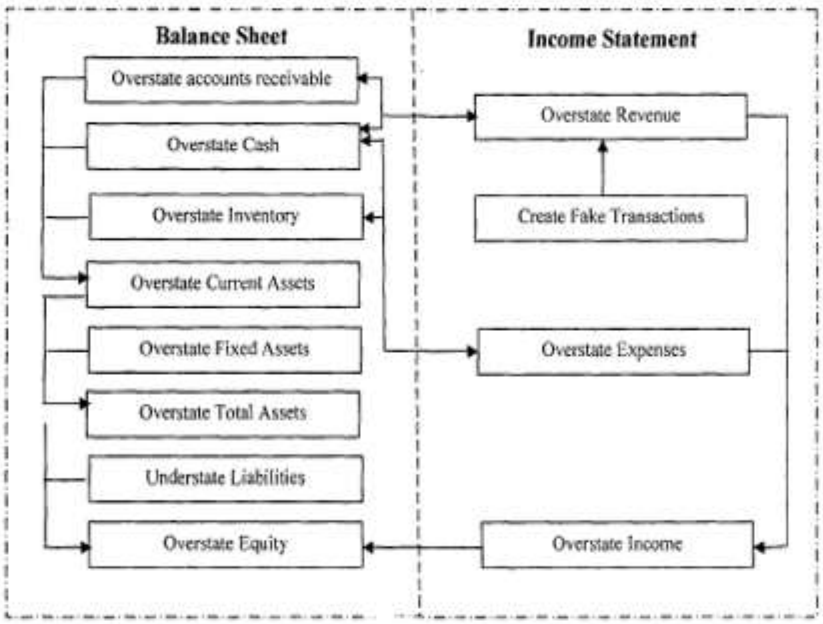

**Figure 3: Flow chart of financial statement of Fraud**

## 3. REVIEW OF LITERATURE

This section discusses and studies the relevant work done by various authors on financial statement fraud.

**Li et al. [16]** stated that with the expansion of the economy comes a new kind of financial theft: credit card fraud, which costs businesses and consumers alike billions of dollars annually. Cardholders and banks were hurt by the credit card fraud case, and the credit management order was put in jeopardy as a result. But fraudsters constantly innovate new methods of deception, which only serves to increase the frequency with which fraud is committed. Credit card fraud detection models that can anticipate future behavior are essential for limiting financial losses. ML is an effective method for identifying fraud since it distinguishes between fraud and non-fraud. SVM has been demonstrated to be a new, high-performing method.

**Priscilla et al. [17]** stated that digitalization has become a significant element in the banking industry as a result of the enormous development of technology. As the number of online transactions rises, so does the rate of fraud. Even though numerous methods exist to detect fraudulent transactions, criminals develop their paradigms. The author aims to provide research on “Credit Card Fraud Detection (CCFD)” by emphasizing the issue of class imbalance and different ML methods, as well as extending the efficient evaluation metrics specifically for CCFD.

**Trivedi et al. [18]** stated that credit card fraud has increased dramatically as a result of contemporary technical advancements and modern communications expressways. Designed to identify credit card transaction fraud seems to be a significant topic in financial analysis of basic economic implications. Credit card fraud results in annual losses of millions of dollars for businesses and individuals. Criminals engaged in fraudulent activities are always on the lookout for novel guidelines and methods. Because of this, fraud detection systems are essential to keeping money in banks and other financial institutions. Throughout the research, the author set up an ML-based system for recognizing fraudulent charges on credit cards, complete with a feedback mechanism. The classifier's detection rate and efficiency could both be enhanced by its feedback approach.

**Sovitkar et al. [19]** stated that the efficacy of transaction fraud detection techniques has a direct impact on user loss in online transactions. But, for low-frequency customers with low transaction volume, current techniques cannot properly characterize each user's transaction activity and result in a high misjudgment rate. As a result, a novel technique for individual behavior creation is suggested, which may improve the accuracy of low-frequency user behavior by migrating the existing transaction group behavior and transaction status recognition model centered on new transaction behavior. The “Naive Bayes model” is applied to compute the likelihood that the existing transaction is fraudulently centered on the results of each behavior, and ultimately to decide if the current transaction is fraudulent.

**Mittal et al. [20]** stated that electronic commerce called e-commerce is a business theory that permits businesses and individuals to purchase and sell items via the Internet. Lots of data are kept and moved from one place to another in the era of the Internet and forwarding to E-commerce. Fraudsters may get access to data that has been transmitted. Fraud is on the increase causing annual losses of billions of dollars across the globe. Various contemporary techniques of detecting fraud are often proposed and used in a variety of business sectors. Fraud detection's primary job is to monitor the activities of many users to identify unwelcome conduct. To identify these different types of assaults, data mining techniques, and ML have been developed and deployed. For example, SVM, K-nearest Neighbor (KNN), NN, fuzzy logic, Decision Trees (DTs), and others, were used for fraud detection systems a long time ago. All these methods have shown acceptable results, but still require to be enhanced by the methods themselves or by employing a hybrid learning method for fraud detection.

**Taha et al. [21]** stated that credit cards have become the possible maximum communal form of payment for both conventional and online purchases resulting in new developments in communication technology and electronic commerce systems. Due to this development, there is substantially increasing fraud connected with these transactions. Fraudulent credit card transactions cost companies and consumers a considerable amount of money each year, and fraudsters are always developing new technologies and methods to commit fraud. The identification of fraudulent transactions has become a major element influencing the adoption of electronic payment systems. As a result, efficient and effective methods for identifying fraud in credit card transactions are required. Applying an improved light gradient boosting machine, the author suggested an intellectual method for identifying fraud in credit card transactions.

**Daliri et al. [22]** emphasized that financial fraud is a major problem because it damages public faith in financial institutions and leads to monetary losses for those same institutions. Since fraud has grown more widespread in recent years, financial institutions have been trying to devise ways to combat it. A great deal of study has gone into developing methods to detect it since fraud strategies have become more complex and varied. The study uses the Harmony Search Algorithm in conjunction with the Artificial Neural Network technique to detect fraud. The proposed method uncovers previously hidden patterns in otherwise indistinguishable sets of normal and fraudulent customer data. Given that fraudulent behavior could be detected and avoided if caught early enough, the results show that the proposed system has sufficient fraud detection capabilities.

**Zheng et al. [23]** stated that AdaBoost is a boosting-based ML technique that assumes the data distribution and input feature space in the testing and training sets are similar. It raises the weight of occurrences that are incorrectly categorized throughout the training phase. The theory does not carry true in numerous real-world data sets. AdaBoost is therefore expanded to “transfer AdaBoost (TrAdaBoost)”, which may successfully transmit knowledge from one domain to another.

**Xuan et al. [24]** claimed that credit card theft is common and causes huge financial losses on a regular basis. Through the use of Trojans and Phishing techniques, criminals could gain access to other people's credit card details. The ability to identify fraudulent purchases made with a stolen card in real-time makes a fraud detection method that much more important. One strategy is to utilize ML techniques to create a model of what constitutes normal/fraudulent behavior based on historical transaction data, which would then be used to validate the legitimacy of a given transaction. Two different kinds of RFs were used by the author for training to learn normal and unusual transaction behavior features.

**Benchaji et al. [25]** stated that financial fraud crimes have risen dramatically in tandem with the growing usage of credit cards, resulting in massive losses in the banking sector. To avoid such losses, all banks must have an effective fraud detection system. Identifying credit card fraud is a huge challenge: data sets for credit card fraud are severely imbalanced because the number of fraudulent transactions is substantially smaller than the number of legitimate transactions. As a consequence, many traditional classification methods for extremely skewed data sets often fail to distinguish objects from the minority class. To improve categorization implementation of the skewed data set's minority of credit card fraud occurrences. The author proposed a sampling strategy based on "K-means clustering" and an evolutionary algorithm. The K-means approach was used to cluster and classify the smaller samples, and the genetic algorithm was employed to collect more data and develop an efficient fraud detection classifier throughout each cluster. Table 1 illustrates a short review of the literature.

**Table 1: Summary of Related Work**

| **Author** | **Year** | **Technique used** | **Outcomes** |
|---|---|---|---|
| **Li et al. [16]** | 2021 | SVM | Losses can be minimized with the use of a predictive model for identifying credit card fraud. |
| **Priscilla et al. [17]** | 2020 | SVM | The author aims to provide research on “Credit Card Fraud Detection (CCFD)” by emphasizing the issue of class imbalance and different ML methods. |
| **Trivedi et al. [18]** | 2020 | Random forest | The detection rate and efficiency of the classifier could be enhanced by the use of the feedback approach. |
| **Sovitkar et al. [19]** | 2020 | Naive Bayes | A novel technique for individual behavior creation is suggested, which may improve the accuracy of low frequency. |
| **Mittal et al. [20]** | 2020 | KNN | The author employs a hybrid learning method for fraud detection. |
| **Taha et al. [21]** | 2020 | Decision Tree | Applying an improved light gradient boosting machine, the author suggested an intellectual method for identifying fraud in credit card transactions “(O Light GBM)”. |
| **Daliri et al. [22]** | 2020 | ANN | The results show that the proposed method can identify fraud effectively enough. |
| **Zheng et al. [23]** | 2020 | AdaBoost | It raises the weight of occurrences that are incorrectly categorized throughout the training phase. |
| **Xuan et al. [24]** | 2018 | Random forest | To train the behavior characteristics of regular and anomalous transactions, the author utilized two types of RFs. |
| **Benchaji et al. [25]** | 2018 | KNN | To avoid such losses, all banks must have an effective fraud detection system. In reality, detecting credit card fraud presents a significant challenge. |

## 4. COMPARATIVE ANALYSIS

### 4.1. Credit Card Fraud Detection (CCFD)

Credit Card Fraud Detection applies the ML technique in which a Data Science team examines information and develops a model to identify and prevent fraudulent transactions. This is done by combining all important elements of cardholder transactions, for example, “the date, user zone, product category, amount, provider, and customer behavioral patterns, among others”. The data is then put into a model that has been trained to examine patterns and rules to evaluate whether a transaction is fraudulent [26]. For the comparison supervised learning methods are carried out and the accuracy, precision, and recall parameters of each method are compared below. The values of these metrics can be calculated using the following formulae.

**i) Accuracy**

The ratio of the total number of values to the sum of the true positive and true negative.

$$Accuracy = \frac{True\ Positive+True\ Negative}{True\ Positive+True\ Negative+False\ Positive+False\ Positive}$$

(1)

TP = True Positive

FP = False Positive

TN = True Negative

FN = False Negative

**ii) Recall (or Sensitivity) and Precision**

It is a True Positive Rate (TPR). It is defined as the proportion of successfully categorized true images to total test images. The recall is determined as follows:

$$Recall = \left(\frac{true\ positive}{true\ positive+false\ negative}\right)$$

(2)

The precision is calculated as

$$Precision = \left(\frac{true\ positive}{true\ positive+false\ positive}\right)$$

(3)

## 4.2. Neural Network

A NN is a “kind of artificial intelligence” that simulates the way the human brain processes and retains information using case-based reasoning and pattern recognition. The fundamental component of this paradigm is a huge number of highly linked processing units known as neurons, which work together to solve particular issues in unison. These neurons are arranged in a few hierarchical levels, with input, hidden, and output layers being the most common. The values are accumulated via applied weights and transformed into an output value by applying an activation function after receiving input through all the neurons in an input layer. The result is then transmitted to the output layer’s neurons, which offer a feed-forward route to the output layer. While training samples are given to the network, an iterative training procedure is used to change the weights between two neurons in two neighboring layers. By comparing fresh information to existing data and identifying hidden patterns within huge data sets, NNs may learn the features of possibly false financial statements. After learning the pattern of input data from sample fraud and non-fraud cases. NNs may evaluate individual data signals to create a unique behavior pattern that classifies input data as fraudulent or non-fraudulent. The resulting pattern is then used to look for signs of financial statement fraud. Table 2 shows the NN frequency based on CCFD.

**Table 2: Neural Network frequency based on CCFD [27]**

| Accuracy | Precision | Recall |
|---|---|---|
| 98.69 | 98.41 | 98.98 |

## 4.3. Support Vector Machine (SVM)

In the context of classification issues, SVM is by far the most widely used and well-known supervised method. The SVM relies on a straightforward linear equation, which is:

$$y = mx + c \qquad (1)$$

Here “m is the gradient and y = c is the value where the line cuts the y-axis. This number c is called the intercept on the y-axis. Key Point”.

Which permits modifying the linear division space. The SVM algorithms can be broken up into linear and nonlinear models. To classify the original data, the linear SVM just cuts it in half along the same linear axis. And if the data cannot be split linearly but can be transformed into a space known as the feature space, where the data domain may be broken linearly to distinguish the groups, we refer to this as the nonlinear SVM [28]. The CCFD-based frequency of the NN is shown in Table 3.

**Table 3: SVM based on CCFD [29]**

| Accuracy | Precision | Recall |
|---|---|---|
| 93.96 | 93.22 | 93.00 |

### 4.4. Naïve Bayes (NB)

The NB algorithm is widely used as a classification tool in data mining. Assuming the qualities are independent of one another inside the specified class, it is reliant on the likelihood of some class. This assumption motivates the need to evaluate multivariate probabilities using the data used for training. Most possible combinations of attribute values do not appear in the training results, or they do not appear often enough to be useful. Therefore, it is not necessary to explicitly estimate each associated multi-variate probability. NB avoids this problem by assuming that the observer has some degree of conditional flexibility. Despite this rigid assumption of freedom, NB is a very competent classifier in many practical contexts [30]. The CCFD-based frequency of the Neural Network is shown in Table 4.

**Table 4: Naïve Bayes based on CCFD [31]**

| Accuracy | Precision | Recall |
|---|---|---|
| 91.62 | 97.09 | 84.82 |

### 4.5. K-Nearest Neighbour (K-NN)

K-NN is the easiest unsupervised ML method. The K-NN method assumes that the existing cases and the new case are similar, and it places the new instance in the most similar category to the existing categories. The K-NN algorithm stores all the accessible data and their classification based on the comparison of a new data point. This implies it can be grouped into a proper suite group by utilizing the K-Nearest Neighbour algorithm as new data develops. The K-NN method can be used at the same time as for Classification for Regression, but it is primarily utilized for problems with classification. The K-NN method is a non-parametric method, which means that the underlying data is not considered. It is often referred to as the lazy learner technique instead of storing the data set, it does not learn with the help of a training data set, and only plays an action on the information set at the time of the classification. The KNN method is presented with a new input data set, and it is kept in a category that is remarkably close to the new data [32]. Table 5 shows the Neural Network frequency based on CCFD.

**Table 5: K-Nearest Neighbour based on CCFD [33]**

| Accuracy | Precision | Recall |
|---|---|---|
| 94.99 | 94.58 | 92.00 |

### 4.6. Logistic Regression (LR)

The LR is the most well-known linear classifier. For a multivariate regression, a link between an independent variable and dependent variables could be formed using logistic regression. As the model of multivariate analysis, Logistic Regression (LR) can be used to infer the presence or absence of a function or consequence given a set of values for a variety of predictor variables. The numerous advantages of the LR machine learning algorithms are through adding a suitable relation function to the normal LR model, variables are either discrete or continuous, and they do not usually have the normal distributions or any combination of both forms. The factors should be numerical in the case of multi-regression analysis, and, in the case of the related statistical model, the arm will fly parallel to the cleaning board, the variables must have a normal distribution, and discriminant analysis. The dependent variable belongs to the binary variable in the current situation that reflects the absence or presence of a landslide. The logistic relation function is applicable when the dependent variable is binary [34]. Table 6 shows the Neural Network frequency based on CCFD.

**Table 6: Logistic Regression based on CCFD [35]**

| Accuracy | Precision | Recall |
|---|---|---|
| 94.84 | 97.58 | 92.00 |

### 4.7. Decision Tree (DT)

In ML, a DT structure is a predictive model that adds observable information about a phenomenon to predictions about the phenomenon's goal value. DT learning is an ML approach for generating a DT from data, and it is the major data mining technique. Each node, which corresponds to a variable, and each arc, which corresponds to a child, represents a potential value for that variable. The values of the variables are represented by the path from the tree's root to the leaf node representing the predicted value of the target variable. The leaves of a tree symbolize clusters, while the branches indicate seasonal combinations of characteristics that result in clusters. Dividing a resource set into subgroups centered on a characteristic value test might be used to learn a tree. In each subfolder created by the separation, this procedure is performed recursively. When the partition is no longer advantageous, or when one class can be utilized to all of the samples in the subclass, the return process is complete [36]. Table 7 shows the Neural Network frequency based on CCFD.

**Table 7: Decision tree based on CCFD [37]**

| Accuracy | Precision | Recall |
|---|---|---|
| 92.88 | 99.48 | 86.34 |

### 4.8. Random Forest (RF)

RF is one of the most popular approaches to classification and regression tasks when it comes to ML algorithms. It is based on the notion of "ensemble learning," in which many classifiers are integrated to better explain and predict a wide range of complex problems. The “RF is a classifier that consists of different DTs on the sub-sets of the dataset and to enhance the accuracy, it assumes the average of the different accuracies of the dataset [38]”. Table 8 shows the Neural Network frequency based on CCFD.

**Table 8: Random Forest tree based on CCFD [39]**

| Accuracy | Precision | Recall |
|---|---|---|
| 99.96 | 96.38 | 81.63 |

## 5. Comparison Based on Accuracy

Table 9 given below shows the comparison of various methods on the accuracy value. Figure 4 shows the comparison graph according to which RF has higher accuracy among all methods.

**Table 9: Accuracy Comparison**

| Methods | Accuracy |
|---|---|
| Neural network | 98.69 |
| Support Vector Machine | 93.96 |
| Naïve Bayes | 91.62 |
| K-Nearest Neighbour | 94.99 |

| | |
|---|---|
| Logistic Regression | 94.84 |
| Decision tree | 92.88 |
| Random forest | 99.96 |

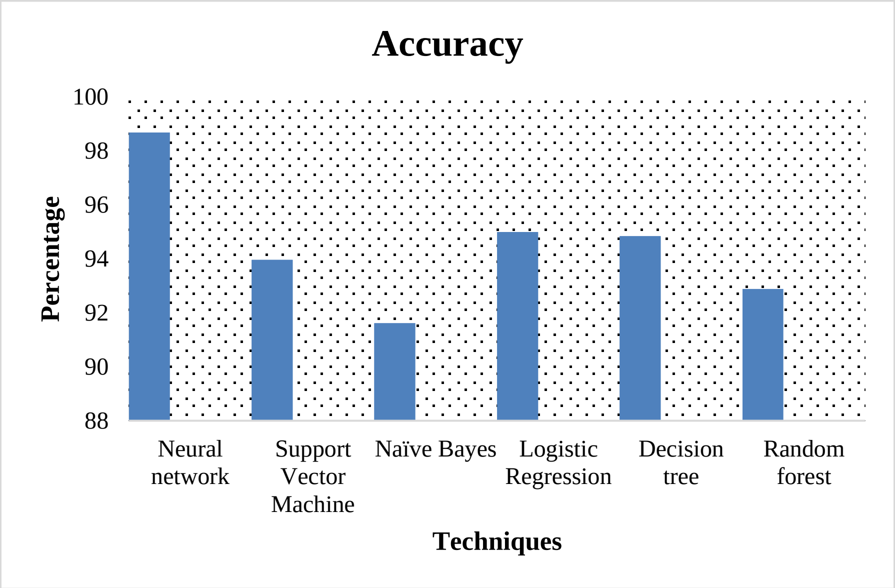

**Figure 4: Accuracy comparison graph**

### 5.1. Comparison Based on Precision

Table 10 given below shows the comparison of various methods on the precision value. Figure 5 shows the comparison graph according to which decision tree has higher precision among all methods.

**Table 10: Precision Comparison**

| **Methods** | Precision |
|---|---|
| Neural network | 98.41 |
| Support Vector Machine | 93.22 |
| Naïve Bayes | 97.09 |
| K-Nearest Neighbour | 94.58 |
| Logistic Regression | 97.58 |
| Decision tree | 99.48 |
| Random forest | 96.38 |

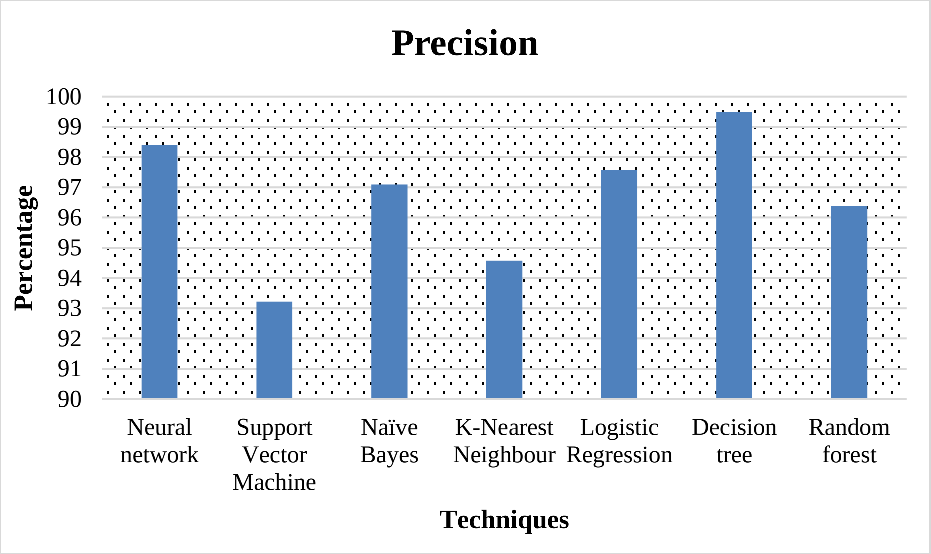

**Figure 5: Graph of Precision**

### 5.2. Comparison Based on Recall

Table 11 given below shows the comparison of various methods on the recall value. Figure 6 shows the comparison graph according to which neural network has higher recall among all methods.

**Table 11: Recall comparison**

| **Methods** | Recall |
|---|---|
| Neural network | 98.98 |
| Support Vector Machine | 93.00 |
| Naïve Bayes | 84.82 |
| K-Nearest Neighbour | 92.00 |
| Logistic Regression | 92.00 |
| Decision tree | 86.34 |
| Random forest | 81.63 |

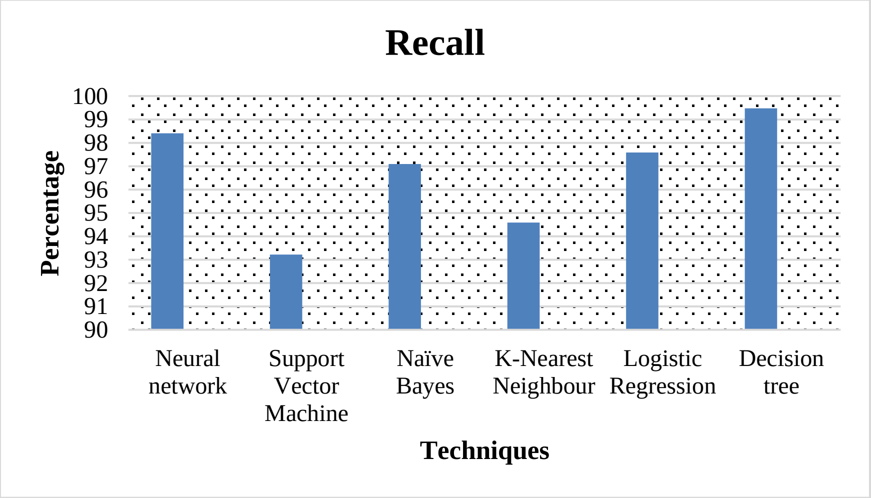

**Figure 5: Recall comparison graph**

## CONCLUSION AND FUTURE SCOPE

Nowadays, knowledge is a significant and strategic resource and an asset for assessment and forecasting which offer solutions for fraud detection. Financial statement fraud is a major issue for today's organizations; therefore, firms are putting a lot of effort into countering it. Accounting fraud is renowned for being difficult to detect. Therefore, finding effective methods for detecting corporate accounting fraud in a timely way, and therefore reducing the amount of fraud-related damage, is a critical area of accounting research.

This paper contains the comparative analysis results of various methods based on parameters, for example, accuracy, precision, and recall. The comparison shows that RFs have higher accuracy, DTs have higher precision values, and NNs have higher recall values. Furthermore, future research must focus on models of trader behavior since this would provide fresh insights into manipulation processes from the viewpoint of trading behaviors. Such knowledge would allow for far faster detection of trade-based manipulations in the stock market than is presently feasible using current techniques.

## REFERENCES

1. Hajek, Petr, and Roberto Henriques. "Mining corporate annual reports for intelligent detection of financial statement fraud–A comparative study of machine learning methods." Knowledge-Based Systems 128 (2017): 139-152.
2. Huang, Shin-Ying, Rua-Huan Tsaih, and Fang Yu. "Topological pattern discovery and feature extraction for fraudulent financial reporting." Expert systems with applications 41.9 (2014): 4360-4372.
3. Javadian Kootanaee, Akbar, Abbas Ali Poor Aghajan, and Mirsaeid Hosseini Shirvani. "A hybrid model based on machine learning and genetic algorithm for detecting fraud in financial statements." Journal of Optimization in Industrial Engineering 14.2 (2021): 169-186.
4. Fu, Kang, et al. "Credit card fraud detection using convolutional neural networks." Neural Information Processing: 23rd International Conference, ICONIP 2016, Kyoto, Japan, October 16–21, 2016, Proceedings, Part III 23. Springer International Publishing, 2016.
5. Bhattacharyya, Siddhartha, et al. "Data mining for credit card fraud: A comparative study." Decision support systems 50.3 (2011): 602-613.
6. Carneiro, Nuno, Gonçalo Figueira, and Miguel Costa. "A data mining-based system for credit-card fraud detection in e-tail." Decision Support Systems 95 (2017): 91-101.
7. Yin, Wenpeng, et al. "Comparative study of CNN and RNN for natural language processing." arXiv preprint arXiv:1702.01923 (2017).
8. Shin, Hoo-Chang, et al. "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning." IEEE transactions on medical imaging 35.5 (2016): 1285-1298.
9. Gurusamy, Ravikumar, and Vijayan Subramaniam. "A machine learning approach for MRI brain tumor classification." Computers, Materials and Continua 53.2 (2017): 91-109.
10. Yuan, Chengsheng, et al. "Fingerprint liveness detection from different fingerprint materials using convolutional neural network and principal component analysis." Computers, Materials & Continua 53.3 (2017): 357-371.
11. Pan, Zhaoqing, et al. "Adaptive fractional-pixel motion estimation skipped algorithm for efficient HEVC motion estimation." ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 14.1 (2018): 1-19.
12. Bahnsen, Alejandro Correa, et al. "Feature engineering strategies for credit card fraud detection." Expert Systems with Applications 51 (2016): 134-142.
13. Javadian Kootanaee, Akbar, Abbas Ali Poor Aghajan, and Mirsaeid Hosseini Shirvani. "A hybrid model based on machine learning and genetic algorithm for detecting fraud in financial statements." Journal of Optimization in Industrial Engineering 14.2 (2021): 169-186.

14. Subramaniyan, Senthilkumar, et al. "Semi-supervised machine learning algorithm for predicting diabetes using big data analytics." Business Intelligence for Enterprise Internet of Things (2020): 139-149.
15. Javadian Kootanaee, Akbar, Abbas Ali Poor Aghajan, and Mirsaeid Hosseini Shirvani. "A hybrid model based on machine learning and genetic algorithm for detecting fraud in financial statements." Journal of Optimization in Industrial Engineering 14.2 (2021): 169-186.
16. Li, Chenglong, et al. "Comparative study on credit card fraud detection based on different support vector machines." Intelligent Data Analysis 25.1 (2021): 105-119.
17. Priscilla, C. Victoria, and D. Padma Prabha. "Credit card fraud detection: A systematic review." Intelligent Computing Paradigm and Cutting-edge Technologies: Proceedings of the First International Conference on Innovative Computing and Cutting-edge Technologies (ICICCT 2019), Istanbul, Turkey, October 30-31, 2019 1. Springer International Publishing, 2020.
18. Trivedi, Naresh Kumar, et al. "An efficient credit card fraud detection model based on machine learning methods." International Journal of Advanced Science and Technology 29.5 (2020): 3414-3424.
19. Sovitkar, Sarika Ashok, and Seema S. Kawathekar. "Comparative study of feature-based algorithms and classifiers in face recognition for automated attendance system." 2020 2nd international conference on innovative mechanisms for industry applications (ICIMIA). IEEE, 2020.
20. Mittal, Sangeeta, and Shivani Tyagi. "Computational techniques for real-time credit card fraud detection." Handbook of Computer Networks and Cyber Security: Principles and Paradigms (2020): 653-681.
21. Taha, Altyeb Altaher, and Sharaf Jameel Malebary. "An intelligent approach to credit card fraud detection using an optimized light gradient boosting machine." IEEE Access 8 (2020): 25579-25587.
22. Daliri, Sajjad. "Using harmony search algorithm in neural networks to improve fraud detection in banking system." Computational Intelligence and Neuroscience 2020 (2020).
23. Zheng, Lutao, et al. "Improved TrAdaBoost and its application to transaction fraud detection." IEEE Transactions on Computational Social Systems 7.5 (2020): 1304-1316.
24. Xuan, Shiyang, et al. "Random Forest for credit card fraud detection." 2018 IEEE 15th international conference on networking, sensing and control (ICNSC). IEEE, 2018.
25. Benchaji, Ibtissam, Samira Douzi, and Bouabid El Ouahidi. "Using genetic algorithm to improve classification of imbalanced datasets for credit card fraud detection." Smart Data and Computational Intelligence: Proceedings of the International Conference on Advanced Information Technology, Services and Systems (AIT2S-18) Held on October 17–18, 2018 in Mohammedia 3. Springer International Publishing, 2019.
26. Sailusha, Ruttala, et al. "Credit card fraud detection using machine learning." 2020 4th international conference on intelligent computing and control systems (ICICCS). IEEE, 2020.
27. Daliri, Sajjad. "Using harmony search algorithm in neural networks to improve fraud detection in banking system." Computational Intelligence and Neuroscience 2020 (2020).
28. Awad, Mariette, et al. "Support vector machines for classification." Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers (2015): 39-66.
29. Najem, Suha M., and Suhad M. Kadeem. "A survey on fraud detection techniques in e-commerce." Tech-Knowledge 1.1 (2021): 33-47.
30. Chen, Shenglei, et al. "A novel selective naïve Bayes algorithm." Knowledge-Based Systems 192 (2020): 105361.
31. Kumari, Priyanka, and Smita Prava Mishra. "Analysis of credit card fraud detection using fusion classifiers." Computational Intelligence in Data Mining: Proceedings of the International Conference on CIDM 2017. Springer Singapore, 2019.
32. Rico-Juan, Juan Ramón, Jose J. Valero-Mas, and Jorge Calvo-Zaragoza. "Extensions to rank-based prototype selection in k-Nearest Neighbour classification." Applied Soft Computing 85 (2019): 105803.
33. Saia, Roberto, and Salvatore Carta. "Evaluating the benefits of using proactive transformed-domain-based techniques in fraud detection tasks." Future Generation Computer Systems 93 (2019): 18-32.

34. Bertsimas, Dimitris, and Angela King. "Logistic regression: From art to science." Statistical Science (2017): 367-384.
35. Ling, Seow Wei, Nowshath K. Batcha, and Rajasvaran Logeswaran. "Machine Learning Model for Predicting Potential Donors Using Logistic Regression." Journal of Applied Technology and Innovation (e-ISSN: 2600-7304) 4.4 (2020): 34.
36. Panhwar, Maryam, et al. "Information technology (IT) application and challenges faced by medical and dental undergraduate students." Rawal Medical Journal 46.2 (2021): 438-438.
37. Makki, Sara, et al. "An experimental study with imbalanced classification approaches for credit card fraud detection." IEEE Access 7 (2019): 93010-93022.
38. Random forest algorithm, javatpoint, (accessed on 24 December 2020) https://www.javatpoint.com/machine-learning-random-forest-algorithm.
39. Patil, Suraj, Varsha Nemade, and Piyush Kumar Soni. "Predictive modelling for credit card fraud detection using data analytics." Procedia computer science 132 (2018): 385-395.