Ingénierie des Systèmes d'Information
Vol. 29, No. 6, December, 2024, pp. 2309-2334
Journal homepage: http://iieta.org/journals/isi

Unveiling Financial Fraud: A Comprehensive Review of Machine Learning and Data Mining Techniques

Rajashekhar K. Rao*, Venkata Naresh Mandhala

Department of CSE, Koneru Lakshmaiah Education Foundation, Andhra Pradesh 522302, India

Corresponding Author Email: krsraohyd@gmail.com

https://doi.org/10.18280/isi.290620

Received: 26 April 2024
Revised: 11 October 2024
Accepted: 25 November 2024
Available online: 25 December 2024

Keywords:
financial fraud, financial statements, fraud detection, machine learning, data mining

ABSTRACT

The financial markets' growing complexity and the exponential growth of data availability have made fraud detection in financial statements a critical and challenging issue. This review article offers a thorough summary of the different approaches and procedures employed in the detection of fraudulent financial statements. It explores traditional statistical methods, machine learning algorithms, and hybrid models, highlighting their strengths and limitations in identifying anomalies and irregularities. The paper also discusses the contribution of artificial intelligence and big data analytics to improving the precision and effectiveness of fraud detection. Furthermore, it underscores the need for robust regulatory frameworks and ethical guidelines to prevent misuse and ensure transparency. The review concludes with a discussion on future research directions, emphasizing the potential of emerging technologies in revolutionizing the field of financial fraud detection.

1. INTRODUCTION

Financial statements are essential instruments that provide important insights into the performance and status of a company's finances. Decisions made by creditors, investors, and other stakeholders are based on them. However, in some cases, financial statements can be manipulated through fraudulent activities, leading to a misrepresentation of a company's true financial condition.

Fraud in financial statements refers to deliberate acts of deception or misrepresentation carried out by individuals within an organization. These fraudulent activities can range from misstating revenues, inflating assets, understating liabilities, or manipulating expenses. The ultimate goal of such fraudulent practices is, too often, to make the company appear more profitable or financially stable than its original state.

While fraud in financial statements can take various forms, it typically involves intentional misstatements or omissions of material information. It is important to note that not all financial reporting discrepancies are indicative of fraud. Sometimes, errors or unintentional mistakes can occur due to accounting errors or misinterpretations of accounting standards.

However, when fraudulent activities take place, they can have serious consequences for both the company and its stakeholders. Investors may make decisions based on false information, leading to financial losses. Creditors may be misled about a company's ability to repay its debts. Moreover, the reputation of the company and its management can suffer irreparable damage.

To combat fraud in financial statements, companies implement internal controls, policies, and procedures designed to detect and prevent fraudulent activities. These measures include segregation of duties, regular internal and external audits, and the establishment of ethical guidelines and whistleblower programs. Additionally, regulatory bodies and accounting standard setters provide important information for making sure that financial reporting is clear and honest.

In recent research, a lot of writers have come up with new models that are very good at finding fraud. The XGBoost model used by Ali et al. [1] did better than Decision Tree (DT), Logistic Regression (LR), AdaBoost, Support Vector Machine (SVM), and Random Forest (RF). Zhou et al. [2] suggested using the Node2Vec graph embedding method along with an intelligent and distributed Big Data approach to find financial fraud on the Internet. The groups of trial results show that the suggested method can make finding Internet financial fraud more effective. In their study, Jan [3] used both recurrent neural network (RNN) and long short-term memory (LSTM) models. The LSTM model was the most accurate, with 94.88% success rate.

In conclusion, fraud in financial statements poses a significant threat to the integrity of financial reporting and the trust of stakeholders. Companies must remain vigilant in implementing effective internal controls to detect and prevent fraudulent activities. Moreover, stakeholders must exercise due diligence and rely on reputable sources of information to make informed decisions about their investments and financial transactions.

2. LITERATURE REVIEW

For reviewing Fraud Detection Techniques in Financial

Statements, we have taken 4 review papers and 31 articles.

Ashtiani and Raahemi [4] have raised 2 questions in their paper which were attempted to answer. The questions reviewed in the paper are:

(RQ1) Which financial statement-related datasets and fraud detection strategies were used in work?

(RQ2) Gaps, Suggestions and tends for further study in this field?

The articles may be found using search terms such as financial statements, fraud, machine learning, artificial intelligence, data mining, according to search algorithms that have been established. The aforementioned keywords have been used to search five well-known digital libraries. Total 187 articles were found, and filtrations were made to remove duplicates and adopted a snowballing method in addition to the earlier automated searches of digital libraries. The completed publications were sorted into groups based on comparable investigations, both in terms of methodology and datasets.

For the question RQ1, the papers considered were between 1995 and 2019 and were limited to conference proceedings and peer-reviewed journal papers. Fewer papers were from conference proceedings, and the majority were journal articles (32%).

As for the question RQ2, the studies confirmed that could provide new paths for next scholars to work in this field. Classification and regression techniques were also extensively used.

Ali et al. [5] used Systematic Literature Review (SLR) method that covers mainly 3 stages, planning review, review conduct, and reporting the review. In the 2nd stage Research Questions are Analysed and reviewed.

1. What are the most common financial frauds addressed by ML approaches?

2. What popular ML-based approaches are used to detect financial fraud?

3. What evaluation metrics are used to detect financial fraud?

4. What are the next research directions, trends, and gaps in the field?

When searched for the related articles 287 articles were found similar and finally 93 articles were selected for review that relates to research questions. The searched articles ranged between 2010 to 2021.

Ali et al. [1] analyzed data from 950 MENA businesses. Manufacturing, technology, telecommunications, energy, real estate, and insurance are some of the sectors that are represented. By identifying fraud on a collection of sample firms, the research seeks to construct an improved model for financial fraud prediction.

During the preprocessing phase, missing values were filled in with within-country mean values instead of deleting them one by one. The data was then normalized using the MinMaxScaler to make sure that all the features were on the same scale. To better understand the dataset, detailed descriptive statistics were gathered, and outliers were found and thrown out using the Isolation Forest method. The Gini importance value for each feature was used to figure out how important it was, and the most important features were chosen based on how well they related to the target variable.

Shen et al. [6] use knowledge graph models in conjunction with conventional features to study correlation information on fraud in financial statements detection. They also learn novel representations that are enhanced with feature embedding of different financial categories (FSFD). Features may be built as knowledge graphs with correlation relations serving as edges and features serving as nodes. Correlation types create feature relations. Research indicates that adding correlation information to financial feature representations enhances SVM and K-NN classification performance, but not as much as decision trees and logistic regression (Kernel).

In the preprocessing methods, fraudulent financial records are gathered, and a set of financial features are chosen to be analyzed. The Pearson correlation coefficient is used to find out how closely two or more financial variables are related. This information is then used to make a feature graph that can be used to spot both fake and real financial statements. Then, the correlations found in both types of statements are used to make a shared latent feature graph. This method finds positive, negative, and irrelevant correlations, which are very important for knowing how different financial factors are related.

Zhou et al. [2] recommended learning and representing topological characteristics in the financial system network into low-dimensional dense vectors using the graph embedding technique Node2Vec. This made it possible for the deep neural network to predict and categorize data samples from a large dataset in an efficient and intelligent manner. This approach provided an intelligent and distributed Big Data method for detecting Internet financial fraud. The strategy processes the massive dataset in parallel using Hadoop clusters and Apache Spark GraphX. The experimental findings demonstrate that by raising accuracy, recall, F1-Score, and F2-Score, the suggested strategy may increase the effectiveness of identifying financial fraud on the Internet.

The goal of this study by Albizri et al. [7] is to investigate, categorise, and synthesise current scholarly work in a variety of scientific fields on data analytics as well as machine learning towards financial fraud detection. A few research questions are generated and answered in this paper.

RQ1: What are the prime research topics and phases covered by data analytics and machine learning research for detecting financial statement fraud?

RQ2: How do researchers propose to investigate financial statement fraud using data analytics and machine learning?

RQ3: How can the findings of RQ2 be presented to guide future academic research?

The search terms were financial statement fraud, financial fraud detection, management fraud, and fraud examination. The above questions were researched in 60 papers. The following research topics were chosen.

•Creating Fraud Detection Models.

•Investigating and Predicting Fraud-Related Factors.

•Analytical Tools for Fraud Prediction.

•Literature review.

For RQ2: These papers are also divided into the following categories based on their research method:

•Analytical (Design Science, Secondary Data Based).

•Behavioural (Survey, Experimental, Field Study).

•Conceptual (Theoretical, Literature Review, Discussion).

Deng [8] use the Nave Bayes classifier to develop an FFS detection model. He chose 44 FFS based on auditing reports and 44 non-FFS based on a particular set of criteria from Chinese listed firms between 1999 and 2002 served as the experiment's training data set. Similarly, from 2003 to 2006, the data set for testing includes 73 FFS and 99 non-FFS. He trains the model with the training data set and then applies it to the testing data set, yielding good experimental results.

One of the preprocessing steps used in the study was data

normalization, which lessened the effect of the different dimensions of factors. This is achieved by adjusting the sample data according to a specific formula to ensure uniformity across the dataset. Additionally, the selection of variables for the input vector is based on prior research and includes 26 candidate financial ratios, which are chosen to enhance the model's predictive capabilities. The study also emphasizes the importance of identifying significant variables that correlate with fraudulent and non-fraudulent financial statements to improve classification accuracy.

Chen et al. [9] suggest a two-step study method that uses both financial and non-financial factors to find an organization's scam early warning system. In this study, the approach includes selecting the data for a stepwise regression analysis, finding the TTF's significant variable after screening, and using that variable as the input for both the logistic regression and SVM. The study compares and looks at the data at the end in order to get better results with FFS classification.

The FFS enterprises from 1998 to 2012 were used as research samples. 66 enterprises were chosen from the Taiwan Economic Journal Data Bank's listed and OTC companies (TEJ). A total of 132 study samples are matched using the one-by-one pair approach, including 66 typical firms.

Chen [10] suggested a number of data mining methods, such as ANN, SVM, BBN, and DT. This investigation investigates OTC and lists Taiwanese businesses that, between 2002 and 2013, released fraudulent financial statements.

The study employs normalization of selected variables as a preprocessing method to prepare data for modeling. This step ensures that the variables are on a similar scale, which is crucial for the performance of many machine learning algorithms. Random sampling without repetition is conducted to create a training dataset, which helps in reducing bias and ensuring that the model is trained on a representative sample of the data. Additionally, rigorous tenfold cross-validation is utilized to test classification accuracy, providing a robust evaluation of the model's performance.

Goel and Uzuner [11] related publications on financial statement fraud detection that use natural language processing (NLP) to analyze textual information. In order to expand the scope of qualitative textual analysis and highlight the importance of emotion in fraud detection, this work looked at sentiments that manifest as linguistic expressions that are observable in written text. There hasn't been much prior study on the function of sentiment in fraud detection. They want to fill this research void by providing an alternative, rational approach to evaluating the qualitative data from annual reports and determining if sentiment is significant for identifying fraud.

Additionally, they selected only the fraudulent businesses that satisfied the following five requirements for the final sample: The claimed fraudulent activity should have an effect on the selected firms' annual reports, or 10-Ks; (2) the selected companies' 10-Ks should be retrieved from the Analysis, Retrieval (EDGAR) database, and Electronic Data Gathering; and (3) the original 10-Ks, not the updated 10-Ks, should be downloaded from EDGAR; (4) the MD&A section of the chosen companies must be included in the 10-Ks; and (5) the chosen companies should have Between 1994 and 2012, this led to the detection of 180 fake firms.

As part of the preprocessing steps, all tables with financial information were taken out of the text of yearly reports' Management Discussion and Analysis (MD&A) sections. Also, all the numerical information that was in the MDAs' text was taken out so that only the qualitative material could be seen. This way of doing things made sure that the research focused on the meaning and language without using numbers, which could change the results. There were 360 MDAs in the final MDA text corpus, which included both fake and real entries.

Huang et al. [12] developed a method to detect and extract FFR topology patterns. In particular, the proposed method uses pseudo- and non-spurious samples to train GHSOM pairs with the same training parameters, and takes advantage of the unsupervised nature of learning and the evolving hierarchical structure of GHSOM to test hypotheses about counter-relationships between groups. This study also presents (1) a topological pattern-based classification rule to detect FFR and (2) a competitive-expert to capture the key features of cheating behavior. Empirical results from 762 144 publicly traded annual financial reports in Taiwan.

The following sources were used between 1992 and 2008 to detect false statements: Major Securities Violations and Indictments via the Securities and Futures Investor Protection Centre, and the lawsuit filed by Tadisian Juwan. Law and regulation search engine.

Before it can be used, the suggested method needs a number of data preparation steps, such as variable measurement, sampling, and choosing which variables are important to feed into the Growing Hierarchical Self-Organizing Map (GHSOM) model. Usually, the matched-sample approach is used in empirical studies that look at fraudulent financial reporting (FFR). There are many tools, like discriminant analysis and logistic models, that can help you choose the important factors. These tools help you find the most relevant features for the GHSOM input.

Kirkos et al. [13] investigated the efficacy of facts Using mining (DM) classification techniques, one may identify items connected to fake financial statements (FFS) and identify businesses that generate them. using statistics Mining strategies to discover control fraud may want to help auditors in their paintings. This observe investigates the detection of fraudulent monetary statements using choice trees, Neural Networks. The input vector consists of financial statement ratios. The 3 fashions' performance is as compared.

The information pattern protected data from seventy-six Greek production firms (no economic organizations had been included). The 38 FFS businesses have been paired with 38 non-FFS businesses. these companies had been categorized as non-FFS because there has been no indication.

The study employed a supervised discretization method for preprocessing, which utilizes class information to improve classification accuracy. This method was chosen after testing various discretization techniques, including equal depth and equal width. Supervised entropy-based discretization is particularly effective as it defines intervals that are more likely to enhance classification performance. Due to software limitations, value discretization was necessary to eliminate the effects of outliers, albeit at the cost of some information loss. This preprocessing step is crucial for the effective application of Data Mining techniques in detecting fraudulent financial statements.

Li et al. [14] discovered financial statement fraud by using data mining algorithms with financial and linguistic elements taken from yearly 10-k filings. Financial statement fraud is detected using a distance weighted discrimination (DWD) model when the sample size is less than the sample dimensions. In terms of generality, this model does well in

HDLSS settings. To improve feature selection and parameter optimization for classifiers such as DWD, Back Propagation. We also used genetic algorithms. The suggested GA-based DWD model showed promise as a tool for identifying fake financial statements as it outperformed existing GA-based classification models in terms of classification accuracy while requiring less input data.

To verify that the data was current and reliable, the authors gathered annual report 10-k filings from publicly traded corporations that were accessible via the Securities and Exchange Commission's (SEC) website's Electronic Data-Gathering, Analysis, and Retrieval (EDGAR) database.

Over 111 companies' 2007 yearly financial reports were obtained from the EDGAR database. They thought that companies that weren't in the AAERs weren't necessarily dishonest. About 21.05 percent of the 57 firms we kept for our project were fake.

The F-score technique is used for filter-based feature selection, measuring the discrimination of features. Data standardization, such as z-score scaling, is applied to handle financial data with different units and scales. The combination of financial and textual features is also utilized to enhance the detection of financial statement fraud.

Wu and Du [15] aim to combine textual information from management comments in the annual reports of 5130 Chinese listed firms with numerical elements extracted from financial accounts to produce a better method for identifying financial fraud. First, we construct a financial index system that incorporates indices that have traditionally been ignored from prior research. The annual filings of Chinese-listed companies are then extracted using word vectors. Strong deep learning systems are then used, and the outcomes are compared with combination, text, and numerical data.

The preprocessing methods for financial fraud detection involve several key steps to ensure data quality and model performance. Data cleaning is performed to remove noise, distortion, or extreme values from the dataset. Missing values are addressed by removing samples with incomplete attributes. Scaling and standardization techniques are applied to bring features to a similar scale, enhancing model suitability. Chinese word segmentation is utilized to transform unstructured text into a numeric format, facilitating algorithmic processing. The Jieba package is selected for effective Chinese word segmentation during text preprocessing.

Andayani and Wuryantoro [16] aim to explain how corporate social responsibility, financial statement fraud detection, and sound company governance affect the decline in false financial statements. The 53 papers that were gathered from scientific journals and subjected to qualitative content analysis yielded results that could not be broadly applied. Quantitative studies usually describe the detection of fraud structure of financial statements through responsive elements in the data to the report frauds detection category system, whereas qualitative studies employ studies to explore more theories regarding fraud alongside additional interdisciplinary concepts over a longer period.

To eliminate the chance of deceptive financial statements, preventive actions are essential. Because there are still opportunities for fraudsters to exploit, the system for good corporate governance, which strengthens governance standards, only sometimes results in a reduction of false financial statements. The governance components subjected to quantitative analysis continue to provide a range of outcomes. In this instance, some areas have managed to put in place an efficient system of good governance to combat fraud, while others have not.

Sabatian and Hutabarat [17] set out to look into the part of the Fraud Triangle that makes it hard to spot fake bank records. The financial records of the Cigarettes and Cosmetics subsectors from the Indonesia stock exchange from 2016 to 2018 were used in this study. This study makes use of requirements-primarily based totally purposive sampling to acquire thirty samples of facts. analysis of records the usage of logistic linear regression. in step with the findings, rationalisation had a significant impact on economic statement fraud. economic stability, outside strain, non-public financial want, financial desires, vain monitoring, and enterprise nature, alternatively, haven't any touching on economic assertion fraud.

This study looked at seven cigarette subsector groups and five cosmetics subsector groups that were listed on the Indonesia stock market between 2016 and 2018. This study uses a method called "purposeful sampling" and has several conditions, such as:

1. Organizations indexed at the IDX inside the cigarette and cosmetics subsectors.

2. The employer became indexed at the IDX 12 months before the research length (2015) and persisted to be listed at the IDX in the course of the studies period (2016-2018).

3. Businesses which have an audited Annual document from 2016 to 2018.

4. The enterprise changed into now not delisted sooner or later of the statement duration.

As samples, the researchers received 18 annual reviews from cigarette groups and 12 annual critiques from beauty organizations. Documentation strategies in conjunction with downloading annual reviews from the website www.idx.co.identity are used to accumulate information. The IBM SPSS Statistic 25 software program software and logistic regression analysis techniques had been used to build up the information.

Sorkun and Toraman [18] examine how data mining techniques may be used to identify fraud in online ledgers using financial information. To do this, data sets were constructed using 72 sample e-ledgers along with a rule-based control system. From these, error percentages were calculated and marked. The tagged e-ledgers created statements of finances that were taught on nine differentiating characteristics using various data mining techniques. The training approach made use of Artificial Neural Networks, Decision Tables, M5P Trees, J48 Trees, Support Vector Machine, K-Nearest Neighbour algorithm, Linear Regression, and Decision Stump. The acquired outcomes are contrasted and explained.

Finding fraud in financial statements is made easier by choosing the right features during the preprocessing step, according to the study. Although there isn't a clear agreement on the best features to use, this shows that different ratios have been mentioned in the research. A rule-based e-ledger control program is used to label financial statements. For rule violations, a fraud score running from 0 to 100 is presented. For successful fraud detection, the preprocessing includes choosing 9 unique features from the bank statements.

Othman [19] conduct an evaluation on contemporary era-primarily based absolutely methods for detecting economic statement fraud. The motive of this paper is to explain the annoying conditions of predicting an extraordinary fraud event and to offer a know-how of the diverse records-mining-

primarily based completely techniques for detecting economic declaration fraud. because con artists are becoming nimbler and are always coming up with a look at gives instructions for future studies in detecting the evolution of fraudulent economic reporting.

Anisykurlillah et al. [20] checked out the real-world evidence that shows how rationalization, industry type, outside pressure, financial goals, and financial security affect financial statement fraud. Don't forget to consider institutional ownership as a moderating factor. The LQ45 was made up of 58 widely traded companies on the Indonesia Stock Exchange in the 2016–2018 population study. With SPSS and deliberate sampling, descriptive and regression analysis were done on 29 companies.

Yadav and Sora [21] develop an improved FSF identification technique for qualitative data in financial reports using deep neural networks. Text is first pre-processed by tokenization, lemmatization, and filtering. The Harris Hawks Optimization (HHO) algorithm is then used to choose the features. Finally, a Deep Neural Network-Based Deer Hunting Optimization (DNN-DHO) algorithm is used to check if there is a fraud report in the financial accounts. Financial statement datasets in a Python environment were used to test the proposed FSF detection algorithm. When compared to conventional classifiers the new technique yields output with a high classification accuracy of 96%. Additionally, it yields superior outcomes across all performance indicators.

The preprocessing methods in text mining include filtering, lemmatization, and tokenization. Filtering involves removing unnecessary characters and stop words from the text, which are common words that do not contribute significant meaning.

Lemmatization breaks down words into their base or root form, which helps make the text data more consistent. Tokenization breaks the text into smaller components, such as phrases, words, or symbols, which are known as tokens. These preprocessing steps are crucial for enhancing the quality of the data before applying further analysis or classification techniques.

Omeir et al. [22] use two well-known fraud detection models created by Beneish [23] and Dechow et al. [24]. This article compares the predictive accuracy of financial fraud for Iranian firm statements between these two models. They start by attempting to determine the statistical description linked to the Beneish [23] and Dechow et al. [24] models' first and fourth quartiles. The t-test and variance analysis are then used using SPSS software to assess the forecasting capabilities of the models. They conducted an 11-year study, from 2009 to 2019, on 197 firms.

Zhao and Bai [25] propose a new method for detecting and forecasting financial fraud among publicly traded groups based totally on system studying. They accumulated 18,060 transactions as well as 363 monetary signs, 362 economic variables, and a category variable. They first ignored nine indicators that had nothing to do with economic fraud before processing the missing data. Subsequently, they identified 13 indicators out of 353 indicators that significantly impact financial fraud. These indicators are only dependent on a few distinctive selection patterns and the frequency of functions occurring across all algorithms. After that, the researchers developed three ensemble models, five single classification models, and ensemble styles with a voting classifier to anticipate economic fraud facts of publicly listed agencies. These models included the LR, RF, XGBOOST, SVM, and DT. After that, they chose the best single version out of five gadget learning algorithms as well as the enjoyable ensemble version from all hybrid models. The most helpful model parameters have been identified by comparing many version assessment metrics and using the grid seek technique.

The preprocessing methods involved handling missing values and outliers in the dataset. Indicators with more than 70% missing values were deleted, while those with 40% to 70% missing values were filled with 0. For indicators with less than 40% missing values, the mean filling method was applied. After preprocessing, 240 effective indicators were extracted, ensuring no missing values remained. Standardization was used to scale the data between 0 and 1 to mitigate accuracy issues due to varying scales of indicators.

To discover fraudulent monetary statements, Indrati and Claraswati [26] employ the idea of fraud diamond. To stumble on monetary announcement fraud, the modified Jones model is used. The changed Jones version is used to calculate the enterprise's accumulated income and receivables from credit income. The authors use the receivables ratio as a proxy variable from the character of the enterprise on this observe, ensuing inside the modified Jones version being the excellent research version for detecting monetary statement fraud. This study's population includes all assets and actual estate sector businesses that had been listed at the Indonesia inventory trade among 2015 and 2019. From 2015 to 2019, the sample consisted of 20 companies in the belongings and actual property sectors indexed on the Indonesia stock exchange (100 organization information with a 5-yr observation length). For statistical techniques consisting of more than one linear regression and speculation checking out, SPSS version 26 is used. economic stability, goal, and auditor exchange don't have any effect on economic announcement fraud, in line with this observe. meanwhile, fraudulent economic statements are inspired via external pressure, enterprise nature, and general accruals.

Humphrey et al. [27] analyze the conceptual framework for financial statement fraud detection used by the Fraud Pentagon. The study investigates the connection between fraud pentagon and financial statement fraud. In order to gather relevant and current information on the elements of the fraud pentagon pertinent to financial statements fraud, the current study used a library research technique. Based on the literature search, financial statement fraud is significantly impacted by the following components of the fraud pentagon: Competence (differences in the number of independent directors on the board and the company's directorship); Arrogance (frequent CEO images and CEO duality); Pressure Opportunity (industry nature and ineffective monitoring); and Rationalization (auditor switching a This paper suggests that the fraud pentagon should be supported by legislation to combat financial statement fraud.

Malik [28] combines three different machine learning algorithms with two feature selection techniques (correlation and wrapper) using data from the Auditor General of India between 2015 and 2016 to find the best algorithm for distinguishing between fraudulent and non-fraudulent firms. Machine learning was implemented via a data science life cycle. The results and analysis demonstrated that machine learning algorithms significantly outperformed traditional fraud detection methods.

Alvadain et al. [29] present a new machine learning method to predict financial fraud. Various machine learning classifiers are fed transaction level features from a synthetic database of 6,362,620 transactions. Correlations between different

characteristics were also investigated. An adversarial system of conditional generation is also used for Table data to generate an additional 5000 data samples (CTGAN).

The database, titled "Synthetic Financial Database for Fraud Detection," was obtained from the Kaggle database.

The study used a synthetic set of financial transactions with 6,362,620 samples, with a big difference between transactions that were false and those that were not. To fix this problem, a Conditional Generative Adversarial Network for Tabular Data (CTGAN) was used to make about 5,000 examples of fraudulent deals. Next, the dataset was divided into two groups: training and testing. The ratio of training to testing was 70:30, which meant that 70% of the data was used for training and 30% for testing. Heatmaps and scatter density plots were used to look at the patterns of association between features and get a better understanding of the data.

Rabade [30] examine and compile the corpus of information about the identification of intelligent fraud in business financial accounting. This study looks at several datasets being considered data mining methodologies. This research illuminates how to select the best approach for various types of datasets while keeping speed, accuracy, and cost in mind.

The primary objective of the study conducted by Saleh et al. [31] was to present scientific evidence regarding the correlation between causes of fraud and false financial statements. The research delved into extensive details on the utilization of Altman's z-score and Dechow f-score in detecting false financial statements among Jordanian industrial owners. Spanning from 2015 to 2019, the study considered false financial statements, while the nature of the firm, external pressure, financial security, and financial objectives were considered separate fraud elements. To assess the theories proposed in the study, a methodological model employing a multiple regression procedure was employed for research analysis. Although certain triangle fraud variables were found to have no association with fraudulent financial reports, other factors exhibited a strong correlation with fraud.

Chukwuma et al. [32] look for signs of accounting fraud that may be used to identify businesses that are more likely to falsify financial statement reporting and support the investigation of risky organizations. It is suggested to use a thorough forensic data analysis technique that incorporates all necessary phases of a data-driven methodology. To identify accounting fraud, the method uses machine learning models, logistic regression models, and financial ratio analysis. Because machine learning models are accurate, comprehensible, and inexpensive, they are very helpful in identifying fraudulent activity.

The preprocessing methods for the study include data cleaning and pre-processing to ensure accuracy and readiness for analysis. This involves removing any missing or duplicate data and correcting any errors in the financial statement data collected from publicly available sources. These steps are crucial as they prepare the dataset for subsequent analyses, such as financial ratio analysis, logistic regression modeling, and machine learning modeling, which rely on high-quality data to produce reliable results. Proper preprocessing enhances the effectiveness of the detection of accounting fraud by ensuring that the data used is accurate and relevant.

Jan [3] conducted a comprehensive analysis of both financial and non-financial data obtained from TWSE/TEPx listed companies spanning from 2001 to 2019. Their study encompassed a sample of 153 companies, out of which 51 were found to have engaged in financial statement fraud, while the remaining 102 companies were free from such fraudulent activities. To construct effective models for detecting financial statement fraud, the researchers employed two robust Remarkably, the empirical findings consistently demonstrated the superior performance of the LSTM model across all performance indicators. Notably, the LSTM model achieved an impressive accuracy rate of up to 94.88 percent, making it the most widely utilized performance indicator in this context. The study used a fake set of financial transactions with 6,362,620 samples, with a big difference between transactions that were false and those that were not. To fix this problem, a Conditional Generative Adversarial Network for Tabular Data (CTGAN) was used to make about 5,000 examples of fraudulent deals. Next, the dataset was divided into two groups: training and testing. The ratio of training to testing was 70:30, which meant that 70% of the data was used for training and 30% for testing. Heatmaps and scatter density plots were used to look at the patterns of association between features and get a better understanding of the data. The study used a fake set of financial transactions with 6,362,620 samples, with a big difference between transactions that were false and those that were not. To fix this problem, a Conditional Generative Adversarial Network for Tabular Data (CTGAN) was used to make about 5,000 examples of fraudulent deals. Next, the dataset was divided into two groups: training and testing. The ratio of training to testing was 70:30, which meant that 70% of the data was used for training and 30% for testing. Heatmaps and scatter density plots were used to look at the patterns of association between features and get a better understanding of the data.

The study employs data normalization and standardization as preprocessing methods, utilizing the MinMaxScaler to scale features within a range of 0 to 1. Random data splitting is conducted to create training and test datasets, with 75% of the data used for training and 25% for testing. The independent and dependent variables are defined, with the dependent variable being a dummy variable indicating financial statement fraud. The data matrix is reshaped to meet the input requirements of the models, ensuring compatibility with the RNN and LSTM architectures.

Cheng et al. [33] create a financial statement fraud detection model that considers missing values and imbalanced classes. First, missing values are removed using pairwise and listwise deletion. Second, it uses nonlinear distance correlation to find relevant qualities and proposes three combined attribute selection techniques. Third, it makes use of both oversampling and under sampling to rectify the unequal class distribution. Lastly, a collection of practical rules is produced by using rule-based classifiers. In actuality, a list of dishonest businesses is used in this research to get data on financial statement fraud.

The preprocessing methods utilized in the study include handling missing values and addressing imbalanced classes. Two techniques for handling missing values were employed: listwise deletion and pairwise deletion. Listwise deletion removes all instances with missing values, while pairwise deletion retains more instances by only deleting those with missing data for selected attributes. To address the imbalanced class problem, the study applied under sampling and oversampling techniques. Oversampling increases the number of records in the minority class, while under sampling reduces the majority class to achieve balance.

Antawirya et al. [34] studied the fraud pentagon's parts to find examples of false financial statements. The companies that were studied were all listed on the Indonesian Stock

Exchange in the financial field between 2015 and 2018. After deliberate selection of the study sample, the data were looked at using multiple regression analysis. The study's findings support the fraud pentagon theory by showing that genuine financial statements can be found using parts of the model.

The fraud pentagon idea is examined by Ariyanto et al. [35] in relation to financial statements that are false. There is no way to test every piece directly. But there are other options. The pressure component is substituted with a personal financial necessity. The industry is starting to have an opportunity-driven character. The external auditors' attributes, together with the director transition, all point to competence and rationalisation. One measure of hubris is the frequency with which CEOs appear in pictures. Testing was done on pharmaceutical businesses that were listed on the Indonesian stock market between 2015 and 2019. They chose the samples using a purpose-driven sampling approach. Panel data regression is used to analyse the data. The analysis's conclusions indicate that some features of the sector have a favourable effect on false financial reporting. A sign of dishonest financial reporting may be a shift in senior management roles, such as directors. The falsified financial accounts of Indonesian pharmaceutical businesses are unaffected by factors such as the number of CEO appearances in images, personal financial need variables, or the calibre of external auditors. For the purpose they have used Dechow et al. [36] and Skousen et al. [37] models.

Ye et al. [38] evaluate many classification models for identifying financial statements that are fake. (FFS). Samples from earlier research were not handled realistically particular area. Thus, to learn unbalanced data, Random Forest is utilised in combination with SMOTE sampling. Some other useful performance measures are also provided. The experimental dataset spans the years 2007–2017 and includes 11726 publicly accessible Chinese financial declarations, 1314 of which the CSRC suspected of being fraudulent. The Random Forest approach is outperformed by Artificial Neural Networks (ANN), Logistics Regression (LR), Support Vector Machines (SVM), CART, Decision Trees, Bayesian Networks, Bagging, Stacking, and Adaboost.

The study uses SMOTE (Synthetic Minority Over-sampling Technique) to prepare the data. This technique mixes over-sampling and under-sampling to make the classifier work better. By setting k=5, random seed=1, and N=500 based on the sample imbalance ratio of 1:7.9, a new training sample of 7884 fraudulent and 10412 non-fraudulent instances is made, giving the system a fraud ratio of 1:1.32. In addition, the mean is used to replace missing numerical attributes, and the mode is used to replace missing word attributes. This works better than Listwise Deletion.

Sawangarreerak and Thanathamathee [39] examined financial ratios from Thai Stock Exchange financial statements for possible fraud trends using financial ratio separation that were connected to false financial statements were found by them. This research is unique in that it examined the probability of fraud for every financial item using mathematics. Six financial elements connected to fraud were also found by this study: Gross profit is the first, primary business revenue is the second, primary company revenue in relation to total assets is the third, capitals and reserves are the fourth, and capitals and reserves are the fifth. Five is the ratio of long-term debt to total capital and reserves; six is the ratio of accounts receivable to main company income. Three additional financial elements set this research apart from others.

The dataset was processed using RapidMiner Studio version 9.8, applying binning discretization methods to classify the data. Two types of binning were utilized: equal width and equal frequency, with varying numbers of bins (3, 5, and 10 for equal width; 3, 5, and 8 for equal frequency). The comparison of these methods aimed to determine the most suitable approach for identifying associated patterns in financial items. Ultimately, binning with equal width using five bins was deemed appropriate for detecting fraudulent financial statements.

Each of the models this paper talks about have its own pros and cons. It's easy to understand and explain linear regression, and it can be adjusted to keep it from overfitting. Using stochastic gradient descent, it is also easy to add new data to linear models. When there are non-linear relationships, linear regression doesn't work very well. They aren't naturally adaptable enough to handle more complicated patterns, and it can be hard and take a long time to add the right interaction terms or polynomials.

Decision trees can learn connections that aren't linear and can handle outliers pretty well. In real life, ensembles work really well and have won a lot of traditional (not deep learning) machine learning events. Individual trees that are not limited can keep growing until they remember the training data, which can lead to overfitting. Ensembles, on the other hand, can help with this.

At the moment, deep learning is the best way to do things in some areas, like computer vision and speech recognition. Deep neural networks are great at processing picture, audio, and text data, and batch propagation makes it easy to add new data to them. The number and structure of their layers can be changed to fit different kinds of situations, and the fact that they have hidden layers makes feature engineering less necessary. Because they need a lot of data, deep learning algorithms aren't generally good for general-purpose tasks. When it comes to standard machine learning problems, tree ensembles usually do better than them. In addition, they take a lot of computing power to learn and a lot more knowledge to tune (set the architecture and hyperparameters).

This means that each training report is saved by nearest neighbours algorithms, which are "instance-based." After that, they guess what will happen with new observations by looking for the most similar training observations and adding up their scores. These methods use a lot of memory, don't work well with data that has a lot of dimensions, and need a meaningful distance function to figure out how similar two sets of data are. Most of the time, it's better to use your time to train regularized regression or tree ensembles.

Regression based on logic The outputs make sense in terms of probability, and the method can be tweaked to keep it from fitting too well. With stochastic gradient descent, it's easy to add new data to logistic models. When there are multiple or non-linear choice boundaries, logistic regression doesn't work as well as it could. They aren't adaptable enough to easily show relationships that are more complicated.

There are a lot of different kernels for SVMs that can be used to model decision boundaries that are not linear. Also, they don't tend to overfit, especially in spaces with a lot of dimensions. But SVMs use a lot of memory, are harder to tune because choosing the right kernel is so important, and don't work well with bigger datasets. Random forests are generally better than SVMs in the business world right now.

Naive Bayes (NB) models work fairly well in real life,

especially for how simple they are, even though the assumption of conditional independence rarely holds true. They are simple to use and can grow as your information does. Because they are so simple, NB models are often beaten by models that have been properly trained and tuned using the methods above.

K-Means is by far the most famous clustering algorithm. It's quick, easy to use, and surprisingly flexible if you prepare your data ahead of time and add features that are useful. It won't always be easy for the user to say how many groups they want. Also, K-Means will not make good clusters if the real underlying clusters in your data are not spherical.

The best thing about hierarchical clustering is that the groups aren't always supposed to be round. It also works well with bigger numbers. You have to pick the number of groups (i.e., the level of the hierarchy to "keep" after the algorithm is done), just like with K-Means.

DBSCAN doesn't depend on globular groups, and it can handle more users. Also, not every point has to be put into a cluster. This makes the clusters less noisy, which could be a weakness based on your use case. The user has to change the hyperparameters 'epsilon' and 'min_samples,' which determine how dense the groups are. You can tell a lot about these hyperparameters by how DBSCAN works.

3. PROBLEM STATEMENT

The financial markets and the economy as a whole are in danger because of financial statement theft, which includes dishonest financial reporting and taking assets without permission. Fraudulent activities still happen, even though auditors and governing bodies are trying to stop them. Often, investors and other stakeholders lose a lot of money. Traditional ways of finding fraud, like reviewing documents by hand and relying on expert opinion, might not be able to spot patterns and oddities of fraud in a business world that is becoming more complicated and globalized.

Machine learning and data mining are two potential ways to make it easier to spot fraud in financial statements. These methods can look at a lot of financial data, find small trends and oddities, and get better at finding things over time by learning from their mistakes. But using machine learning and data mining to find fraud in financial statements comes with some problems. These include the fact that fraud data isn't always balanced, financial reporting standards are hard to understand, and fraud detection models need to be clear and easy to use.

Research questions:

1. How can methods like machine learning and data mining be used to find patterns and oddities in financial statements that aren't normal?
2. What kinds of financial factors and features are most useful for using machine learning and data mining to find fraud?
3. How can the problems of uneven data, uneven class sizes, and idea drift be fixed in fraud detection models based on machine learning?
4. How can we make machine learning models better at finding fraud in financial records by making them clearer and easier to understand?
5. How well do different machine learning and data mining methods compare when it comes to finding fraud in financial statements? What are their weaknesses?
6. How can forensic accounting, ongoing auditing, machine learning, and data mining be combined with each other to make it easier to spot fraud in financial statements?

Objectives:

1. Create and test models that use machine learning and data mining to find false financial reports and asset theft.
2. Use machine learning and data mining to find out how well different financial traits and variables work for finding fraud.
3. Find ways to deal with the problems of uneven data, uneven class distribution, and idea drift in machine learning-based fraud detection models.
4. Improve how clear and easy to understand machine learning models are for finding fraud in financial records.
5. Compare how well and how poorly different machine learning and data mining methods find fraud in financial statements.
6. Check out how machine learning and data mining can be combined with other methods to make it easier to spot fraud in financial records.

4. RESEARCH METHODS

In their pursuit, Ashtiani and Raahemi [4] endeavored to answer RQ1 by introducing different deceitful exercises that tended to utilize ML methods in view of the chosen articles. As indicated by the articles explored, false monetary exercises can be comprehensively named Visa, contract, budget report, and medical care extortion.

Numerous specialists have concentrated on monetary extortion locations involving ML techniques for RQ2. Examples of such techniques include SVM, ANN, HMM, KNN, Decision Tree, the Fuzzy Logic-based method (FL), the Bayesian model (BL), the Genetic Algorithm, the Ensemble method, clustering-based methods, and logistic regressions.

A few presentation assessment measurements for RQ3 have been utilized lately by different scientists, including exactness, accuracy, review, F1 measure, bogus negative rate (FNR), a region under the bend (AUC), particularity, etc. Table 1 shows the recurrence with which different methods are utilized in the articles picked.

Table 1. Types of financial fraud

Fraud Type	Description	Technique Used	Reference
Fraud_ on_ Financial_ Statements	In order to make the firms seem more favourable, their accounting records have been fraudulently manipulated in this corporate scam.	Support Vector Machine Clustering based method Decision Tree Logistic Regression Naïve Bayes Artificial Neural Network	[20]

Table 2. Methods used in different articles

Techniques	Short Description	No. of Articles
SVM	A method of categorization that is linear in nature.	10
HMM	To create more sophisticated random processes, a dual embedded random process is used.	8
ANN	A multilayer network with a functioning mechanism akin to human cognition.	10
Fuzzy Logic	A line of reasoning suggesting that thought processes are imprecise and approximate.	5
KNN	Data is categorised based on the closest and most comparable classifications.	7
Decision Tree	A classification technique combined with a regression tree for decision assistance.	5
Genetic Algorithm	It looks for the most effective strategy to resolve issues pertaining to the recommended fixes.	3
Ensemble	Meta algorithms are prediction techniques that integrate many intelligence techniques into one.	8
Logistic Regression	They are mostly used in classification tasks involving binary and many classes.	8
Clustering	Similar occurrences are grouped into sets using an unsupervised learning technique.	6
Random Forest	Techniques for classification that work by merging many decision trees.	7
Naïve Bayes	A classification method with the ability to forecast group affiliation.	11

Table 3. Methods with research gaps and future study

Paper No.	Methodology	Algorithms Used	Research Gaps / Future Study
1	Ensemble methods	XGBoost, AdaBoost, RF, SMOTE, LR, DT, SVM	•Decentralized model •Non-financial attributes
2	NLP	Node2Vec	•GraphSage, PinSage
3		RNN, LSTM	•need for incorporating additional deep learning algorithms, such as DNN, DBN, CNN, CDBN, and GRU
4	Refer Table 1		•Clustering Methods •Anomaly Detection Approach (Unsupervised Approach) •Artificial immune systems (AIS) •Genetic algorithms (GA) •Word2Vec, Doc2Vec, and BERT •Oversampling methods •Sentiment and semantic analysis
5	Refer Table 2		•Oversampling and under sampling techniques •Word2Vec, Doc2Vec, and BERT •Drift
6		SVM, KNN, DT, LR, NB	•Poor performance of Naïve Bayesian classification
7	NODE2VEC		•lack of exploration into inductive graph embedding network algorithms, such as GraphSage and PinSage
8		Naïve Bayes classifier with other models like logistic regression, decision trees, and neural networks.	• lack of exploration into qualitative factors, such as auditors' qualifications or board composition
9	Machine learning Ensemble methods	Regression, Neural networks,SVM, Decision Tree, Bayesian network, Clustering, Stacking, bagging, boosting Generic programming, EGB2	•Under sampling •Machine learning •Artificial intelligence •Classification algorithms
10	Naïve Bayesian classifier	Decision Trees (DT), Bayesian Belief Networks (BBN), Support Vector Machines (SVM), and Artificial Neural Networks (ANN)	•Need for a two-stage statistical treatment and the use of tenfold cross-validation
11	NLP	$\chi 2$, Information gain, SVM	•Intensifiers (diminishers)
12	GHSOM		•need for further exploration of alternative methods like SVM
13		Decision Trees, Neural Networks, and Bayesian Belief Networks	•gap in existing literature compared to other areas like bankruptcy prediction
14	DWD mdel	genetic algorithm	
15		RNN, LSTM, GNU	•gap in the consideration of non-financial factors
16	qualitative content analysis		•Increase sample size
17		IBM SPSS Statistic 25 software and logistic regression	•inconsistent results regarding the influence of factors such as Financial Stability, External Pressure, Personal Financial Need, and Effective Monitoring
18		Linear Regression, ANN, KNN, SVM, Decision Stump, M5P Tree, Random Forest and J48	•Use Deep learning methods
19		regression, neural networks, decision trees, and Bayesian belief networks.	•need for innovative detection techniques that can effectively address emerging financial scams
20		Regression analysis	•suggesting that a longer duration data set could yield more accurate results
21		Harris Hawks Optimization (HHO), Deer	•lack of focus on qualitative data

		Hunting Optimization (DNN-DHO)	
22	Beneish [23] and Dechow et al. [24] models.		•More usage of Beneish [23] and Dechow et al. [24] models
25	SMOTE, Machine Learning	LR, RF, XGBOOST, SVM, and DT	•Increase sample size •Use multiple sectors
26	Fraud Diamond	Modified Jones Model	•Increase sample size •Measure rationalization
27	Fraud Pentagon		•More usage of Fraud Pentagon
28	Machine Learning	Naïve Bayes, Support Vector Machine and Random Forest	•Increase sample size •Use hybrid model
29	Machine Learning	XGBoost	•Increase fraudulent transactions
30	Machine Learning	decision trees, logistic regression, K-Nearest Neighbour, and Random Forest	•Add additional factors •Unsupervised, semisupervised techniques •Textual and auditory data
31	Fraud Triangle	Altman's z-score, Dechow f-score	•Sample size and sample time •Implement Fraud Triangle
32	Data Analytics	Random Forest, Support Vector Machine, Neural Network	•Hybrid techniques •Deep learning, ensemble methods •NLP •Use of data analytics techniques
33	Deep Learning	RNN, LSTM	•Use Deep Learning algorithms •Consider non-financial variables
34	Fraud Pentagon		•Scope of other sectors •Use other factors
35	Data regression techniques	Chow test, Hausman test, and Lagrange Multiplier test	•pharmaceutical industry is particularly prone to fraud, necessitating further exploration of this sector's dynamics
38	SMOTE	Random Forest, Artificial Neural Network (ANN), Logistics Regression (LR), Support Vector Machines (SVM), CART, Decision Trees, Bayesian Networks, Bagging, Stacking and Adaboost.	•Use non structured data •Consider different datasets
39	Data Analytics	Association rule mining, FP growth algorithm, discretization	•Use clustering methods •Implement other datasets

Ali et al. [5] used the SLR method, this is a thorough process for gathering and examining all research that addressed certain research issues. It is used to find and compile data that is focused on certain concerns in order examine the reasoning behind the results and judgements of reviewers. Review reporting, review execution, and review planning are the three main phases that this SLR research addresses.

For RQ1 only Financial Statement Fraud is taken for consideration and the result is displayed in Table 2.

A lot of researchers have looked into how to use ML methods for RQ2 to find financial fraud. Artificial Neural Networks (ANN), Decision Trees, Genetic Algorithms (GA), Fuzzy Logic (FL), Support Vector Machines (SVM), Hidden Markov Models (HMM), K-nearest neighbours (KNN), Bayesian models (BN), Decision Trees, and Logistic Regression are some of these. Table 3 lists the methods that were used and how often they were used.

Previous studies have shown the effectiveness of SVM, NN, DT, and LR in the job of fraud detection, according to Ali et al. [1]. The models SVM, DT, and LR were selected to compare with XGBoost. Since Small along with medium-sized data sets get the greatest results from decision tree-based algorithms, they selected the RF and AdaBoost algorithms from the ensemble techniques. The comparison phase techniques were tested, and the models were trained using the Scikit-learn Python package.

Extreme Gradient Boosting (XGBoost) is the main technique used in the study to find Financial Statement Fraud (FSF). The Synthetic Minority Oversampling Technique (SMOTE) is used to improve the representation of the minority class and fix the mismatch between classes. For comparison, different machine learning methods are used, such as AdaBoost, Random Forest, Decision Tree, Support Vector Machine, and Logistic Regression (LR). Putting these methods into practice is done in Python, with the Scikit-learn tool being used for training and testing models.

Shen et al. [6] used the Pearson's correlation coefficient to analyse correlation information among financial parameters. Knowledge graphs use entities and their interactions to represent information. They postulate that the link between financial elements in knowledge graphs has statistical qualities, and that the hierarchical information in these graphs may have an impact on the identification of financial statement fraud. They start by randomly choosing financial characteristics and generating vector space representations for every financial statement. Once they have made a shared latent feature network with correlation links, they use the ANALOGY method to learn how to embed financial features. They use matrix-vector multiplication to project the financial records onto the feature latent space as the third step.

To find financial statement fraud (FSFD), this study uses both traditional machine learning classifiers and knowledge representation learning methods. We use five standard models to compare them: naive Bayes with kernel (NB), Support Vector Machines (SVM), k-nearest neighbour (K-NN), Decision Trees (DT), and Logistic Regression (LR). This model, called ANALOGY, learns how to embed financial features from a shared latent feature graph that is made up of correlations between financial features. Using the Pearson correlation coefficient to measure correlation and make feature graphs for both fraudulent and non-fraudulent financial statements is part of the approach.

Zhou et al. [2] proposes an intelligent and distributed bigdata approach to increase the effectiveness of Internet

financial fraud detections. One of the biggest Internet financial service providers in China provided the first trial dataset. The dataset includes 192586 data samples, including 4375 fraud samples, after pre-processing. More than sixty data columns make up the dataset, including beginning amount, currency, income level, payment history, balance sheet, sale status, and more. The machine learning algorithms Node2Vec, DeepWalk, and SVM are compared in experiment groups, and the outcomes are assessed.

The paper proposes an intelligent and distributed Big Data approach for Internet financial fraud detection, which includes four main modules: data preprocessing, normal data feature extraction, graph embedding, and prediction module. The graph embedding module utilizes the Node2Vec algorithm to learn and represent the topological features of vertices in the network graph into low-dimensional dense vectors. The approach is implemented on Apache Spark GraphX and Hadoop to process large datasets in parallel, enhancing the classification effectiveness of deep neural networks. The methodology emphasizes the use of structural equivalence and homophily properties of Node2Vec for better feature representation.

Albizri et al. [7] used a variety of machine learning techniques, such as neural networks, support vector machines (SVM), Bayesian belief networks, decision trees, and regression-based algorithms, to identify fraud.

The study methods are put into four main groups in the paper: analytical, behavioural, conceptual, and design science. A large number of the papers (72) used analytical methods, mostly statistical analysis with secondary data or the design science study paradigm. Surveys, experimental schemes, and field studies were all examples of behavioural methods. Theoretical discussions and book reviews were all part of conceptual methods. The study aims to summarize what has already been written, find gaps, and suggest a framework for further work in the area of detecting fraud in financial statements.

To test how well the model could classify things, Deng [8] used ten variables. Six of these variables had high values in both the training and testing data sets. The FFS classification accuracy rates are 100% with ten variables and with six variables. The average accuracy rates with six variables are 95.45% and with ten variables they are 86.36 percent. The FFS classification accuracy rates in the testing data set are 89.04 percent with six variables and 83.56 percent with ten variables. The average accuracy rates with ten variables are 92.44%, and with six variables they are 90.70%. The results of the experiments show that the variables in Table 1 are very important and that the Nave Bayes classifier is good at recognizing FFS.

The financial and nonfinancial variables are tested for using as input variables in logistic regression and SVM by Chen et al. [9] using the stepwise regression screening approach. The research then looks at all model training and assessment techniques. Lastly, the research evaluates the benefits and limitations of the categorization correct ratio and provides suggestions in light of the analytical findings. Stepwise regression is used in the study to filter the variables and keep the research variables that have the most impact. The three models—Decision Tree (DT), Logistic Regression, and Support Vector Machines—have their prediction accuracy computed (SVM).

The study uses a mixed method that combines several ways to predict fake financial statements (FFS). To find the most important factors, it uses stepwise regression. These variables are then fed into logistic regression, support vector machine (SVM), and decision tree (DT) models. Out of these, the decision tree C5.0 model does the best job of classifying, with a 93.94% success rate. It then moves on to the logistic regression and SVM models, which are 83.33% and 78.79% accurate, respectively. Through variable screening and model comparison, the method tries to improve the percentage of correct classifications.

In this work, Chen [10] used tenfold cross validation, which is often regarded as cautious in the academic sector, along with a two-stage statistical treatment (data mining methods are applied). DT is utilised to determine which variables are the most significant and representative since this research has a greater number of variables. In this work, the DT variable selection programme is SPSS Clementine, and the variable selection tools are CART and CHAID.

Decision Trees (DT), Bayesian Belief Networks (BBN), Support Vector Machines (SVM), and Artificial Neural Networks (ANN) are some of the data mining methods used in the study to find fake financial records. A two-stage statistical treatment method is used, which is thought to be safer than earlier one-stage methods. Tenfold cross-validation is used in the study to make the classification models more accurate. The Chi-squared Automatic Interaction Detector (CHAID) and Classification and Regression Trees (CART) are also used to choose variables in the study.

Goel and Uzuner [11] study uses natural language processing (NLP) and machine learning to look at the role of emotion in finding fraud. Techniques for gathering information and rating features were used to figure out which traits are most crucial for finding scam. To automatically find mood traits, they used a lexicon-based method, which is also sometimes called a dictionary-based method.

A methodology for supervised machine learning called SVM was used to create a fraud sentiment classification model, using the Waikato Environment for Knowledge Analysis (WEKA) platform the fraud emotion classification model was made. They also used a linear kernel and the sequential minimal optimization method to make the SVM available in WEKA.

By dividing the total number of categorized occurrences by the number of properly identified instances, the accuracy of the fraud sentiment classification model was determined.

Sentiment lexicons are lexical resources containing words with their associated semantic orientation. Sentiment lexicons provide pre-made sets of words with labels indicating sentiments. which may be used to match input words in the text to establish their semantic orientation. They evaluated a number of sentiment dictionaries, lexicons, and tools for sentiment analysis before choosing just those which were either robust across numerous domains or specialized to a particular domain for automated sentiment feature detection.

They use the financial emotion lexicon by Loughran and McDonald (henceforth, LM), which is specialized to our industry and contains terms that are often seen in financial material. Next, I retrieved sentiment characteristics from genuine and fraudulent MD&A corpora for each of these LM categories using DICTION.

They explore the function of more complex linguistic features in identifying sentiment features that can be used to differentiate between genuine and fraudulent MD&As.

Natural language processing (NLP) and machine learning are used in the study to look at how people feel about fraud

identification. It uses part-of-speech (POS) and sentiment lexicon-based features to tell the difference between fake and real Management Discussion and Analysis (MD&A) parts. To find the most useful features for finding scam, feature ranking methods like chi-squared (χ^2) and information gain are used. We use the WEKA platform to build a support vector machine (SVM) classification model and a 10-fold cross-validation method to check how well the model works. The study looks at how people feel about things based on their polarity, subjectivity, and passion.

Huang et al. [12] suggested a unique use of unsupervised neural networks for dichotomous data (fraudulent and non-fraudulent, for example) efficiently. In order to investigate the dual GHSOM technique is suggested for general pattern detection of dichotomous data. The non-fraud-central geographical hypothesis is addressed first, followed by the fraud-central spatial hypothesis, using two dual GHSOMs.

•The sum of the variances between a node's weight vector and all of the transmitted input data is known as the node's mean quantization error, or MQE.

•Train each individual map: During each map's training phase, input data is loaded one at a time. The winner is the node that has the least distance between its weight vector and the freshly provided input data after all nodes' distances are computed.

•Horizontally expand every single map: Every map expands independently until, by Eq. (1), a new, separate node across it and its neighbor in the column or row if the stop condition is not fulfilled.

$$\text{Avg(MQE)} < \tau_1 \times \text{MQE}_p \quad (1)$$

•Extend or abolish the hierarchical framework: The following layer will be developed using the node whose MQEi is bigger than $\tau 2 \times$ MQE0. The hierarchy continues in this manner until every leaf node satisfies the stop requirement given in Eq. (2).

$$\text{MQE}_i < \tau_2 \times \text{MQE}_0 \quad (2)$$

Data pre-processing activities like variable measurement, sampling, and the identification of important variables to serve as GHSOM input variables must be completed before to putting the suggested strategy into practice. Traditionally, matched-sample designs are used in FFR empirical investigations. There are several helpful techniques for choosing important variables, such the logistic model and discriminant analysis.

The dual GHSOM technique has two mechanisms: (1) a feature-extraction mechanism that extracts FFR-related patterns/features for decision support, and (2) a classification mechanism that identifies fraudulent samples.

The study employs a novel dual GHSOM (Growing Hierarchical Self-Organizing Map) approach for detecting fraudulent financial reporting (FFR) and feature extraction. This method utilizes both fraudulent and non-fraudulent samples to train dual GHSOMs under identical parameters, allowing for the examination of topological patterns among their subgroups. Additionally, the approach incorporates a classification mechanism to identify fraudulent samples and a feature-extraction mechanism to capture FFR-related patterns. Principal Component Analysis (PCA) is also utilized for dimensionality reduction and feature selection.

Based on artificial intelligence and statistics, Kirkos et al. [13] give a lot of different ways to group things into groups. This study uses three methods that are well-known for their ability to sort data into groups. Some of these ways are neural networks, decision trees, and Bayesian belief networks.

Three different models were made, each using a different method. Sipina Research Edition software was first used to build the Decision Tree model.

The authors built the Neural Network model in the second experiment. They used Nuclass 7 Non-Linear Networks for Classification to create a multi-layer feed-forward perceptron network.

In the third experiment, they built a Bayesian Belief Network (BBN). Our Programme was TheBN Power Predictor.

The Bayesian Belief Network model produced the best results in a 10-fold cross-validation procedure, correctly classifying 90.3% of the validation sample. The accuracy rate of the neural network model was 80%, compared to 73.6 percent for the decision tree model. All the models had a reduced Type I error rate.

The paper employs three primary Data Mining classification techniques to detect fraudulent financial statements (FFS): Decision Trees, Neural Networks, and Bayesian Belief Networks. Decision Trees are utilized for their ability to represent acquired knowledge and extract IF-THEN classification rules. Neural Networks, particularly backpropagation networks, are chosen for their efficiency in handling noisy data and their popularity in prediction tasks. Bayesian Belief Networks are applied to capture dependencies among attributes and improve classification accuracy through conditional independence tests. The performance of these methods is compared based on their predictive accuracy in identifying FFS.

Li et al. [14] looked into Traditional multivariate analysis is useless when there is sample size, and classification techniques, such as support vector machines, encounter "data piling." Marron et al. proposed distance weighted discrimination (DWD) as a feasible approach to address HDLSS problems.

DWD could be improved by addressing the data piling problem from the distinct standpoint of classifier performance via the projection of samples in the normal vector's direction. With samples weighted according to distance, the goal of the DWD optimization problem is to minimize the sum of the reciprocals of the residuals. In HDLSS scenarios, a linear classifier achieves good performance. GA-based classification models are utilized in this work for feature selection and parameter optimization.

Three cross-validations were performed. To evaluate the classifier's performance, 5-fold cross validation was used to calculate the average accuracy of the training data for each fold. Next, the testing sets prediction accuracy is calculated using the optimal parameter and chosen characteristics. Three equal-sized subsets of the raw dataset D are randomly selected to serve as independent testing sets T, while the other two subsets are used as training sets k = D - T.

Several classification techniques are employed to model the link between the input samples and the fraud outcomes. Optimizing parameters and feature subsets using genetic algorithms leads to better classifier performance.

The classifiers' accuracy determines the fitness score. Using the training and testing sets, are applied using individually produced parameters and chosen features.

The paper employs a combination of financial and textual

features extracted from 10-k filings for detecting financial statement fraud. It utilizes the Distance Weighted Discrimination (DWD) model, which is effective in high dimension low sample size (HDLSS) contexts. A genetic algorithm (GA) is applied for feature selection and parameter optimization to enhance the performance of classifiers. The methodology includes grid search for optimal parameter settings and F-score technique for filter-based feature selection. Various classification models, including DWD, Support Vector Machine, Neural Networks, and Decision Trees, are compared for their effectiveness.

Different types of machines learning models, including random forest, SVM, XGB (eXtreme Gradient Boosting), ANN, and deep learning models, including CNN, LSTM, GRU, and transformer, were used with real-world data by Wu and Du [15] to come up with classification results. These algorithms classify fraud based on financial, non-financial, and textual factors found in the Business Situation Discussion & Analysis (BSDA) sections of publicly traded companies' yearly reports. The study showed that GRU got a score of 94.49 percent and LSTM got a score of 93.98 percent.

The paper employs deep learning methods for financial fraud detection, specifically utilizing Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and Transformer models. It combines both numerical and textual data from Chinese listed companies' annual reports as input for the models. The study also incorporates techniques for Chinese word segmentation and embedding processing to transform unstructured text into a numeric format suitable for algorithmic processing. Additionally, the paper evaluates model performance using metrics such as AUC, sensitivity, specificity, and F1-score.

Andayani and Wuryantoro [16] Content analysis is the process of methodically reading a text, taking note of its recurring themes, and then extrapolating inferences on the meaning and purpose of the document (Hall & Wright, 2008). To get a more complete view of the governance (or set of procedures) and policies in place inside the organization, a qualitative content analysis was conducted (GCG). In addition to GCG, businesses should now use fraud detection systems and social responsibility.

The three system categories that comprise the coding scheme are social responsibility, governance systems, fraud detection, and faked financial statements. Every article emphasizes reading holistically and is coded. Additionally, there were debates about a number of papers where it was difficult to come to an agreement on whether the data and analysis were adequate.

The research employs qualitative content analysis as its primary methodology, which involves systematically reading and analyzing documents to identify consistent characteristics and derive conclusions about their purpose and meaning. This method was applied to a collection of 53 research articles from both Indonesia and abroad, focusing on themes related to financial statement fraud, governance (GCG), and corporate social responsibility (CSR). The coding scheme utilized in the analysis links falsified financial statements with governance mechanisms, fraud identification, and social responsibility, ensuring a holistic reading of the articles.

Sabatian and Hutabarat [17] as a stand-in for the Fraud Financial Statement, utilised the Benesh M-Score Formula.

The Fraud Triangle, which has three components—Pressure, Opportunity, and Rationalization—is used as the independent variable in this research. The concept of "opportunity" is separated into two categories: ineffective monitoring, represented by the proportion of independent commissaires (BDOUT), and industry nature, represented by the ratio of total inventory (INVENTORY). TATA, or the Total Accrual Ratio, is a stand-in for "rationalization."

A method called "purposive sampling" was used to look at 7 companies in the cigarette industry and 5 companies in the cosmetics industry that were traded on the Indonesia Stock Exchange between 2016 and 2018. Companies have to have been on the IDX since 2015 and have audited yearly reports for the time period of the study. Documentation methods, like downloading yearly reports from the IDX website, are used to gather data. To test theories about financial statement fraud, the study uses IBM SPSS Statistic 25 software and logistic regression analysis.

Sorkun and Toraman [18] use several types of machines learning to train the data. Specifics Data from the Balance Sheet and the Income Table made by 72 e-ledger were used during the training phase. Artificial Neural Networks (ANN), Decision Tables, Decision Stump, M5P Trees, K-Nearest Neighbor (KNN), Support Vector Machines (SVM), and Linear Regression were all used in the training process. These methods were put to the test in two different ways, and the results showed how well they worked. Right-sized. The Normalized Root Mean Squared Error (NRMSE) measure is used to figure out how well these methods work.

Apart from these investigations, three sets of labels were created and evaluated using the ranges of 0–40 is regarded as low, 41–80 as medium, and 81–100 as high. This grouping produced a fresh labelled data set, which was used to test several categorization techniques. Linear regression and M5P trees were substituted with measurements are used to assess the success rates of these techniques' results.

Several data mining techniques were used in the study to look for signs of fraud in financial statements drawn from e-ledgers. Artificial Neural Networks (ANN), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Decision Stump, M5P Tree, Random Forest, and J48 Tree are some of the methods that are used. There were two different methods used: one put fraud numbers on financial statements so that regression analysis could be done, and the other put statements into three groups: low, medium, and high fraud. Metrics like Normalized Root Mean Squared Error (NRMSE) and True Positive Rate (TP-Rate) were used to judge how well these methods worked.

The limited predictability of unusual occurrences, current technology-based techniques for identifying financial statement fraud, and upcoming research in identifying changing false financial reporting are all covered by Othman [19].

It is plausible to argue that catastrophic rare fraud occurrences are unlikely to occur given the limited predictability of rare events. On the other hand, offenders are still vital in order to alert interested parties to red flags of fraudulent activity early on. The shortcomings of earlier detection algorithms could be gradually improved with ongoing attempts to create a fraud detection model. Future fraud forecasts may become more accurate as a result of this.

The paper discusses various technology-based methods for detecting financial statement fraud, including data mining techniques such as regression, neural networks, decision trees, and Bayesian belief networks. Logistic regression is highlighted as a commonly used statistical method for predicting fraud patterns. The Bayesian classifier is noted for

its effectiveness in assessing financial misstatement risks. Additionally, the random forest model is recognized for its accuracy and ability to handle large datasets. The study emphasizes the need for innovative techniques to keep pace with evolving fraud tactics.

Moderated regression analysis is used by Anisykurlillah et al. [20] to evaluate the model's hypothesis, assuming that the model's classical assumptions are followed beforehand. In this investigation, an absolute value difference test was used. The difference between the standardized absolute values of the two independent variables is found using the absolute value difference test. The next step determines the direction of the moderating effect of institutional ownership on dependent variable.

The study chose 29 companies from a group of 58 publicly traded companies on the Indonesia Stock Exchange using purposive sampling. These companies were exactly those that made up the LQ45 index from 2016 to 2018.The data were analyzed with SPSS, which did both descriptive and regression studies. To test the hypothesis about how independent factors affect financial statement fraud, moderated regression analysis was used, with institutional ownership acting as a moderating variable. In order to make sure the regression model was correct, classical assumption tests like normality, multicollinearity, autocorrelation, and heteroscedasticity tests were also run.

Yadav and Sora [21] use the four-step procedure for detecting financial statement fraud: of text classification. We present here DNN-DHO approaches to improve the DNN model for FSF detection. A performance comparison is made between the suggested technique and other classifiers.

Omeir et al.'s [22] approach of systematic deletion is utilized for sampling. The analysis was made with a number of fraud factors, and a sample of all the firms in the statistical population throughout the study period with one of those variables is selected for this purpose, with the other companies being removed.

To find the necessary variables pertaining to the financial statements of the firms under examination, they consulted the financial statements available in the Tehran Stock Exchange's computerized archive. Part of the information required came from the widely used database software developed. The most current updated accounting standards provided the data.

The M-score, or Beneish [23] model, was created by Prof. Messod D. Beneish to identify financial fraud. To assess the possibility of financial or earnings manipulation in a company's financial statements, the model uses eight financial measures.

The research employs two primary fraud detection models: the Beneish model [23] and the Dechow model [24], to compare their precision in predicting financial statement fraud in Iranian companies. A sample of 197 companies was analyzed over an 11-year period from 2009 to 2019, utilizing SPSS software for statistical analysis, including t-tests and variance analysis. The methodology includes identifying companies with specific instances of fraud as noted in their audit reports, which were then selected for further analysis. The study also examines the predictive accuracy of both models against actual cases of fraud confirmed by judicial verdicts.

Zhao and Bai [25] use the four-step procedure for detecting financial statement fraud: text pre-processing, feature selection, feature extraction, and text classification. We present here DNN-DHO approaches to improve the DNN model for FSF detection. A performance comparison is made between the suggested technique and other classifiers, such as DNN.

The Synthetic Minority Oversampling Technique (SMOTE) is used in this work to fix sample imbalance in financial data. It uses several feature selection methods to find the most important indicators that affect financial fraud. In the end, 13 key indicators were chosen. Five single classification models are created by the authors. These are XGBoost (XGBOOST), Decision Tree (DT), Random Forest (RF), and Logistic Regression (LR). They also use ensemble models with a vote classifier to improve the accuracy of their predictions. Performance measures like accuracy, recall, precision, and AUC are used to judge the models.

Indrati and Claraswati [26] Jones Model as one of their calculation models to calculate earnings management. Discretionary and non-discretionary accruals are the two categories into which the Modified Jones Model separates accruals. The acknowledgment of accruals of unbound and unregulated income or expenses that are the decision of management policy is known as discretionary accruals. The quality of financial statements may deteriorate if non-discretionary accruals are not followed, even when they are legitimate accruals that adhere to generally accepted accounting rules. The Modified Jones Model is utilised since credit income is reflected on an accrual basis. The study's element fraud diamond makes use of the receivable ratio as a proxy variable. The Modified Jones Model is therefore a highly effective technique for identifying financial statement fraud.

The study employs the Modified Jones Model to measure financial statement fraud, which distinguishes between discretionary and non-discretionary accruals. A purposive sampling technique is utilized to select a sample of 20 property and real estate companies listed on the Indonesia Stock Exchange from 2015 to 2019, resulting in 100 company data observations. Multiple linear regression analysis is conducted using SPSS version 26 for hypothesis testing, allowing the examination of relationships between independent variables and fraudulent financial statements.

Using the fraud pentagon theory as a tool, Humphrey et al. [27] looked into financial statement fraud. The fraud triangle model, which was used in the research, was investigated in accordance with Cressey's 1953 aspect of the fraud triangle. This feature focuses on situations when dishonest persons in positions of trust have betrayed trust. In 2004, Wolfe and Hermanson developed the fraud diamond model, which included the capability component in addition to the fraud triangle. The study also looked at the shortcomings of the previously mentioned model as discovered via information research of relevant and recent literature. According to the study's findings, pressure indicators such as institutional ownership, financial stability, financial aim, and external pressure are reliable and sufficient for spotting potential financial statement fraud. The research also examined opportunity and discovered that it depends on several elements, such as the industry and the size of the audit committee. The research findings validated the significance of these variables in mitigating and identifying financial statement fraud inside a business. The study also found that rationalization helps identify potential financial statement fraud. This is demonstrated by factors such as accrual levels and the number of auditors who switch. Additionally, the study demonstrates that factors like the shift in corporate

directorship, which has a big impact on determining the likelihood of financial statements, can be used to estimate competency.

The study uses a library research method and focuses on looking at relevant and already published literature about the five parts of fraud that are involved in financial statement fraud. It looks at different parts of the fraud pentagon, like chance, pressure, rationalization, competence, and arrogance, and how they have big impacts on financial statement fraud. The study is divided into parts that explain what the study is about, talk about financial statement fraud, look at the fraud pentagon, and finally, say what they found about how these things relate to financial statement fraud.

Malik [28] used the data science life cycle to implement a conceptual framework for machine learning (DSLC). They asked two questions in the first phase:

How much can be predicted about fraud detection using machine learning techniques?

Which algorithms work best for this kind of task?

The second phase included gathering data from the UCI machine repository's publicly available Indian audit dataset of businesses. Between 2015 and 2016, the dataset was obtained from the Auditor General Office (AGO) in India. There are 776 occurrences and 26 numerical attributes in all. Enterprises from 14 different industries are included in the collection. Table 3 enumerates the industries and the number of companies in each.

Using Weka software, an exploratory analysis of the dataset was carried out in the third step. Because risks were unrelated to the prediction model, they were eliminated. The numeric attribute Money value had one missing value, which the exploratory analysis revealed. They used the ReplaceMissingValues filter to replace it with the mean.

The fourth phase involved NB, SVM, and RF are the three different modelling strategies that are used before proceeding with the experimental setup and choosing features for dimensionality reduction.

Alwadain et al. [29] used machine learning to guess how much financial crime there was. They used a conditional generative adversarial network for tabular data (CTGAN) to make over 5,000 cases after looking at a fake dataset of financial transactions. After teaching 27 machine learning algorithms, they carefully looked at how well the best model did.

To predict financial fraud using machine learning methods, the study used a made-up set of financial transactions. The Conditional Generative Adversarial Network for Tabular Data (CTGAN) was used to make about 5,000 more examples. With the usual settings and implementations, a total of 27 machine learning classifiers were trained and tested. Over and over, 10-fold cross-validation was used to make sure the evaluation was strong. The sample was split into training and testing groups with a 70:30 ratio. It was thoroughly tested and found that the model that worked best, XGBoost, was the most accurate.

Research trends, data sources, and ML/DM approaches were recognized by Rabade [30]. There are certain restrictions and questions about legitimacy, though. Utilizing an unbalanced data set and other supervised machine learning algorithms, the most effective and acceptable technique for identifying fraud. Following a comparison examination, they found that the K-Nearest Neighbour method outperforms Logistic Regression and Nave Bayes for detecting fraudulent transactions based on metrics like Precision/Recall and F1-Score.Owing to its benefits like increased dimensionality and precision, the Random Forest algorithm can be a wise option. Both regression and classification issues can be handled by it. Subsequent investigations have to concentrate on enhancing the precision of the models through the incorporation of intricate factors, utilization of unorganized data, and optimal training.

The paper explores various machine learning and data mining techniques for financial fraud detection, emphasizing their application in analyzing corporate financial accounts. It discusses the use of algorithms such as decision trees, logistic regression, K-Nearest Neighbour, and Random Forest for identifying fraudulent activities. The Random Forest algorithm is highlighted for its ability to handle both classification and regression tasks effectively, while K-Nearest Neighbour is noted for its superior performance in precision, recall, and F1-Score compared to other methods.

The study also emphasizes the importance of using historical data and confusion matrices to evaluate the performance of these algorithms.

A sample of manufacturing companies registered on the Amman Stock Exchange (ASE) between 2015 and 2019 was used by Saleh et al. [31]. 53 companies were included in the sample. The secondary data used in the analysis were taken from yearly reports that were released on (ASE). After subtracting the results that did not fit the study's criteria and were not part of the research period, there were 238 financial findings in all.

The research employs a thorough analytical method to determine firm bankruptcies and, as a result, financial statement fraud by using the Altman z-score model. According to the model's creator, according to research made 95% of firms' financial collapses may be accurately predicted one year prior to their demise. accuracy drops to 72% after two years and to 52% after three.

According to Altman's revised model, the likelihood of a business filing for bankruptcy is either too low or too high, depending on whether the z-score was accepted above 2.9 or below 1.2.

Dechow model states that a company's financial statements are thought to be clean if the F-Score is less than one (1); financial statements are thought to be clean if the F-Score is less than one If the F-Score is greater than one (>1), the company is probably lying. If the F-Score is zero, the company is probably telling the truth. (F-Score = 1), which means that there is a 1 in the chance that a certain event will end in a certain way, no matter what else may happen. As long as the F-score (FScore>1) is greater than 1, the expected probability may be greater than the unconditional probability. One more thing it could mean is that the company's books have been changed.

To test the study's theories, the research uses a methodological model with a multiple regression process. It mainly looks at the Altman Z-score model and the Dechow F-score model, which are two well-known mathematical models for finding fraud. The sample is made up of 53 manufacturing companies that are listed on the Amman Stock Exchange. Secondary data came from the companies' yearly reports. As part of the analysis, financial statements are put into groups based on how likely they are to fail, which makes it possible to look at fraud signs in a structured way. Financial ratios are also used in the study to look for fraud and judge the security of the finances.

Chukwuma et al. [32] 's research methodology includes the following steps:

1. Data collection: A sample of companies' financial statements will be collected from publicly available sources. Financial ratios, financial statement items, and information on accounting fraud cases will be included in the data.

2. Data cleaning and pre-processing: To ensure that the collected data is accurate and ready for analysis, it will be cleaned and pre-processed. This will include the removal of any missing or duplicate data as well as the correction of any errors.

3. Financial ratio analysis: For each company in the sample, financial ratios will be calculated. The purpose of these ratios is to find anomalous trends in financial data that might point to fraud.

4. Logistic regression modelling: Models of logistic regression will be trained to estimate the probability of fraud for every organisation in the sample using the financial ratio data. The models will be calibrated through the use of regularisation and cross-validation techniques in order to enhance their predictive performance.

5. Machine learning modelling: Machine learning models will be developed to forecast the probability of fraud for every organisation in the sample using the financial ratio data. The models will be calibrated through the use of regularisation and cross-validation techniques in order to enhance their predictive performance.

6. Model evaluation: To see how well the logistic regression and machine learning models work, we will look at their accuracy, precision, recall, and the area under the receiver operating characteristic curve (AUC-ROC). We will also look at how easy the models are to understand and how much they cost.

7. Data analysis: To find important patterns and trends in the data, the results of the financial ratio analysis, logistic regression modelling, and machine learning modelling will be evaluated. The results will be put to use in identifying businesses that are more prone to falsify financial statement reports and in supporting the investigation of higher-risk enterprises.

The study employs a thorough forensic data analytic approach that includes several key methods for detecting accounting fraud. Data collection involves gathering financial statement data from publicly available sources, including financial ratios and fraud case information. Financial ratio analysis is conducted to identify abnormal patterns in financial data that may indicate fraudulent activity. Logistic regression models are utilized to predict the likelihood of fraud based on financial ratios, with calibration techniques to enhance performance. Machine learning models are also trained on the financial ratio data to improve fraud detection accuracy and interpretability.

In order to evaluate financial statement fraud, this study, referenced as Jan [3], selected 18 criteria that are frequently employed. The set comprises four non-financial variables, often referred to as corporate governance variables, and fourteen financial variables. To build an ideal model, deep learning and optimization techniques like RNN and LSTM are applied. 60% dataset. The dataset's parameters are fitted during the learning process, and continuous optimization leads to the optimal prediction model. In order to validate the status and convergence with consistent accuracy, a random sample of 15% of the total data is used to create the validation dataset. Adjusting the hyper-parameters prevents overfitting. The test dataset, or the remaining 25% of the data, is used to assess the model's performance—that is, its ability to identify and generalize.

The study employs two deep learning algorithms: Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) for detecting financial statement fraud. RNN is utilized for processing sequential data, capturing relationships between data points, and preserving important past information during modeling. LSTM, an advanced version of RNN, is used for its superior performance in handling long-term dependencies in data. The research samples data from 153 TWSE/TEPx listed companies, combining both financial and non-financial data for model training and validation. Performance indicators such as accuracy, precision, sensitivity, and ROC curve AUC values are used to evaluate the models.

Cheng et al. [33] outline the six-step research approach that is being suggested. The following are the specific steps:

Getting the data ready: Thus, a survey of the literature served as the foundation for this research in order to identify the pertinent elements of financial statement fraud. Next, the collected dataset was assembled to include a total of 72 properties. Only 437 quarterly reports (181 fraudulent as well as 256 non-fraudulent) were kept after the 72 financial fraud characteristics were computed from the acquired financial fraud dataset. This was necessary since the query returns data had many missing values.

Dealing with null values Numerous missing variables in the gathered financial fraud dataset will have an impact on the categorization outcomes. The two approaches utilized in this study to address missing values are as follows.

(a) Pairwise deletion: After obtaining significant attributes through the use of five attribute selection techniques, the data from the chosen attributes was kept.

(b) Listwise deletion: Prior to applying the attribute selection technique to identify the crucial attributes, they eliminated every missing value from the financial fraud dataset that had been gathered.

Handling classes unevenly Unbalanced classes can be handled by many techniques; however, this study employed the SMOTE approach and random oversampling as two oversampling strategies to address the imbalanced class issue. In order to produce balanced classes, there are more minor class records when using the oversampling approach. By comparison, the minority class has almost the same number of entries as the majority class when the under sampling strategy is used.

Evaluation is a common technique for determining a model's efficacy, and in this step, 10-fold cross-validation is used to do so. After categorization, the efficacy evaluation metrics can be computed using a confusion matrix.

To deal with missing values in the dataset, the study uses both listwise and pairwise elimination. To find the important characteristics, it uses three combined attribute selection methods, one of which is a nonlinear distance correlation method. Both under sampling and oversampling techniques are used in this study to fix the imbalanced class problem. Random oversampling and the SMOTE technique are used especially. To make a useful set of rules for finding financial scams, rule-based classifiers like decision trees and random forests are used.

The study's population was defined by Antawirya et al. [34] as all financial sector businesses that were listed on IDX between 2015 and 2018. The method for gathering data is purposeful sampling. There are 47 firms that satisfy the requirements based on the sample selection criteria. To

evaluate the data, multiple linear regression analysis was used.

In the present research, Ariyanto et al. [35] examine fake financial reports using Dechow et al. [36] 's fraud score model. "The F-score model consists of the sum of two variables, namely accrual integrity and financial performance," according to Skousen et al. [37]. A stand-in for the variable of personal financial status is insider ownership of shares (OSHIP). An insider's cumulative percentage ownership in a corporation is determined by multiplying their shares by the total number of shares that are outstanding. The present study employs total inventory divided by total sales of the research year reduced by total inventory divided by total sales of the preceding year as a proxy for the industry's characteristics. Additionally, an ordinal scale (dummy variable) is used to evaluate the calibre of independent auditors. Companies audited by the Big Four firms—Ernst & Young, Deloitte Touche Tohmatsu, KPMG, and Pricewaterhouse Coopers—are coded 1 while companies audited by other firms receive code 0. The proxy for competence in relation to changes in corporate directors is examined in the present research. An ordinal scale (dummy variable) is used in this assessment, with code 1 denoting a change in the company's directing roles over the observed time. If there hasn't been a shift in the company's directing roles over the observed time, it is coded 0. The total number of CEO images on the business's annual financial report for the study year is a measure of arrogance. The data was analysed using panel data regression analysis.

The study chose a group from all the financial companies listed on the Indonesia Stock Exchange from 2015 to 2018 using a method called "purposeful sampling. “There were 47 businesses that met the requirements for the study. Multiple linear regression analysis was used to analyze the data and figure out how the parts of the fraud pentagon relate to fake financial statements. The goal of the study was to find out how well the fraud pentagon model can find theft in financial statements.

Ye et al. [38] use the Random Forest technique to identify FFS. In practice, it is not uncomplicated to utilise a single basic classifier due to the large class imbalance issue. Moreover, the data shown here indicates that the Random Forest is not a good option for logistic regression in this scenario. SMOTE was selected in order to reduce biases resulting from a greater probability of past fraud. When dealing with an unbalanced dataset, it is also suggested to include performance measurements like precision, recall, F-Measure, Kappa Statistics, and AUC to accuracy.

The Random Forest algorithm is used in this study to find financial statement fraud (FFS). This algorithm is good at learning from data that isn't balanced. By making fake samples for the minority class, SMOTE (Synthetic Minority Over-sampling Technique) is used to fix the class mismatch. Normalizing numeric characteristics and imputation for missing values are part of the method and are used to improve the quality of the data. In order to judge how well the model works, performance measures like recall, F-Measure, Kappa Statistics, and AUC are used. The study also looks at how well Random Forest works compared to other predictors, such as ANN and logistic regression.

Association rule mining was used by Sawangarreerak and Thanathamathee [39] to identify fraudulent trends in financial statements from the Thai Stock Exchange.

The study uses a method with three key steps: processing and discretizing the data, using FP-Growth to find related patterns, and looking for and analyzing fraud patterns in financial statements. As part of data preprocessing, binning discretization sorts the information into different groups that make it easier to analyze. Association rule mining is used to find patterns of fraud in financial records from the Stock Exchange of Thailand for this study. Another idea from the study is to use different data analysis techniques, like clustering, to help find fraud trends that are linked.

5. RESEARCH GAP AND FURTHER RESEARCH

As per Ashtiani and Raahemi [4], clustering algorithms have not been used in unsupervised approaches lately. Thus, further research into anomaly detection methods is an interesting area. Furthermore, heuristics and meta-heuristics algorithms may be further researched in combination with bio-inspired algorithms, including genetic algorithms (GA) and artificial immune systems (AIS), to identify financial statement fraud.

The study acknowledges potential search bias due to the possibility of missing relevant studies from additional digital libraries not consulted during the review process. This was somewhat mitigated by employing the snowballing technique to include related articles. Limitations regarding the size of datasets were noted, particularly in smaller markets like Taiwan, which could affect the generalizability of findings. The issue of imbalanced datasets was highlighted, as fraudulent financial statements are significantly fewer than non-fraudulent ones, complicating model training. The paper also mentions the need for further exploration of unsupervised and semi-supervised methods for anomaly detection.

The research identified a lack of attention towards unsupervised and semi-supervised machine learning techniques for fraud detection in financial statements, suggesting a need for further exploration in these areas. There is a notable gap in the use of unstructured data, such as textual and audio data, which could provide valuable insights for intelligent fraud detection. The reviewed literature showed limited application of clustering methods and other anomaly detection approaches, indicating these as potential areas for future research. Additionally, the issue of imbalanced datasets remains underexplored, with few studies addressing the use of various oversampling and under-sampling techniques.

The majority of financial transaction datasets, according to Ali et al. [5], include millions of transactions, and they are all plagued by unbalanced datasets. However, compared to the amount of legitimate financial transactions, the fraudulent financial transaction count is far smaller.

The systematic literature review (SLR) acknowledges various limitations and validity threats encountered during the study. Although protocols were developed to enhance external and internal validity, some limitations remain unaddressed. The review may not encompass all relevant studies due to the exclusion criteria applied, which filtered out articles not focusing on financial fraudulent transactions or those not utilizing machine learning methods. Additionally, the review is limited to articles published in English and those that are peer-reviewed, potentially excluding valuable non-English studies.

The paper identifies several research gaps in the area of financial fraud detection using machine learning, including the need for more comprehensive studies that address various types of financial fraud beyond credit card fraud. There is a lack of standardized evaluation metrics across different

studies, which complicates the comparison of results. The review highlights the necessity for more robust methodologies that can enhance the external and internal validity of findings in this field. Future research directions suggest exploring advanced machine learning techniques and their applicability to emerging fraud types.

Ali et al. [1] compared to non-fraudulent financial statements, there are much less fraudulent financial statements. Consequently, a significant issue with the datasets is their extreme imbalance. In certain papers, oversampling techniques were utilized to balance the datasets. The evaluated literature did not include any papers that used under-sampling methods. Moreover, oversampling was done exclusively using the SMOTE approach (Synthetic Minority Oversampling Technique).

The dataset used in the study is limited due to the scarcity of publicly available companies in the MENA region, which may affect the generalizability of the findings. The research focuses solely on financial attributes, leaving out the analysis of non-financial attributes, which could provide a more comprehensive understanding of financial statement fraud. The study acknowledges that traditional regression analysis methods have limitations in adequately detecting fraud, indicating a need for more advanced techniques. The performance of the model may be influenced by the inherent class imbalance in the dataset, despite the application of SMOTE.

The research highlights a limited dataset, as there are not many publicly available companies in the MENA region, which may affect the generalizability of the findings. It suggests that future studies could incorporate non-financial attributes and decentralized models to enhance the understanding of financial statement fraud. The paper indicates that while ensemble models have shown promise, there is still uncertainty regarding their performance compared to traditional methods across different data environments. Additionally, the study emphasizes the need for further exploration of quantitative techniques specifically targeting financial statement fraud detection.

Shen et al. [6] demonstrated the effect of applying correlation information using feature selection. The results of the FSFD comparison between the original and the models we created are shown. Bold text indicates which results are improved as a result of our developed tactics. Furthermore, compared to the other techniques, the SVM performs much better in terms of most classification metrics and is more sensitive to correlation data. NB performed poorly in the majority of classification criteria while using the learning feature embedding approach. This is mostly because the NB requires higher independence feature assumptions than the other four classifiers.

According to the paper, standard methods for finding financial statement fraud (FSFD) often miss correlation information between financial features, which could cause classifiers to not work as well as they could. The paper doesn't say exactly what the limitations are, but it points that way. In contrast to other classifiers, the naive Bayes classifier did not do well with the learning feature embedding method. This suggests that its assumptions about feature independence may not be as strong as they could be. This research mostly looks at data from publicly traded Chinese companies, which might make it harder to use the results in other situations or parts of the world. A weakness in current machine learning approaches is that they don't look at how different financial factors are related to each other. This can have a big effect on how well classifiers work in finding financial statement fraud (FSFD). One important point is that current methods mostly focus on choosing the right features or lowering the number of dimensions, and they don't always look at the different ways that financial features are linked to hidden factors. Traditional methods, according to the study, assume that features are conditionally independent, which means they don't take into account what we know about how features affect FSFD when they are added together.

In order to find financial fraud more easily, Zhou et al. [2] recommended that inductive graph embedding network algorithms like GraphSage, PinSage, and others should be improved and used in extra work. This would help them learn about the properties of new nodes that are added to a changing network graph.

The paper highlights that existing rule-based expert systems and traditional machine learning models struggle to detect financial fraud due to the dramatic increase in the scale of financial data. It notes that fraudsters continuously evolve their methods, making it challenging for detection systems to keep up. Additionally, the paper mentions that the complexity of financial interactions in networks necessitates advanced graph computing techniques, which can be resource intensive. The limitations of traditional algorithms, such as SVM, are also discussed, indicating their lower precision and recall rates compared to the proposed Node2Vec approach.

The study shows that the pattern of Internet financial fraud is always changing. This shows that we don't fully understand how these patterns change over time and that we need more flexible ways to find them. Not enough research has been done on inductive graph embedding network methods like GraphSage and PinSage, which could make it easier to find new nodes in dynamic networks. The paper says that current methods aren't up to par with how sophisticated fraud techniques are getting. This shows that we need to make fraud detection systems that are stronger and more adaptable.

In predictive modelling, observation under sampling (OU) works better than bootstrapping, according to Albizri et al. [7]. It becomes sense to investigate further categorization techniques considering these findings. Although previous study findings guided the selection of classification algorithms, it remains plausible that other algorithms may provide better results in identifying financial statement fraud.

Future study on this unstructured data should focus more on it since it might provide intriguing findings on fraud detection. Additionally, seldom used were unsupervised learning strategies and clustering algorithms, which are recommended for further study.

One limitation noted is that certain studies can only identify misstatements publicly acknowledged by the SEC, potentially overlooking many undetected cases. The rarity of misstatements revealed by the SEC poses challenges in developing effective detection models, leading to a high frequency of false positives. Additionally, the research excludes papers that do not establish or create fraud indicators, which may limit the comprehensiveness of the findings. The focus on specific journals and types of papers may also restrict the breadth of the literature reviewed.

The paper identifies a need for a structured framework to organize existing literature on financial statement fraud detection, highlighting gaps in current research methodologies and findings. There is a lack of comprehensive studies examining the characteristics of high-F-score firms that

remain undetected by the SEC, suggesting a gap in understanding earnings management within GAAP. Future research should explore the application of data analytics and machine learning to improve fraud prediction models, particularly in addressing issues related to noisy data and false positives. The generalizability of fraud detection models across different countries remains underexplored, indicating a need for comparative studies.

According to Deng [8], financial ratios make up the whole of their input vector. Accuracy may be increased by adding qualitative data to the input vector, such as the administrative board's makeup or the credentials of prior auditors.

The study's input vector is limited to financial ratios, which may restrict the model's accuracy; incorporating qualitative information could enhance results. The classification of non-fraudulent financial statements (non-FFS) does not guarantee that these statements are free from falsification, as future FFS behavior may still emerge. The model's effectiveness assumes that the selected variables are informative, but some may be independent of the sample class, potentially affecting discrimination. The research is primarily based on data from listed companies in China, which may limit its generalizability to other contexts.

The application of data mining techniques for management fraud detection has been minimal, indicating a gap in research in this area despite its potential. The study primarily focuses on financial ratios extracted from formal financial statements, suggesting a lack of exploration into qualitative factors, such as auditors' qualifications or board composition, which could enhance detection accuracy. There is a need for further investigation into the effectiveness of other data mining algorithms compared to the Naive Bayes classifier, as the study primarily emphasizes this method.

In the future, researchers may anticipate the FFS using other techniques, according to Chen et al. [9], to provide a more accurate reference. To further increase the method's classification accurate ratio, future researchers might try using other variable screening techniques.

The research on forecasting fraudulent financial statements (FFS) is insufficient, indicating a need for further exploration and alternative methods to enhance prediction accuracy. Some nonfinancial variables are challenging to measure and acquire, leading to their exclusion from the study. The focus on FFS may overlook certain instances, as some FFSs might not be identified, potentially affecting the study's accuracy. The study's sample selection may influence results, as pair companies could also become FFS companies in subsequent years.

The research identifies a gap in the forecasting of fraudulent financial statements (FFS), indicating that existing studies are insufficient in providing effective predictive tools for auditors beyond traditional analysis methods. It highlights the challenge of incorporating nonfinancial variables due to difficulties in measurement and material acquisition, suggesting that future research could explore different variable screening methods to enhance classification accuracy. Additionally, the study acknowledges that the sample may not encompass all FFS cases, which could affect the accuracy of the findings, indicating a need for broader sample selection in future studies.

In this study, Chen [10] introduced two-stage research models as a solution to the drawbacks of single-stage research models. With an accuracy of 88.59 percent for fraudulent financial statement identification and 83.19 percent for overall accuracy, the CART-CART model demonstrated the greatest level of performance. At 92.69 and 87.97 percent, respectively, the CHAID-CART model has the greatest accuracy for both overall and fraudulent financial statement identification.

The paper points out that most of the earlier research on finding fake financial statements only used a few statistical methods and didn't compare different models, which makes the results less reliable. It says that a lot of studies set up their detection models with a one-stage statistical method, which is not a good idea. Furthermore, traditional statistical models are criticized for making a lot of mistakes when trying to spot fake financial accounts. This shows that we need better methods. The study aims to fix these problems by using a two-stage statistical method and several data mining methods to make the results more accurate.

The study finds that most earlier studies only used one or two statistical methods and didn't compare models, which makes the results less reliable. It is not recommended to use a one-stage statistical method to set up detection models as was done in many previous studies. Multiple data mining techniques aren't being used enough to find fake financial statements, which means that analyses aren't full or sufficient. There should be a two-stage statistical treatment and tenfold cross-validation to make the model more accurate and reliable, says the study.

According to Goel and Uzuner [11], while factual information is often assumed to predominate in corporate disclosures, it might be challenging to employ sentiment characteristics to identify language patterns in papers that are both fraudulent and true. Future studies might examine the ways in which the expression of a feeling can be strengthened or weakened by the employment of intensifiers and demisters.

One problem with the study is that sentiment analysis is very dependent on the domain from which the training data are taken, because the same word can have different polarities in different situations. Because of this, the results may be more useful in the accounting domain and less useful in other domains. The study also admits that a model chosen based only on high accuracy may still have low predictive power for the specific domain problem being looked at.

The study fills in a gap in the existing literature by looking at how mood and fraud detection work together. This is because previous empirical work in accounting has not fully looked into this relationship. Its goal is to fill in the gaps in our knowledge about how to use qualitative predictors, especially the mood found in annual reports, to find fraud. This study also shows that we need more advanced natural language processing (NLP) methods to look at non-factual data like feelings and emotions in order to find fraud. Additionally, it says that sentiment analysis depends on the topic, which means that it can't be used in all situations.

Huang et al. [12] Future studies might also examine non-written business communication channels to learn more about how emotion can be used to mislead people. One of the drawbacks of the research is the significant sensitivity of sentiment analysis to the domain from which the training data is derived.

There needs to be more research into other approaches that might make the suggested dual GHSOM approach work better and faster. For example, supervised learning techniques like SVM could be used in future studies to create a nonlinear decision border. The need to fix the Type I and Type II mistakes in the classification rules is acknowledged, showing

possible ways to make scam detection more accurate. For better detection of possibly fraudulent activities, the study stresses the need for stronger early warning signals.

More study is required, according to Kirkos et al. [13], to address the problem of optimum discretization techniques as well as the effect of data discretization on model performance. Research is also required to ascertain the circumstances in which DM approaches are superior to other approaches.

The paper highlights that detecting management fraud is challenging due to a lack of knowledge regarding its characteristics and the infrequency of such occurrences, which results in auditors lacking the necessary experience to identify it. It notes that managers often attempt to deceive auditors, making standard auditing procedures insufficient for detection. Additionally, the increased focus on system assessment conflicts with the reality that most significant frauds originate from top management levels, where controls are often weak. The study also mentions software limitations that affected the transparency of the Bayesian Belief Network's synaptic weights.

This study shows that Data Mining methods aren't used very much for finding management fraud. This means that there isn't as much written about this topic as there is about other topics, like predicting bankruptcy. According to reports, not enough is known about the signs of management fraud, which makes it harder to find. The study shows that normal auditing methods might not be enough to find management fraud, which means that more analytical methods are needed. The study also shows that AI methods have some potential, but they don't always do better than statistical methods. This means that more research needs to be done in this area.

Li et al. [14] state that as there are 3.6 to 11.6 years between the filing date and the publication date of the AAERs (which were issued between 2006 and 2008), semi-supervised learning will be taken into consideration in the future to cope with a weakly-labeled dataset of enterprises. In addition, oversampling strategies will be used to tackle the problem of class disparity. Additionally, new elements like stock data may be leveraged to better identify financial statement fraud and gauge the possibility of fraud risk.

Only Chinese datasets with a sample length of around five years were utilized in this analysis after Wu and Du [15] removed certain datasets for different reasons. Prediction performance may also be increased by enhancing text mining algorithms.

The study's sampling period is limited to five years, which may result in the exclusion of companies that have been delisted for various reasons, potentially affecting the prediction results. Some annual reports had to be eliminated due to incompleteness, which could also impact the accuracy of the predictions. The applicability of the models is restricted as the data source only includes Chinese listed companies, excluding those from other markets, which may limit the generalizability of the findings. There is no consensus on the best group of variables for financial fraud detection, indicating a gap in existing research.

The research identifies a gap in the consideration of non-financial factors related to corporate governance in financial fraud detection, which are often overlooked in existing studies. This includes aspects such as ownership structure and management structure that could provide additional fraud clues. There is a limitation in the dataset used, as it only involves Chinese listed companies, which may affect the applicability of the models to other markets. The study also notes the challenge of handling high class imbalance in datasets, which complicates the evaluation of model effectiveness.

According to Andayani and Wuryantoro [16], increasing the sample size to around 100 would help to lower the margin of error. A larger sample size leads to a greater accuracy rate since it increases population diversity.

The research faced limitations due to some articles being inaccessible, which reduced the sample size used in the study. There were no major adversities affecting the study, but the difficulties in data access highlighted areas for improvement. The elements of governance tested through quantitative research yielded varying results, indicating inconsistency in the effectiveness of good corporate governance systems in reducing fraudulent financial statements. The study suggests that expanding the sample size in future research could enhance accuracy and variability in the population.

The research identifies a gap in the accessibility of articles, which has resulted in a decreased sample size, impacting the comprehensiveness of the findings. There is a need for future studies to increase the sample size to around 100 to minimize the margin of error and enhance the accuracy rate. The varying results from quantitative research on governance elements indicate inconsistencies in the effectiveness of good corporate governance systems in reducing financial fraud. Additionally, the relationship between CSR activities and financial fraud remains ambiguous, with some studies showing a negative correlation.

A few other proxy factors, such the number of commissioners, the sales-to-receivables ratio, and director changes, are suggested by Sabatian et al. [17]. Samples from other sectors, such as manufacturing or commerce, need to be included in future study. Increase the duration of the company's investigation.

The study says that to better understand how to spot financial statement fraud, more research should include more variables, such as the number of commissioners, the ratio of sales to receivables, and changes in directors. For more complete results, it suggests adding companies from other industries to the sample, like trade or production, and keeping the study going for longer. The study shows that some things, like Financial Stability, External Pressure, Personal Financial Need, and Nature of Industry, don't have a big effect on financial statement fraud. This shows that the Fraud Triangle model has some holes.

The study recommends further research to include additional variables such as the number of commissioners, the ratio of sales to receivables, and changes in directors to enhance the understanding of financial statement fraud detection. It suggests expanding the sample to include companies from other sectors, such as manufacturing or trade, to provide a broader perspective on fraud detection. The research indicates a need for a longer observation period to better capture the dynamics of fraudulent activities over time. There is a lack of significant findings regarding several components of the Fraud Triangle, indicating potential areas for deeper investigation.

The studies conducted by Sorkun and Toraman [18] illustrated the significance of feature selection. It is intended to broaden the feature set and use Deep Learning techniques in subsequent research to identify the traits that are capable of identifying fraud.

Choosing the right traits is an important part of detecting fraud, but there isn't agreement on which ones are the best.

This could make the methods used less effective. The study also says that scam methods change over time, which suggests that a system that can adapt and update itself might be needed to keep up with these changes. These things suggest that it might be hard to keep scam detection methods accurate and useful over time.

The study shows that we need to learn more about how to choose features because there isn't a single agreement on which features are best for finding fraud in financial records. It seems to show that even though different ratios have been suggested in the books, it is still hard to find the unique features. The study suggests that there might be a hole in the way unsupervised learning methods are used because systems that combine them with supervised learning methods are not good enough. In the future, researchers might also look into how to use deep learning techniques to make it easier to spot scams.

Othman [19] say that even though a lot of time and work has gone into finding fraud, both the rate and number of fraud finds have gone down by a large amount. The problem is that CEOs who commit financial fraud are more likely to change their plans to avoid being caught if they know about the tools and methods that can be used to find fraud.

Later research could look into finding ways to make the program fit the specific needs of a company. When trying to find scam, a model might not give the most accurate prediction. Some say that problems with corporate governance have led to a string of corporate financial crises. Because of this, study might want to look at governance variables, Exogenous indicators would also help identify and find financial fraud more accurately. Along with internal firm-specific factors, these also include external factors related to the economy, industry, and institutional setting.

The prediction of rare events, such as financial statement fraud, is inherently challenging due to the small reference class, which complicates the extraction of relative frequency information and increases the likelihood of biases in judgment. Existing fraud detection models may oversimplify complex systems, leading to inaccurate estimations of probabilities and uncertainties. Many current detection techniques have not evolved alongside the tactics used by fraudsters, making it increasingly difficult to identify fraudulent activities. The reliance on red flags for fraud detection is criticized for being imperfect and potentially limiting auditors' ability to discover other fraud causes.

There is a big study gap in this area because of how hard it is to predict rare fraud events. This paper stresses the need for better detection tools that can let people know about fraud activities early on. Not much study has been done on how well fraud detection techniques work in real time, which is important because fraud techniques are always changing. In future research, it should be looked into how to improve the accuracy of fraud prediction and detection models by adding control factors and exogenous parameters. The paper also shows that current models aren't able to keep up with the tricks that scammers use.

As stated by Anisykurlillah et al. [20], there may be restricted data analysis throughout time due to the short research duration of three years. Determining the independent variables in the regression with more precision may be achieved with a longer data analysis time.

Other recommendations included extending the research duration by a minimum of five years and increasing the number of samples. The following study's very low value of determination makes it necessary to include other factors that may have an impact on financial statement fraud, which is a significant predictor of the crime. It is advised to consider the opportunities proxied by supervisory efficacy, organizational structure, and auditor change in many studies on the growth of the fraud triangle, including the fraud diamond. Additional theories that might reinforce the existence of internal control include the fraud hexagon, which adds capacity, ego, and collaboration to this investigation. To improve on this study, researchers could include additional moderating factors that affect independent variables related to financial statement fraud, such as capacity or financial difficulty.

The study is limited by its short duration, covering only three years, which may restrict the depth of data analysis over time. A longer study period is recommended to enhance the accuracy of determining independent variables in the regression analysis. The low R value of 46.2 indicates a need for further investigation into additional variables that could influence financial statement fraud. The research suggests incorporating other moderating variables, such as capability or financial distress, to better understand their impact on financial statement fraud.

The research identifies inconsistencies in previous studies regarding the influence of financial stability, external pressure, and rationalization on financial statement fraud, indicating a need for further investigation in these areas. The role of institutional ownership as a moderating variable is underexplored, particularly its potential to mitigate the effects of financial targets on financial statement fraud. The study's short observation period limits the generalizability of findings, suggesting that longer-term studies could provide more comprehensive insights. The low levels of institutional ownership in the sample may have affected the effectiveness of monitoring, highlighting a gap in understanding the optimal levels of institutional ownership for fraud prevention.

The findings of Yadav and Sora [21] show that the suggested method is the best way to find FSF. The created method has an AUC of 0.94, a score of 96% for accuracy, 97% for sensitivity, 98% for precision, 97% for F1 score, 94% for specificity, 0.03% for BER, 0.05 for FPR, and 0.026 for FNR. These numbers were calculated and compared to existing models. It is hoped that the suggested method will work better than Bayes, BP-NN, DNN, CART, SVM, LR, and KNN, which are some of the current models.

The paper identifies a lack of research focusing on qualitative data, such as auditor's remarks, in the detection of Financial Statement Fraud (FSF), which has predominantly relied on quantitative data like financial ratios. There is a need for improved methodologies that can effectively process and analyze textual data in financial reports, as existing methods have not fully addressed this aspect. The study highlights the absence of comprehensive models that integrate both financial and non-financial variables for enhanced accuracy in FSF detection. The paper suggests that current approaches do not sufficiently leverage advanced optimization techniques for feature selection in text mining.

Omeir et al.'s [22] statistical findings indicate that compared to the Dechow model, there is reduced estimate error and more forecast accuracy with the Beneish model. This implies that the Beneish model is superior to the Dechow model. has a better detection capacity for the likelihood of financial statement fraud. As a consequence, businesses with a history of managing profits run the danger of financial statement fraud. Fraud detection is made simpler by the Beneish model.

while evaluating the significance level while auditing firms, auditors should give particular attention to these two models because of the importance of the link. It is advised that investors, financial institutions, and credit agencies adopt the Beneish model to calculate loan and investment risk.

There may be some problems with applying the research's results to other situations or markets because it only looks at companies that were accepted to the Tehran Stock Exchange before 2009.Based on data from only 197 companies over 11 years (2009–2019), the study may not fully reflect all financial reporting methods. These businesses were not involved in investments, financial services, loans, or money, so important areas of the economy may not have been included in the study. Relying on old data might not take into account how scam detection methods have changed over time.

The research highlights a lack of comprehensive scientific and professional literature specifically addressing the differentiation between earnings management and earnings manipulation, which is crucial for accurate fraud detection. There is an indication that existing fraud detection techniques are improving, yet the increasing number of fraud cases suggests a need for further analysis using diverse tools and methodologies. The study emphasizes the necessity for auditors to consider the implications of a company's previous earnings management record, suggesting a gap in current auditing practices regarding fraud risk assessment.

As society changes, many topics still need in-depth investigation, according to Zhao and Bai [25]. There are three issues that need to be resolved.

•They need to gather as much information as they can in terms of data collecting.

•The financial fraud detection and prediction model may be used not just in the stock market but also in business operations to track a company's financial state from anywhere at any time.

•In order to replicate the experiment and evaluate the model impact, they want to gather financial metrics from additional publicly listed firms in other sectors.

As society changes, the paper admits that many issues still need more study. This shows that there is a need for ongoing research in the field of detecting financial fraud. It stresses how important it is to gather as much information as possible to make models more accurate and reliable. The authors want to make sure their model works by using it on financial measures of other publicly traded companies in different industries. This means that the current model might only work with the dataset that was used. The study also says that the financial fraud dataset is very unequal, which can have an effect on how well the model works.

The paper acknowledges that many issues in financial fraud detection require deeper research, particularly as society continues to evolve. This indicates a gap in the comprehensive understanding of fraud detection methodologies in dynamic environments. There is a need for further exploration of imbalanced data sets, as the paper primarily focuses on the application of SMOTE without delving into other potential methods or hybrid approaches. The authors suggest that future research could enhance the prediction of fraudulent behavior in company management, indicating a gap in practical applications of the proposed models.

Due to the fact that not all of the fraud diamond's variables were included in the study by Indrati and Claraswati [26], as well as the absence of statistical bias—a side effect of quantitative research methods—fraud risk factors, the study had limitations. A broader demographic and a sample of businesses registered on the Indonesia Stock Exchange are anticipated for future research. It is advised that future research assess rationalisation and ability using qualitative techniques and that they employ Likert scale questionnaires as main data to represent the factors related to rationalisation and ability.

The study admits that it didn't use all of the variables in the fraud diamond, which shows that there is a need for a more complete look at the risk factors for fraud. It has been pointed out that there is no statistical bias, which is a result of the quantitative study methods used. This could mean that the risk factors for fraud are not accurately reflected. More research should be done with a larger population than the 20 companies that were studied in this study. This might make the results more applicable to a wider range of situations. The study mostly looks at the real estate industry. It doesn't look into other fields, which might give us more information about financial statement scams.

The fraud pentagon literature is put out by Humphrey et al. [27] who contend that the model is adequate and trustworthy for identifying the possibility of financial statement fraud.

The study highlights that the fraud pentagon model, while comprehensive, may still have limitations compared to other models like the fraud triangle and fraud diamond, particularly in addressing all aspects of financial statement fraud detection. It notes that the effectiveness of the fraud pentagon model relies on the accurate measurement of its components, such as pressure, opportunity, rationalization, competence, and arrogance, which can be subjective and vary across different contexts. Additionally, the paper suggests that the dynamic nature of corporate environments may affect the applicability of the model over time.

The study shows that earlier research about what causes financial statement fraud was not always consistent, especially when using the fraud pentagon model. This means that these relationships need more empirical research to be made clear. Not enough studies have put together all the different parts of the fraud pentagon—pressure, chance, rationalization, competence, and arrogance—into a single framework for finding financial statement fraud. The paper says that the current research hasn't looked into how corporate governance systems affect how well the fraud pentagon model works.

The dataset used by Malik [28] is very limited, which is a significant restriction. Future research will thus need a big dataset.

Alwadain et al. [29] claim that adding additional data on fraudulent transactions would improve the model's dependability.

The study primarily focuses on the application of machine learning for financial fraud detection using a synthetic dataset. Potential gaps include the need for validation of the model on real-world datasets, exploration of additional machine learning algorithms beyond the 27 tested, and addressing the limitations of synthetic data generation methods like CTGAN in reflecting real-world fraud patterns. Further research could also investigate the impact of feature selection on model performance and the integration of temporal data for improved predictions.

Future research, in the opinion of Rabade [30], should concentrate on how to train the models to provide the best outcomes by including more features to increase the models' accuracy. Subsequent investigations into fraud detection need to concentrate on heuristic algorithms that are unsupervised, semi supervised, bio-inspired, and evolutionary. It is

anticipated that future study will make use of both textual and audio information. Further research is required because, while posing additional difficulties, this unstructured data might provide interesting insights for sophisticated fraud detection.

The paper highlights the high costs associated with building and maintaining a machine learning system, requiring companies to hire skilled data scientists and invest in data management and storage. It notes the risk of fraudsters circumventing poorly constructed machine learning models, leading to increased fraudulent transactions. The difficulty in gathering reliable, high-quality data is emphasized, as sufficient and meaningful datasets are crucial for effective model performance. Additionally, the need for technical proficiency in developing error-free models is identified as a significant barrier.

The paper identifies limitations and validity threats in the current machine learning and data mining techniques for fraud detection, indicating a need for further research to address these gaps. There is a lack of sufficient high-quality data necessary for developing long-term functional machine learning models, which poses a challenge in gathering reliable datasets. Future research should focus on improving model accuracy by incorporating complex parameters and utilizing unstructured information, which has not been adequately explored. The paper suggests that existing algorithms may not effectively handle unbalanced datasets, highlighting a critical area for improvement.

Saleh et al. [31] suggest many actions for investors, enterprises, decision makers, and scholars in terms of suggestions and further research: 1. The fraud detection model should be applied consistently to every new instance of questionable conduct. 2. Consider specific elements such individual financial requirements, possibilities, and organisational structure. Adhere to a new fraud factor, such as the one suggested by the diamond principle.

The research paper identifies a limitation concerning the lack of detailed information for various listed firms, which may affect the comprehensiveness of the analysis. Another limitation pertains to the sample size; although the study covered five years with 130 observations, this number is considered too limited in some testing scenarios to generalize the results effectively. These limitations did not impact the overall outcome of the analysis, but they suggest areas for improvement in future research.

The research identifies a limitation concerning the shortage of details for various listed firms, which may affect the comprehensiveness of the findings. Another gap is related to the analysis sample size; with only 130 observations over five years, this may be too limited to generalize the results effectively. The study suggests the need for further exploration of discrete factors such as personal financial needs and the division of pressure factors into internal and external categories. Additionally, it highlights the importance of focusing on the rationalization aspect of the fraud triangle and exploring new fraud factors.

Chukwuma et al. [32] proposed the following future improvements.

1. Accounting fraud may be found by combining several data analytic approaches, such as logistic regression, financial ratio analysis, and machine learning.

2. Interpretability and cost-efficiency should be taken into consideration when choosing data analytic approaches for accounting fraud detection.

3. More study may be done to look at the possibilities of different machine learning models, such ensemble approaches, deep learning, and so on.

4. It is also possible to look at the use of sophisticated methods like natural language processing (NLP) in conjunction with real-world data to identify accounting fraud in monetary reports.

5. In reaction to modifications in the banking system and fraudulent tactics, it is essential that the models and procedures for identifying accounting fraud be updated often.

6. Regular monitoring of financial ratios and financial statement components may help identify any unusual trends in the data that can point to fraud.

7. The government and regulatory agencies should take action to enhance financial statement data accessibility and openness to aid in the discovery of accounting fraud.

8. Additional study on the use of data analytics in developing countries and small and medium-sized enterprises (SMEs) is also recommended.

9. In order to shield stakeholders like consumers, workers, and investors from the damaging effects of accounting fraud, it is also essential to educate and create awareness about this crime among these groups.

There is a lack of research on the combination of different data analytic techniques for detecting accounting fraud, particularly the integration of machine learning models with traditional methods like financial ratio analysis and logistic regression. The study highlights the need for further investigation into the interpretability and cost-efficiency of various techniques used for fraud detection. Additionally, there is a call for future research to explore the potential of advanced machine learning models, such as deep learning and ensemble methods, as well as the application of natural language processing (NLP) in detecting accounting fraud.

The following research suggestions are made for the future by Jan [3]: First off, utilizing various deep learning algorithms. Second, include as study variables measures of economic expansion or contraction (e.g., GDP) or other macroeconomic variables. Third, to address particular problems includes context of economic and financial imbalance, take into consideration using additional econometric models. Fourth, in assessing financial statement fraud, non-financial elements must be considered in addition to the often utilized financial variables. Lastly, to take into consideration modifications to capital markets, company structures, financial laws, and economic situations.

Traditional approaches to financial statement fraud detection have shortcomings, which the study aims to address through deep learning algorithms. The research highlights the increasing complexity of financial statement fraud, and the challenges faced by regulatory schemes in preventing it. Additionally, it notes that despite advancements, financial statement fraud still occurs, indicating a need for continuous improvement in detection methods. The study emphasizes the importance of robust corporate governance and internal controls to mitigate these issues but does not detail specific limitations of the models used.

The study found that there isn't a lot of writing about how to use deep learning algorithms to find fraud in financial statements, which shows that more research is needed. It shows that more deep learning algorithms, like DNN, DBN, CNN, CDBN, and GRU, need to be added to scam detection models to make them better. This study shows how important it is to include macroeconomic indicators like GDP and change financial and non-financial variables to reflect changing

economic conditions. These topics are not fully studied yet. It is suggested that econometric models be used in future studies to look into certain types of financial inequality. This shows that there is a need for more methodological methods.

In subsequent research, the following problems from the Cheng et al. [33] study may be resolved: In order to address the imbalanced class problem, First, the meta cost cost-sensitive learning method can be applied; Second, missing values can be filled in using imputation technology; Third, performance can be assessed using alternative classifiers; and fourth, the TEJ dataset and SFIPC lists keep expanding, enabling more research on the data samples.

The study acknowledges that the methods used for handling missing values were limited to only two simple techniques: listwise deletion and pairwise deletion, which may not be comprehensive enough for all datasets. It suggests that future work could explore cost-sensitive learning methods, such as metacost, to better address the imbalanced class problem. The paper also indicates that imputation technology could be further applied to improve the handling of missing values. Additionally, it notes the potential for evaluating performance with other classifiers beyond those used in the current study.

The paper identifies a lack of comprehensive rules to support auditors in detecting financial fraud, despite previous studies performing well in exploring fraud factors. It highlights that many studies have not adequately addressed the imbalanced class problem in financial statement fraud datasets. The research also points out that existing methods for handling missing values are limited, with only listwise and pairwise deletion being utilized. Future work is suggested to include cost-sensitive learning methods, imputation technologies, and the evaluation of other classifiers to enhance the detection of financial fraud.

The Antawirya et al. [34] test shows that the chance of false financial statements goes up as the company's goal return on assets (ROA) goes up. On the other hand, the chance of statement fraud goes down as the number of audit committee meetings goes up. This study doesn't look at rationalization, competency, or hubris. Instead, it looks at direction changes, which is measured by auditor turnover, and how often photos of the CEO show up in false financial statements.

The study shows that not all parts of the fraud pentagon theory were shown to be useful for finding fake financial records. Only two of the five hypotheses were supported, which suggests that the theory isn't completely useful. The study didn't find any major effects of auditor turnover, direction changes, or the number of times a CEO was pictured on fraudulent financial statements. This means that these factors need to be looked into more. Because the study only looked at companies in the financial field that are listed on the Indonesia Stock Exchange, the results may not be applicable to other industries or parts of the world.

Ariyanto et al. [35] contend that the scope needs to include other industries including banking, finance, and mining. Future researchers may choose to use additional proxies derived from study factors that impact the indicators of financial statement fraud inside an organization.

The research addresses a gap in the application of the fraud pentagon theory specifically within the pharmaceutical industry in Indonesia, as previous studies focused on other sectors such as manufacturing and banking. It highlights the lack of research examining the relationship between personal financial needs and fraudulent financial statements, particularly in the context of pharmaceutical companies during the COVID-19 pandemic. The study also identifies a need for further exploration of the impact of external auditor quality and the frequency of CEO photo appearances on financial statement fraud, as previous findings were inconsistent.

Ye et al.'s [38] work may be expanded by using several datasets and imputation techniques for missing information. Additionally, non-structured data and qualitative information, such the composition of the board of directors or more details on internal controls, might be included in the predicators in the subsequent research. We urgently need to add ongoing auditing to the process in light of FFS's evolving nature. To create a real-time model that is both efficient and successful, based on past financial data and capable of anticipating prospective fraudsters before they commit the crime, further research will be needed in the future.

The study highlights the issue of high class imbalance in the dataset, with a ratio of 1:7.9, which complicates the detection of fraudulent financial statements (FFS) using traditional classifiers. It notes that the Random Forest algorithm may suffer from overfitting, leading to unreliable accuracy measures despite achieving high classification performance. The paper also points out that previous methods, such as Listwise Deletion for handling missing values, can result in biased conclusions. Additionally, the reliance on a single classifier may not adequately address the complexities of the data.

The paper identifies that previous studies have not effectively addressed the issues of missing values and severe class imbalance in financial statement fraud detection, often leading to biased conclusions due to methods like Listwise Deletion. It highlights that many prior works focused on unrealistic prior fraud ratios, which resulted in models that performed well under controlled conditions but poorly in real-world scenarios. Additionally, the paper notes that traditional classifiers struggle with recognizing minority classes due to limited information, indicating a need for more robust models.

Sawangarreerak and Thanathamathee [39] found the following constraints:

(1) They hired companies that were listed on the Thai Stock Exchange, regardless of the nature of the business.

(2) The links that were discovered were the common connected patterns of Thai companies registered on the Thai Stock Exchange.

(3) Operating structures, financial statement data, and capital and asset management systems vary depending on the kind of business. The nine financial items we discovered connected to fraud were:

(4) Total asset value. Typically, these solutions alert customers to possible financial statement fraud. Other organizations may generally follow our research methodology.

Future research will have access to a distinct dataset of the many business categories represented by businesses listed on the Thai Stock Exchange, which will enable the proper and unambiguous identification of related fraud tendencies for every category of company. To find other fraud trends, new datasets of businesses that aren't registered on the Thai Stock Exchange should be consulted. Furthermore, data should be grouped using other data analytics techniques, such clustering, before looking for related patterns.

The study points out a problem with the way standard statistical methods are used: they may not be able to find fraud because they depend on certain assumptions, like keeping independent variables from being too similar to each other.

This problem shows that we need more advanced data analysis methods, like machine learning, to make it easier to spot scams. Also, the study shows that datasets from companies that aren't listed on the Stock Exchange of Thailand need to be looked into more in order to find more fraud trends that are linked to them. There are also calls for adding other data analysis techniques, such as clustering, to make it easier to spot trends of fraud.

The publications that have been checked and data retrieved about the technique utilized, the algorithms that were used, and the research gaps and future studies are summarized in Table 3.

6. CONCLUSIONS

Most of the researchers have used various models from Machine Learning and Data Mining to propose their solutions which proved to the best in detecting financial fraud. Even many researchers have given their conclusions that still a lot of research has to be conducted in different areas.

Still a lot of research gaps are available in the current research area where some methods are left untouched, and some are rarely implemented, and some generate unsatisfactory results.

This review provides a detailed examination of various methodologies employed in detecting fraudulent financial statements. It highlights the growing complexity of financial markets and the exponential increase in data availability, which have made fraud detection a critical yet challenging task. The review underscores the effectiveness of traditional statistical methods, machine learning algorithms, and hybrid models in identifying anomalies and irregularities in financial data. It also emphasizes the role of artificial intelligence and big data analytics in enhancing the precision and effectiveness of fraud detection processes.

A significant portion of the paper is dedicated to classifying research methods into four main categories: Analytical, Behavioural, Conceptual, and Design Science. The majority of the studies reviewed (72 papers) utilized analytical methods, primarily involving statistical analysis with secondary data or the design science research paradigm. The paper also discusses the challenges in reaching consensus on the adequacy of data and analysis in some studies, indicating ongoing debates in the field.

The paper identifies several research gaps and areas for future exploration. It notes the lack of comprehensive rules to support auditors in detecting financial fraud and highlights the imbalanced class problem in financial statement fraud datasets as a significant issue that has not been adequately addressed. Additionally, the paper points out the limitations of existing methods for handling missing values, suggesting the need for cost-sensitive learning methods, imputation technologies, and the evaluation of other classifiers to improve fraud detection.

Furthermore, the paper emphasizes the importance of robust regulatory frameworks and ethical guidelines to prevent misuse and ensure transparency in fraud detection practices. It concludes by discussing future research directions, particularly the potential of emerging technologies to revolutionize the field of financial fraud detection. The authors plan to validate their model by applying it to financial indicators of other listed companies across various industries, acknowledging that the current model's applicability may be limited to the dataset used.

REFERENCES

[1] Ali, A.A., Khedr, A.M., El-Bannany, M., Kanakkayil, S. (2023). A powerful predicting model for financial statement fraud based on optimized XGBoost ensemble learning technique. Applied Sciences, 13(4): 2272. https://doi.org/10.3390/app13042272

[2] Zhou, H., Sun, G., Fu, S., Wang, L., Hu, J., Gao, Y. (2021). Internet financial fraud detection based on a distributed big data approach with node2vec. IEEE Access, 9: 43378-43386. https://doi.org/10.1109/ACCESS.2021.3062467

[3] Jan, C.L. (2021). Detection of financial statement fraud using deep learning for sustainable development of capital markets under information asymmetry. Sustainability, 13(17): 9879. https://doi.org/10.3390/su13179879

[4] Ashtiani, M.N., Raahemi, B. (2021). Intelligent fraud detection in financial statements using machine learning and data mining: A systematic literature review. IEEE Access, 10: 72504-72525. https://doi.org/10.1109/ACCESS.2021.3096799

[5] Ali, A., Abd Razak, S., Othman, S.H., Eisa, T.A.E., Al-Dhaqm, A., Nasser, M., Elhassan, T., Elshafie, H., Saif, A. (2022). Financial fraud detection based on machine learning: A systematic literature review. Applied Sciences, 12(19): 9637. https://doi.org/10.3390/app12199637

[6] Shen, Y., Guo, C., Li, H., Chen, J., Guo, Y., Qiu, X. (2021). Financial feature embedding with knowledge representation learning for financial statement fraud detection. Procedia Computer Science, 187: 420-425. https://doi.org/10.1016/j.procs.2021.04.110

[7] Albizri, A., Appelbaum, D., Rizzotto, N. (2019). Evaluation of financial statements fraud detection research: A multi-disciplinary analysis. International Journal of Disclosure and Governance, 16(4): 206-241. https://doi.org/10.1057/s41310-019-00067-9

[8] Deng, Q. (2010). Detection of fraudulent financial statements based on Naïve Bayes classifier. In 2010 5th International Conference on Computer Science & Education, Hefei, China, pp. 1032-1035. https://doi.org/10.1109/ICCSE.2010.5593407

[9] Chen, S., Goo, Y.J.J., Shen, Z.D. (2014). A hybrid approach of stepwise regression, logistic regression, support vector machine, and decision tree for forecasting fraudulent financial statements. The Scientific World Journal, 2014(1): 968712. https://doi.org/10.1155/2014/968712

[10] Chen, S. (2016). Detection of fraudulent financial statements using the hybrid data mining approach. SpringerPlus, 5(1): 89. https://doi.org/10.1186/s40064-016-1707-6

[11] Goel, S., Uzuner, O. (2016). Do sentiments matter in fraud detection? Estimating semantic orientation of annual reports. Intelligent Systems in Accounting, Finance and Management, 23(3): 215-239. https://doi.org/10.1002/isaf.1392

[12] Huang, S.Y., Tsaih, R.H., Yu, F. (2014). Topological pattern discovery and feature extraction for fraudulent financial reporting. Expert Systems with Applications, 41(9): 4360-4372. https://doi.org/10.1016/j.eswa.2014.01.012

[13] Kirkos, E., Spathis, C., Manolopoulos, Y. (2007). Data mining techniques for the detection of fraudulent

financial statements. Expert Systems with Applications, 32(4): 995-1003. https://doi.org/10.1016/j.eswa.2006.02.016
[14] Li, X., Xu, W., Tian, X. (2014). How to protect investors? A GA-based DWD approach for financial statement fraud detection. In 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC), San Diego, CA, USA, pp. 3548-3554. https://doi.org/10.1109/SMC.2014.6974480
[15] Wu, X.G., Du, S.Y. (2022). An analysis on financial statement fraud detection for Chinese listed companies using deep learning. IEEE Access, 10: 22516-22532. https://doi.org/10.1109/ACCESS.2022.3153478
[16] Andayani, W., Wuryantoro, M. (2023). Good corporate governance, corporate social responsibility and fraud detection of financial statements. International Journal of Professional Business Review, 8(5): e01897. https://doi.org/10.26668/businessreview/2023.v8i5.1051
[17] Sabatian, Z., Hutabarat, F.M. (2020). The effect of fraud triangle in detecting financial statement fraud. Jurnal Akuntansi, 10(3): 231-244. https://doi.org/10.33369/j.akuntansi.10.3.231-244
[18] Sorkun, M.C., Toraman, T. (2017). Fraud detection on financial statements using data mining techniques. Intelligent Systems and Applications in Engineering, 5(3): 132-134. https://doi.org/10.18201/ijisae.2017531428
[19] Othman, I.W. (2021). Financial statement fraud: Challenges and technology deployment in fraud detection. International Journal of Accounting and Financial Reporting, 11(4): 1-11. https://doi.org/10.5296/ijafr.v11i4. 19067
[20] Anisykurlillah, I., Ardiansah, M.N., Nurrahmasari, A. (2022). Fraudulent financial statements detection using fraud triangle analysis: Institutional ownership as a moderating variable. Accounting Analysis Journal, 11(2): 138-148. https://doi.org/10.15294/aaj.v11i2.57517
[21] Yadav, A.K.S., Sora, M. (2021). An optimized deep neural network-based financial statement fraud detection in text mining. 3c Empresa: Investigación y Pensamiento Crítico, 10(4): 77-105. https://doi.org/10.17993/3cemp.2021.100448.77-105
[22] Omeir, A.K., Vasiliauskaitė, D., Soleimanizadeh, E. (2023). Detection of financial statements fraud using Beneish and Dechow models. Journal of governance and regulation, 12(3): 334-344. https://doi.org/10.22495/jgrv12i3siart15
[23] Beneish, M.D. (1997). Detecting GAAP violation: Implications for assessing earnings management among firms with extreme financial performance. Journal of Accounting and Public Policy, 16(3): 271-309. https://doi.org/10.1016/S0278-4254(97)00023-9
[24] Dechow, P.M., Ge, W., Larson, C.R., Sloan, R.G. (2011). Predicting material accounting misstatements. Contemporary Accounting Research, 28(1): 17-82. https://doi.org/10.1111/j.1911-3846.2010.01041.x
[25] Zhao, Z., Bai, T. (2022). Financial fraud detection and prediction in listed companies using SMOTE and machine learning algorithms. Entropy, 24(8): 1157. https://doi.org/10.3390/e24081157
[26] Indrati, M., Claraswati, N. (2021). Financial statement detection using fraud diamond. Journal Research of Social Science, Economics, and Management, 1(2): 148-162. https://doi.org/10.59141/jrssem.v1i2.13
[27] Humphrey, E.A., Isenmilia, P.A., Omoye, A.S. (2023). Fraud pentagon: Detection of financial statement fraud in a firm. Mediterranean Journal of Social Sciences, 14(6): 102-113. https://doi.org/10.36941/mjss-2023-0040
[28] Malik, E. F. (2022). Fraud detection by machine learning techniques. Applied Mathematics and Computational Intelligence, 11(1): 88-103.
[29] Alwadain, A., Ali, R.F., Muneer, A. (2023). Estimating financial fraud through transaction-level features and machine learning. Mathematics, 11(5): 1184. https://doi.org/10.3390/math11051184
[30] Rabade, S.U. (2022). Use of machine learning in financial fraud detection: A review. International Journal of Advanced Research in Science, Communication and Technology, 2(3): 38-44. https://doi.org/10.48175/IJARSCT-7595
[31] Saleh, M.M.A., Aladwan, M., Alsinglawi, O., Salem, M.O. (2021). Predicting fraudulent financial statements using fraud detection models. Academy of Strategic Management, 20(3): 1-17.
[32] Chukwuma, O.V., Okolie, P.I., Eneh, N.A., Ejike, S.I. (2023). Using data analytics techniques for the detection of accounting fraud in financial statements. International Journal of Multidisciplinary Research and Growth Evaluation, 4(1): 212-214.
[33] Cheng, C.H., Kao, Y.F., Lin, H.P. (2021). A financial statement fraud model based on synthesized attribute selection and a dataset with missing values and imbalanced classes. Applied Soft Computing, 108: 107487. https://doi.org/10.1016/j.asoc.2021.107487
[34] Antawirya, R.D.E.P., Putri, I.G.A.M.D., Wirajaya, I.G.A., Suaryana, I.G.N.A., Suprasto, H.B. (2019). Application of fraud pentagon in detecting financial statement fraud. International Research Journal of Management, IT and Social Sciences, 6(5): 73-80. https://doi.org/10.21744/irjmis.v6n5.706
[35] Ariyanto, D., Jhuniantara, I.M.G., Ratnadi, N.M.D., Putri, I.G.A.M.A.D., Dewi, A.A. (2021). Detecting fraudulent financial statements in pharmaceutical companies: Fraud pentagon theory perspective. Accounting, 7: 1611-1620. https://doi.org/10.5267/j.ac.2021.5.009
[36] Dechow, P. M., Hutton, A. P., Kim, J. H., Sloan, R. G. (2012). Detecting earnings management: A new approach. Journal of Accounting Research, 50(2): 275-334. https://doi.org/10.1111/j.1475-679X.2012.00449.x
[37] Skousen, C.J., Smith, K.R., Wright, C.J. (2009). Detecting and predicting financial statement fraud: The effectiveness of the fraud triangle and SAS No. 99. In Corporate governance and firm performance. Emerald Group Publishing Limited., 13: 53-81. https://doi.org/10.1108/S1569-3732(2009)0000013005
[38] Ye, H., Xiang, L., Gan, Y. (2019). Detecting financial statement fraud using random forest with SMOTE. IOP Conference Series: Materials Science and Engineering, 612(5): 052051. https://doi.org/10.1088/1757-899X/612/5/052051
[39] Sawangarreerak, S., Thanathamathee, P. (2021). Detecting and analyzing fraudulent patterns of financial statement for open innovation using discretization and association rule mining. Journal of Open Innovation: Technology, Market, and Complexity, 7(2): 128. https://doi.org/10.3390/joitmc7020128