IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 4 May 2021, accepted 21 June 2021, date of publication 13 July 2021, date of current version 15 July 2022.

Digital Object Identifier 10.1109/ACCESS.2021.3096799

Intelligent Fraud Detection in Financial Statements Using Machine Learning and Data Mining: A Systematic Literature Review

MATIN N. ASHTIANI AND BIJAN RAAHEMI

Knowledge Discovery and Data Mining Laboratory, Telfer School of Management, University of Ottawa, Ottawa, ON K1N 6N5, Canada

Corresponding author: Matin N. Ashtiani (mnaja036@uottawa.ca)

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant RGPIN/341811-2012.

ABSTRACT Fraudulent financial statements (FFS) are the results of manipulating financial elements by overvaluing incomes, assets, sales, and profits while underrating expenses, debts, or losses. To identify such fraudulent statements, traditional methods, including manual auditing and inspections, are costly, imprecise, and time-consuming. Intelligent methods can significantly help auditors in analyzing a large number of financial statements. In this study, we systematically review and synthesize the existing literature on intelligent fraud detection in corporate financial statements. In particular, the focus of this review is on exploring machine learning and data mining methods, as well as the various datasets that are studied for detecting financial fraud. We adopted the Kitchenham methodology as a well-defined protocol to extract, synthesize, and report the results. Accordingly, 47 articles were selected, synthesized, and analyzed. We present the key issues, gaps, and limitations in the area of fraud detection in financial statements and suggest areas for future research. Since supervised algorithms were employed more than unsupervised approaches like clustering, the future research should focus on unsupervised, semi-supervised, as well as bio-inspired and evolutionary heuristic methods for anomaly (fraud) detection. In terms of datasets, it is envisaged that future research making use of textual and audio data. While imposing new challenges, this unstructured data deserves further study as it can show interesting results for intelligent fraud detection.

INDEX TERMS Fraud detection, financial statement, machine learning, data mining, outlier detection, systematic literature review.

I. INTRODUCTION

Financial fraud refers to the use of fraudulent and illegal methods or deceptive tactics to gain financial benefits. Fraud can be committed in different areas of finance, including banking, insurance, taxation, and corporates, and more. Fiscal fraud and evasion, including credit card fraud, tax evasion, financial statement fraud, money laundry, and other types of financial fraud, has become a growing problem. Despite efforts to eliminate financial fraud, its occurrence adversely affects business and society as hundreds of millions of dollars are lost to fraud each year. This significant financial loss has dramatically affected individuals, merchants, and banks.

Nowadays, fraud attempts have increased drastically, which makes fraud detection more important than ever. The Association of Certified Fraud Examiners (ACFE) has announced that 10% of incidents concerning white-collar crime involves falsification of financial statements [56]. They classified occupational fraud into three types: asset misappropriation, corruption, and financial statement fraud. Financial statement fraud resulted in the most significant losses among them. Although the occurrence frequency of asset misappropriation and corruption is much higher than financial statement fraud, the financial implications of these latter crimes are still far less severe. In particular, as reported in a survey from EisnerAmper, which is among the prominent accounting firms in the U.S., "the average median loss of financial statement fraud ($800,000 in 2018) accounts for over three times the monetary loss of corruption ($250,000) and seven times as much as asset misappropriation ($114,000)" [43].

The focus of this study is on financial statement fraud. Financial statements are documents that describe details about a company, specifically their business activities and financial performance, including income, expenses, profits, loans, presumable concerns that may emerge later, and managerial comments on the business performance [25], [59].

The associate editor coordinating the review of this manuscript and approving it for publication was Sergio Consoli.

All firms are obligated to announce their financial statements in a quarterly and annual manner. Financial statements can be used to indicate the performance of a company [59]. Investors, market analysts, and creditors exploit financial reports to investigate and assess the financial health and earnings potentials of a business.

Financial statements consist of four sections; income statement, balance sheet, cash flow statement, and explanatory notes. The income statement places a great emphasis on a company's expenses and revenues during a specific period. The company's profit or net income is provided in this section, which subtracts expenses from revenues. The balance sheet provides a timely snapshot of liabilities, assets, and stockholders' equity. The cash flow statement measures the extent to which a company is successful in making cash to fund its operating expenses, fund investments, and pay its debt obligations. Explanatory notes are supplemental data that provide clarification and further information about particular items published financial statements of a company. These notes cover areas including disclosure of subsequent events, asset depreciation, and significant accounting policies, which are necessary disclosures that demonstrate the amounts reported on the financial statements.

Financial statement fraud involves falsifying financial statements to pretend the company more profitable than it is, increase the stock prices, avoid payment of the taxes, or get a bank loan. Fraud triangle in auditing is a framework to demonstrate the motivation behind an individual's decision to commit fraud. It is built upon the fraud triangle theory that was proposed by [13]. The fraud triangle has three elements that increase the risk of fraud: incentive, rationalization, and opportunity, which, together, lead to fraudulent behavior. Auditing professionals have extensively used this theory to explain the motivation behind an individual's decision to commit fraud. It is indispensable to understand the fraud triangle to evaluate financial fraud [59]. Gupta and Singh [27] suggested that when there are incentives such as the obligation to achieve an outcome or cover losses, the potential for fraud increases. The company will encounter temptations or pressures to adopt fraudulent practices. Moreover, the lack of inspections or unsuccessful controls provides a favorable occasion for committing fraud. Rationalization happens when the fraudster aims to justify the fraudulent action, and it could be affected by the others and the conditions. Dbouk and Zaarour [13] remarked that people who perpetrate fraud incline to stay inside their moral safe zone. Therefore, the fraudster inwardly attempts to legitimatize and defend the fraudulent behavior in preparation for committing the first fraud. Dbouk and Zaarour [13] indicated that rationalization occurs when the committer constructed a justification for the fraud and not desired to be deemed an offender. This situation enables fraudsters to consider their dilemma as a particular exemption rather than criminal behavior.

Traditional methods of fraud detection, including manual detection, are not only costly, imprecise, and time-sapping, but also impractical [73]. Activities are conducted to minimize losses resulting from fraudulent actions, but they are not too effective. Artificial intelligence, especially machine learning technologies, turned out to be one of the greatest thriving methods in fraud discovery. Data mining contributes to identify fraud and act immediately to lower overheads. Millions of statement documents can be searched through data mining techniques to spot patterns and identify fraudulent statements [65].

In most cases, prevailing fraud detection techniques have a common data mining rationale, but they may differ in many facets with specific domain knowledge [7]. The goal of financial statement fraud detection (FSFD) is to categorize financial statements into fraudulent or non-fraudulent. Both supervised and unsupervised methods were used to predict fraudulent statements. Classification has been the most popular technique to identify fraudulent financial statements [79]. Most FSFD practices employ supervised machine learning strategies [4], [5], [22] that generally have a two-stage scheme. A model is trained on a dataset containing feature vectors and the class labels in the first stage. Afterward, in the next stage, test samples are classified using the trained model. The performance of machine learning/data mining (ML/DM) algorithms is directly associated with the way feature vectors are extracted from the input data and how informative they are. Selecting inappropriate features may lead to irrelevant or meaningless features and weak performance [7].

This work presents a systematic literature review (SLR) in the scope of intelligent financial statement fraud detection. The primary focus of this systematic literature review is on identifying the ML/DM techniques and datasets employed for FSFD. Furthermore, we aim to analyze the gaps and uncover the trend of research in this area (from the beginning to the most recent studies). To the best of our knowledge, there is no systematic literature review in intelligent FSFD that investigates the datasets and ML/DM methods. More recently, a review article is published in financial statements fraud detection [2]. They combined prior multi-disciplinary literature on FSFD. There are several significant differences between their article and our study, as we will mention some of them in the following. First, their main goal is to combine findings from different domains such as accounting, information systems, and analytics, while our purpose is to discover the datasets and applied ML/DM methods for FSFD. Second, their review protocols (i.e., including search string, explored digital libraries, selection criteria, year span) are quite different than ours. For instance, their search string is ''Management Fraud OR Financial Statement Fraud'' which also includes management frauds. Moreover, they did not consider conference articles while we include them in our study.

The other recently published reviews have focused on the specific areas of finance, such as credit card fraud detection [53], fraud prediction in bank credit administration [60], online banking fraud detection [64], and payment card fraud detection [62]. The scope of our systematic literature review, however, is different from previous ones. Figure 1 shows

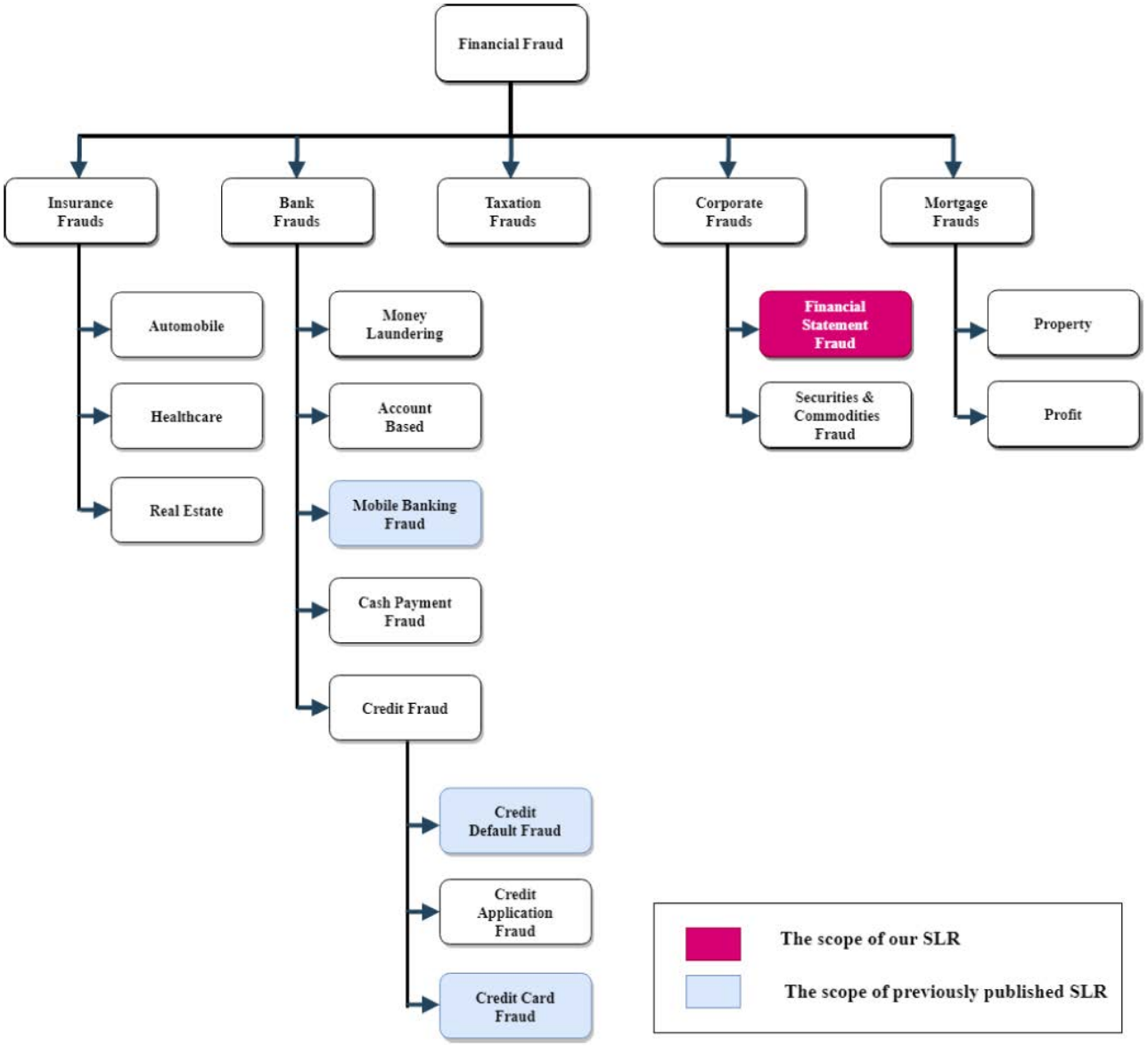


FIGURE 1. The scope of our systematic literature review compared to the published literature.

the focus of the other review articles in the finance area, compared to the focal point of our study.

The purpose of the current SLR is to review the state-of-the-art ML/DM methods applied for FSFD. Furthermore, it will guide researchers in their choice of high-performance ML/DM techniques, as well as the datasets to be considered to predict fraudulent actions in financial statements.

This SLR follows a well-defined sequence of systematic research steps that follow the Kitchenham methodology [42]. By Kitchenham's definition, an SLR is ''a means of identifying, evaluating, and interpreting all available research relevant to a particular research question, or topic area, or phenomenon of interest'' [42]. We made use of Covidence application (https://www.covidence.org/) in order to perform this SLR. Covidence is a web-based platform that standardizes and simplifies the production of systematic literature reviews, including Cochrane reviews.

The primary goal of this study is to seek the answers to our two research questions. Concerning the defined search questions, we explored our specified search string in five major digital libraries. This leads to obtaining 187 papers, including journal articles and conference proceedings. Additionally, we employed the snowballing technique to manually include other related articles manually, which we did not catch during the automatic searching [74]. We have followed the references in the articles, especially review articles, to mature our articles and find the best answers for our research questions.

Finally, 47 papers were selected, according to the SLR procedure, for further analysis. We extract the required information to address the research questions and conclude the results.

The rest of the paper is organized as follows: In Section 2, we outline our research method: search string, study selection, data extraction, and quality assessment. Section 3 provides the results of the SLR and the answers to the research questions. Section 4 addresses possible threats to the validity of this SLR. Finally, we conclude by summarizing the outcomes of this study in Section 5.

II. RESEARCH METHOD

A systematic literature review (SLR) is undertaken to study the current status of research in the FSFD area and address the research questions.

Inspired by Kitchenham, we followed the following steps to develop our SLR protocol [42].

A. DEFINITION OF THE RESEARCH QUESTIONS

This study attempts to answer the following questions:

- (RQ1) What fraud detection techniques and datasets related to financial statements were employed in the literature?
- (RQ2) What are the gaps, trends of research, and future research directions in this area?

The first question helps us to compile and analyze the state-of-the-art machine learning techniques and the most popular data sources used for intelligent FSFD. The second one guides us to identify trends of the research, gaps in the present studies, and the future direction of the researches in this domain.

1) DEVELOPING SEARCH STRATEGY

It is necessary to identify appropriate search concepts and keywords. In order to find the most related studies, we choose the below search string to obtain them in five different digital libraries:

"fraud*"
AND
"financial *statement*"
AND
"Artificial Intelligence" OR "machine learning" OR "data mining"

We included the "Artificial Intelligence" OR "Machine Learning" OR "Data Mining" to make sure that we can discover the studies that used intelligent techniques. Besides, we embedded the "financial *statement*" term to focus on articles that worked on the financial statement fraud.

We searched the above search string in five prominent digital libraries, i.e., ACM Digital Library, IEEE Xplore, Scopus, Web of Science, and ScienceDirect. The search string is modified and translated to the proper input query for searching each digital library. Appendix A presents the search queries in detail. We only focused on the peer-reviewed journal and conference articles and excluded the book chapters and other types of publications. The search was conducted in April 2020, and there was no limitation for the publication year. A total number of 187 articles were retrieved from all the search libraries. We declared the distribution of the articles over the libraries in Table 1. We discovered 63 duplicate articles among the explored articles. After removing the duplicates, we moved forward to the selection process based on the remaining 124 articles. Additionally, a snowballing technique was adopted as a supplement to the previous automatic digital library searches. The snowballing technique requires an initial collection of articles. Backward snowballing refers to discovering more related papers from the reference lists of this collection. Identify more papers that have cited the collection of papers found during the search called forward snowballing. This study only considered backward snowballing, where a total of 8 articles is obtained during this procedure.

B. SELECTION OF PRIMARY STUDIES

Further studies were included during backward snowballing based on the inclusion criteria. The retrieved articles from the digital libraries were excluded based on the exclusion criteria. Table 2 presents both inclusion and exclusion criteria.

- **Duplicate removal:** Our primary search included 187 articles. After removing the 63 duplicated articles, 124 studies remained to be screened.
- **Title and abstract screening:** Articles are screened out by filtering titles, abstracts, and keywords based on the inclusion and exclusion criteria, and 47 irrelevant papers are removed.
- **Full text screening:** From the 77 articles that were left, we excluded 35 studies according to the exclusion criteria.
- **Quality assessment:** From the 42 articles that were left, we excluded three studies because they were not qualified according to our quality assessment criteria.
- **Snowballing inclusion:** We included eight articles by using the backward snowballing technique.

C. DATA EXTRACTION

This step entails deriving relevant data and information from the selected papers. Table3 shows our data extraction form.

We exploit the extracted data in order to answer our two main research questions. In the last column of Table3, we specified the purpose of extracting the corresponding data. We used the details about techniques and datasets to answer RQ1. We analyzed this information to group similar studies together, in terms of their datasets and techniques. Extracting the objective and the conclusion of each study will help us recognize the trends of the works, analyze the gaps, and determine future research (RQ2). Therefore, we summarized the articles according to their goals and conclusions to find the gaps and recognize the direction of future research. We reported the results of our synthesis in two parts. The first part included a meta-analysis of the outcomes of the selected papers. The second part elaborated on the answers to the two research questions according to the extracted data.

TABLE 1. The number of Retrieved articles from five digital libraries.

No	Database Name	Web Address	# Initially Retrieved Articles
1	Scopus	https://www.scopus.com/	74
2	IEEE Xplore	http://ieeexplore.ieee.org/	31
3	ScienceDirect	http://www.sciencedirect.com/	10
4	Web of Science	https://webofknowledge.com/	64
5	ACM Digital Library	http://dl.acm.org/	8
Total Number of Retrieved Articles			187
Total Number of Duplicates			63
After Removing Duplicates			124

TABLE 2. The inclusion and exclusion criteria applied in this SLR.

Inclusion	I.1: Studies discovered via snowballing upon meeting quality and exclusion criteria
Exclusion	E1: Studies that do not focus on the fraudulent financial statement area (i.e., the studies that focused on other aspects of financial fraud detection such as bank frauds, telecom frauds, tax frauds, or even the other types of corporate frauds)
	E2: Purely theoretical studies that do not apply or implement any ML/DM techniques in their works.
	E3: Studies that did not focus on using ML/DM techniques for FSFD.
	E4: Studies in the form of posters, short papers, abstracts, and book chapters.
	E5: Studies that do not mention their data sources.
	E6: Studies that were not in English.
	E7: Review and survey studies.
	E8: For studies that found to have been published twice, and the authors are not different, the latest of the two was chosen.

TABLE 3. Data extraction form.

Strategy	Category	Description	Purpose
Automatic Extraction	Identifier	Identifier number (DOI)	Meta-analysis and supplementary information
	Title	The title of the study	
	Authors name(s)	Name(s) of the study authors	
	Publication year	Publication year	
	Document type	Conference proceedings, journal article	
	Venue	Name of the conference or journal	
Manual Extraction	Objectives	The objectives of the study	RQ2
	Conclusion	The outcome of the study	RQ2
	Technique name	The ML/DM approach used to support the objectives	RQ1
	Algorithm type	Supervised, unsupervised or semi supervised	RQ1
	Dataset	The name and type of the dataset used in the study	RQ1
	Future directions	Future directions, trends and gaps	RQ2

D. META-ANALYSIS

This section demonstrates the primary investigation results of the selected articles using a statistical approach to combine the outcomes from multiple studies. We answer the defined research questions based on the ultimately determined papers (47 papers) in the Results section.

1) DISTRIBUTION OF SELECTED PAPERS OVER THE YEAR

We set no restrictions on the year of publication. This enabled us to broadly investigate the entire published articles in the field of study, from the most traditional techniques to the newest ones. The publication years of the selected papers that met our criteria are varied between 1995 and 2019, while it suggests a rising interest in this area of research over the last four years, namely from 2016 to 2019. Figure 2 presents the distribution of the papers over the publication years, which confirms that a significant number of articles was published in 2016, 2017, and 2019. It is worth to mention that since our study was concluded in April 2020, papers published after that date were not included in our study.

2) PUBLICATION TYPE

We considered only peer-reviewed journal articles and conference proceedings. The majority of papers were journal

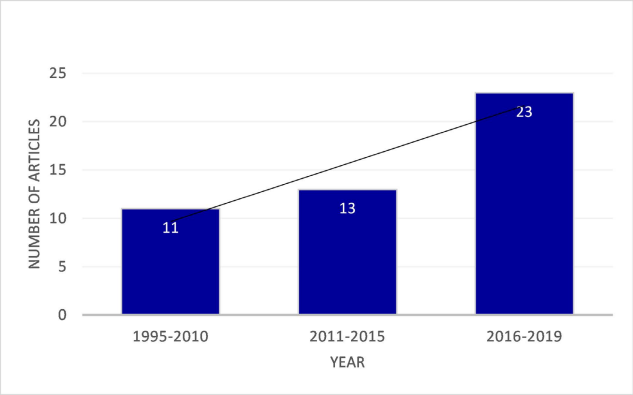


FIGURE 2. Distribution of the selected papers over the publication years.

articles and a fewer number of them were from conference proceedings (32%).

III. DATA SYNTHESIS RESULTS

In this section, we investigate all final selected articles (47 articles). The data is synthesized to address the two mentioned research questions.

RQ1: What fraud detection techniques and datasets related to financial statements were employed in the literature?

A. DATASETS

Prior studies employed various data to identify fraudulent statements. The data structure and the falsification cases of financial statements could vary notably along with the data sources. The performance of fraud detection models could be seriously affected by any small variations in the number of cases. Therefore, investigating the datasets and their specifications is valuable. The data differ in terms of data type (number, text, voice) and data source from which they have been obtained (financial statements, or external sources). Moreover, articles used various financial references to label the input vectors and determine if the data from one company is fraudulent. We analyzed these references in our SLR as well. We synthesized the employed datasets in the literature in terms of the following factors:

- Data type
- Data source
- Labeling reference

1) DATA TYPE

Generally, data can be sorted into one of the two categories: structured and unstructured data. Structured data has a predefined schema, and it can be stored and searched in relational databases structured query language (SQL). In contrast, there is no predefined schema for unstructured data at the time of storing the data in the database. As such, the format of the database and the query language to search and retrieve data is based on Not-only SQL (NoSQL) techniques. Unstructured data is much more challenging to be aggregated, processed, and analyzed. A large number of fraud detection techniques employing ML/DM techniques are concerned with structured data performing quantitative methods, which is in line with our findings in this study [41]. Machine learning algorithms can imitate the procedure of processing unstructured data, namely speech and text, to enhance the performance of evaluating financial statements [6], [23], [67]. The studies relied on structured more than unstructured data. Figure 3 illustrates that structured data (such as financial ratios and non-financial ratios) were employed more frequently than unstructured ones (i.e., textual or vocal data). The second level of this figure shows the number of articles used structured data compared to articles used unstructured data. It should also be noted that several articles utilized a combination of structured and unstructured data in their studies [8], [21], [30], [31], [47].

TABLE 4. Data types.

Data type	References	Frequency
Financial ratios only	[58, 40, 15] [17, 50, 80] [9, 59, 55] [81, 33, 48] [57, 49, 37] [52, 38, 61] [30, 54, 69] [51, 39, 29] [14, 11] [44]	27
Textual content only	[18, 25] [19, 26]	4
Vocal speech and textual content	[32]	1
Vocal speech, textual content, and financial ratio	[71]	1
Financial and non-financial ratios	[10, 16, 36] [58, 76, 78]	6
Financial ratios and textual content	[47, 37, 20] [30, 8]	5
Non-financial ratios and textual content	[34]	1

We categorized the datasets in terms of the data type into four categories. Figure3 presents these four categories as well as the distribution of the articles over these categories. The following is a description of these data types. It should be noted that ten articles used a combination of two data types in their works. The other finding regarding the data type is that no articles were used non-financial ratios alone for the detection purpose. Table 4 shows more details regarding the data types in the reviewed articles.

a: FINANCIAL RATIOS

Financial ratios are calculated based on numerical values collected from financial statements to obtain meaningful information about a company. These numbers can be observed in various sections of a company's financial statements, particularly balance sheets, cash flow statements, and income statements. They are utilized to conduct quantitative analysis, and evaluate the performance of a company, assess a company's growth, profitability, and valuation. They are measurements to compare similar businesses in their industry.

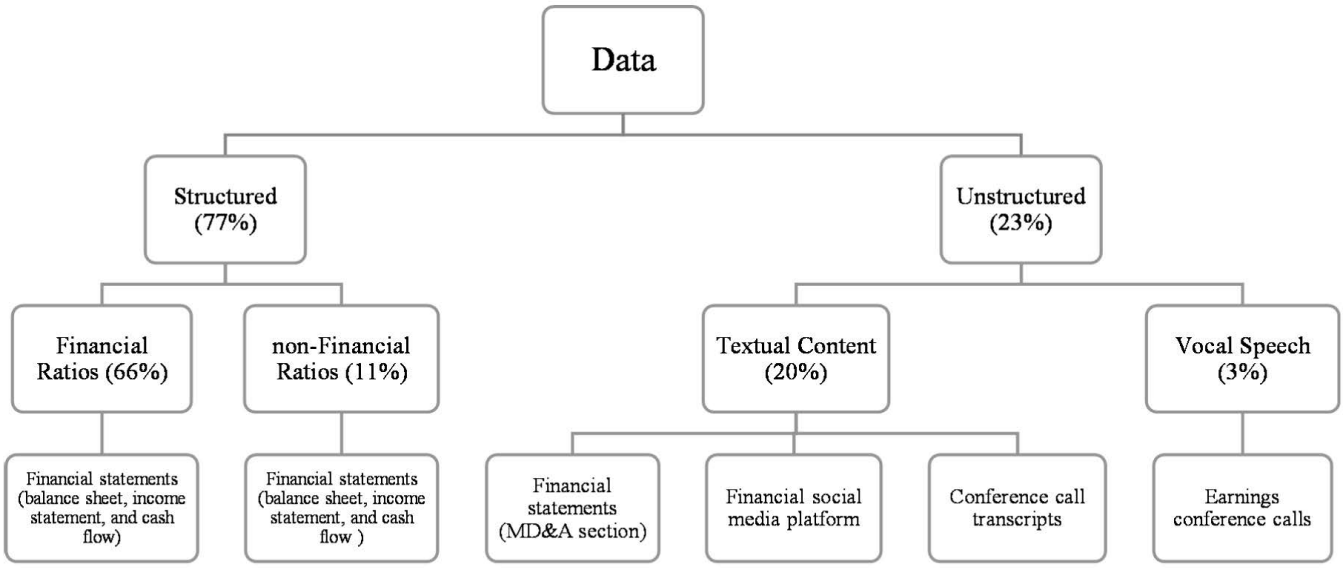


FIGURE 3. Data types. The second row shows the categorization of the data types based on the structure of the data. The third row represents the examples of data employed in the literature. Finally, the last row demonstrates that each data type is originated from which sources.

Liquidity	Leverage	Efficiency	Profitability	Market Value
•Current •Acid-test • Cash •Operating cash flow	•Debt •Debt to equity •Interest coverage •Debt service coverage	•Asset turnover •Inventory turnover •Accounts receivable turnover •days sales in inventory	•Gross margin •Operating margin •Return on assets •Return on equity	•Book value per share •Dividend yield •Earnings per share •Price-earnings

FIGURE 4. Financial ratios.

Financial ratios are grouped into five categories: leverage ratios, efficiency ratios, market value ratios, liquidity ratios, and profitability ratios. Figure 4 shows some of ratios within financial statements for each category. The first attempts for detecting fraud in financial statements using automated techniques were employed only the financial variables/ratios from the financial statement documents. Later, multiple studies have integrated both financial and non-financial ratios as an input data source for feature extraction (Table 4).

b: NON-FINANCIAL RATIOS

It is essential for investors to review non-financial information, as well. Non-financial information could influence a company's return, like the quality of the company's management, its competitors, and its status. Information such as competitive considerations, workforce, technological changes, quality of management, the state of the economy, and market forces are not immediately manifested in a company's financial statements. These are considered as non-financial information, and some studies benefit from such variables as input data in their prediction method. According to our analysis, the non-financial data is targeted far less than the financial variables in the literature (Figure 3). A list of the non-financial ratios employed in the reviewed articles is shown in Table 5.

c: TEXTUAL CONTENTS

Public companies must outline a managerial report of their decisions and operations during the past year as well as the landscapes of the forthcoming year in Management's Discussion and Analysis section (MD&A) of the financial statements.This information gives a comprehensive understanding of the business and financial state of corporations to the readers (investors). We will discuss it in the next section in more detail. More recently, the textual contents and the data from a CEO's vocal speech are taken into consideration in the literature [32], [71]. Textual data can play an essential role in detecting fraudulent and misrepresented financial reports for a couple of reasons. First, there is a significant amount of textual data, including the data from the MD&A section of the financial statements and the transcripts of the earnings conference call available and accessible to be processed. The abundance of this textual data promoted text mining applications in FSFD and associated financial decision-making problems such as stock market prediction, FOREX prediction, and customer relationship management [46]. Second, former studies stated that text mining approaches are investigated relatively less for fraud detection problems rather than the other problems in the financial domain [28]. Third, the textual data from the MD&A section of the financial statements or CEO speeches in earning conference calls are managerial reports and directly related to the managers. On the other hand, it is easier for senior managers to perpetrate fraudulent behaviors due to their ability, opportunity, and incentives to commit such actions [71]. Therefore, the textual information is indispensable and empowers intelligent fraud detection methods.

TABLE 5. Non-financial ratios.

ID	Non-financial ratios
1	The major stockholders' stockholding ratio [16, 76, 77]
2	Duality of board director and CEO [16]
3	Size of the board of directors [16, 76, 77, 78]
4	The proportion of independent directors [76, 77]
5	The ratio of pledged stocks held by directors and supervisors [10, 16, 78]
6	The ratio of stocks held by directors and supervisors [16]
7	Audited by BIG4 (the big four CPA firms) [16]
8	Number of outside supervisors [16, 76]
9	Audited by BIG4 (the big four CPA firms) or not: 1 for companies audited by BIG4, otherwise is 0. [36, 77] (Big 4 accounting firms refer to PWC, KPMG, DTT, and EY)
10	Restatement of financial statements or not: 1 is for restatement; 0 for non-restatement [36]
11	Type of audit report: 1 is for qualified opinion; 0 is for unqualified opinion [36, 77]
12	LHSR [76]
13	Ratio of external directors [77, 78]
14	Total shareholding ratio of the top ten shareholders [77]
15	The percentage of board members who are family members [78]
16	The percentage of board members who are executives [78]
17	The proportion of shares owned by institutional StartFragmentshareholders [78]
18	Percentage of shares held by the largest StartFragmentstockholders [78]
19	Percentage of shares held by family members [78]
20	Ownership concentration [78]

d: VOCAL SPEECH

Two studies used a vocal speech to analyze financial statement frauds. Humpherys *et al*. [32] is one of the earliest studies proposing that CEO's speeches are an informative source of data for FSFD. They extracted vocal cues (verbal and non-verbal) from the CEO's speeches and performed experiments to find a relationship between the vocal data and human emotions. They employed a technology called layered voice analysis (LVA) to investigate if the financial reports are fraudulent. In LVA technology, the goal is to associate vocal variables with critical human sentiment to identify misleading intentions in real-world scenarios. This study revealed that certain key nonverbal signals in the initial parts of CEOs' speeches in the earnings conference calls enable distinguishing fraudulent financial statements from non-fraudulent ones. [32]. The other research employed a combination of textual, financial ratios, and vocal features to develop an FSFD method [71]. Financial ratios, textual data from the MD&A section, and vocal attributes are potentially worthwhile in FSFD. Yang [71] investigated to what extent the simultaneous incorporation of these three sources of data improves the performance of an FSFD tool. They analyzed a combination of three types of features and compared the performance of the detection approach in this case with isolated features from only one source. Their results demonstrated that employing a fusion of numeric, textual, and vocal features enhances the accuracy of the fraud detection approach. Their research was the only study that combined three data categories for FSFD.

2) DATA SOURCE

Researchers extract the data from different parts of the financial statement documents. The fourth row of Figure 3 demonstrates that each data type is originated from which sources. Financial statements contain financial ratio/variables data that has been used widely in the literature. These data were gathered from the cash flow, income statement, and balance sheet parts of financial statements. Furthermore, the articles in which textual contents have been analyzed provided the data from the Management Discussions and Analysis (MD&A) section of the financial statements. This section provides a general picture of the performance of the company, management's upcoming projections, and the company's current financial state. It supports potential investors to figure out the company's management's speculations, beliefs, financial fundamentals, and performance. MD&A is a mandatory disclosure for publicly traded companies under the U.S. Securities and Exchange Commission jurisdiction.

Two studies among the reviewed literature applied financial social media data to perform fraud detection [19], [20]. Dutta *et al.* [20] exploited the unstructured data from a financial social media platforms called SeekingAlpha as their data source and extracted lexical, sentiment, and social media features from the input data. Financial social media platforms incorporate valuable commentary and analysis information about firms, including analyst reports, merger and acquisition, and management disclosures. Natural language processing and machine learning techniques are employed in their study to evaluate companies' fraud risks [20]. Furthermore,

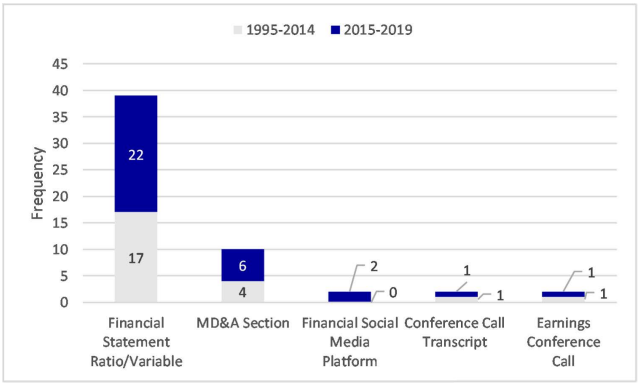


FIGURE 5. Data sources.

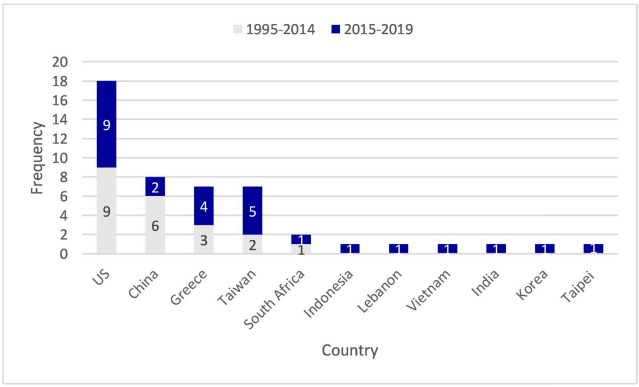


FIGURE 6. Data source by country.

Dong *et al.* [19] employed SeekingAlpha in addition to other financial information datasets to detect fraudulent financial statements.

Both textual data from earnings conference calls (transcripts), and vocal data from earning conference calls were considered as the data source in literature [32], [71]. An earnings call is a conference call between the management of a public company, investors, analysts, and the media to discuss the company's financial outcomes during a given reporting interval, namely a quarter or a fiscal year. Earnings transcripts are available through various financial platforms like SeekingAlpha. The length of the earnings calls are variable and it depends on the firm's market capitalization. Most of the conference calls last for 46 to 60 minutes. Figure 5 shows the distribution of the different data sources among literature for two categories of years period. We will discuss the annual trend in the next section.

3) DATASET BY COUNTRY

Studies worked on various sources of data from a range of countries. Figure 6 shows the countries that the data from their firms employed for FSFD. It is evident from the figure that the data from United States firms have been used widely in the literature. Furthermore, China, Greece, and Taiwan are the next most frequent countries that worked on the FSFD problem.

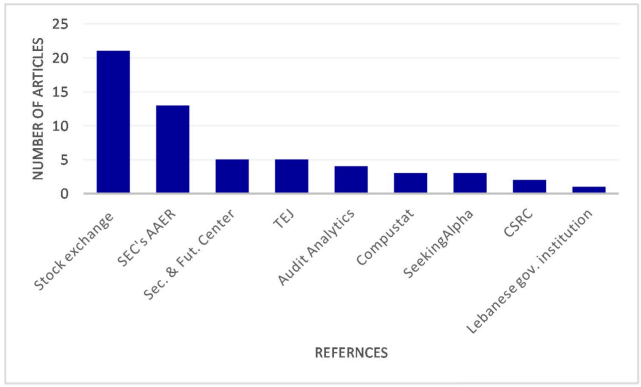


FIGURE 7. Reference databases for detecting fraud status of companies.

4) FINANCIAL REFERENCES

Generally, fraudulent companies are distinguished from the non-fraudulent ones based on any evidence from financial databases confirming financial statement fraud in the auditors' reports. There are various financial databases that studies relied upon them to find fraudulent firms. Figure 7 represents these financial references, which are outlined in the following.

Most of the articles gathered the fraudulent firm data from the Stock Exchange. It could not be certified that the financial statements of the non-fraudulent companies would not be misreported or any fraudulent comportment would not be disclosed subsequently. It only assures that no fraudulent financial statements were detected in an extensive relevant search. Since the studies were conducted to investigate fraud in different countries, the datasets are diversified against countries.

The US government created an independent federal regulatory agency called the Securities and Exchange Commission (SEC) to maintain regular operation of the securities markets and protect investors. The SEC summarizes the accounting-based enforcement efforts in Accounting and Auditing Enforcement Releases (AAERs). These AAERs give an overview of violating financial statements during or after an investigation against a company for alleged auditing or accounting violations.

Five articles exploited Taiwan Securities and Futures Bureau and the group litigation samples proclaimed by the Securities and Futures Investors Protection Center to find the fraudulent firm financial statements in Taiwan. In four studies, fraudulent enterprises were selected from the listed and OTC (trading through decentralised dealer networks) companies of the Taiwan Economic Journal Data Bank (TEJ) [10], [33], [48], [78].

Audit Analytics is a financial database that covers all restatement cases since 2000. The database incorporates an extensive set of both unintentional and intentional restatement cases. Researchers in [39] stated that the Audit Analytics database is the most suitable because of its comprehensive inclusion of all restatement cases. The target of their work was to utilize all restatement cases, including both fraudulent and

non-fraudulent ones. Consequently, they were required to use comprehensive data sources on restatement cases, like Audit Analytics.

COMPUSTAT is another database of fundamental financial and market information for global companies (www.compustat.com). It is a comprehensive database including both active and inactive global companies, indices, and industries. Several studies used the firm's data from COMPUSTAT database [18], [21], [50], [58].

SeekingAlpha (seekingalpha.com) is a service launched in early 2004 and powered by crowd-sourced content for investment research. It has developed to be a good target for financial market analysis, discussions, and opinions. There are four million registered users on this platform. For each firm, contents under breaking news, analysis reports, and StockTalk sections are extracted into a database. Breaking news and analysis reports are created by SeekingAlpha editors and Analysts, and platform contributors, respectively. In addition, StockTalk contains user-generated information and opinions like tweets on Twitter. In summary, SeekingAlpha is a representative financial, social media platform for investment research that is selected as a data source to illustrate how to utilize social media information in two studies [19], [20]. Contents for each firm in SeekingAlpha were separated into two parts according to the time point of fraud disclosure by SEC.

Five out of eight articles from China used the firms punished by the China Securities Regulatory Commission (CSRC) for violating financial statement disclosure standards [16], [33], [36], [48], [76].

Finally, there is only one study from Lebanon, where the financial statements from the database of a Lebanese governmental financial audit institution were extracted using its audit software. [49].

5) DATASET SIZE

Figure 8 presents the number of fraudulent and non-fraudulent samples for each dataset. A study with the largest dataset, including 15985 financial statements(51 fraudulent and 15934 non-fraudulent) and the lowest ratio (0.003), was excluded from the histogram to make it more readable [55]. According to Figure 8, most of the studies made use of datasets with a size of less than 260 data samples. Furthermore, among the studies with a larger size of data, generally, the number of fraudulent samples is much less than the number of non-fraudulent ones. 14 studies used a balanced dataset, which means an equal number of fraudulent and non-fraudulent financial statements were used.

B. MACHINE LEARNING AND DATA MINING METHODS

Machine learning refers to analytic techniques that uncover patterns in data without the guidance of a human analyst or expert. Machine learning efficiently facilitates the determination about which statements are most likely to be fraudulent. These techniques are incredibly useful in fraud detection and prevention since they enable automated pattern recognition over a large amount of data. Adopting proper machine learning models can distinguish fraudulent and legitimate behaviors. These intelligent methods could be adapting over time to unseen, new fraud tactics. Since there is a demand for interpreting patterns in the data and applying data science to constantly improve the capability of distinguishing normal behavior from abnormal behavior, this can become quite a complex task. To this end, thousands of computations are needed to be accurately performed in milliseconds. Both supervised and unsupervised methods make important contributions in FSFD and must be woven into next-generation, comprehensive fraud strategies. In the next sub-section, we discuss the methods in terms of their learning type and techniques types.

1) LEARNING TYPE

The most common machine learning technique across all disciplines is supervised models, which refers to models trained on a rich set of properly labeled financial statements. Each statement is tagged as either fraudulent or non-fraudulent. In order to learn patterns that best reflect legitimate behaviors, the models are trained by feeding extensive amounts of tagged statement details. When developing a supervised model, the number of relevant training data is directly correlated with the accuracy of the model. Generally, the purpose of developing an unsupervised model is to spot anomalous behaviors in situations where the number of tagged data is limited or missing. Therefore, a class of self-learning must be applied to surface patterns in the data which are invisible to analytics. Taking a combination of supervised and unsupervised techniques, previously unseen types of suspicious behaviors could be detected in a manner called semi-supervised learning methods [37].

Several studies employed multiple ML/DM approaches. Machine learning and data mining algorithms are used 141 times in the current literature. Figure 9 presents a hierarchy of the algorithms in the reviewed literature and demonstrates how many times each technique type is applied. It should be mentioned that some articles employed several ML/DM methods. The algorithms are primarily categorized based on the learning method into three main general classes: supervised, semi-supervised, and unsupervised. At the next level, we divided the ML/DM approaches into eight categories of techniques, five supervised techniques, one semi-supervised technique, and two unsupervised techniques (Figure 9).

Most FSFD methods used supervised machine learning methodologies. Notably, among 47 articles, 42 articles employed supervised methods, four studies used unsupervised approaches, and only one article employed semi-supervised algorithms.

2) TECHNIQUE TYPES

As we mentioned earlier, we classified the employed ML/DM methods of the reviewed literature into eight categories. The

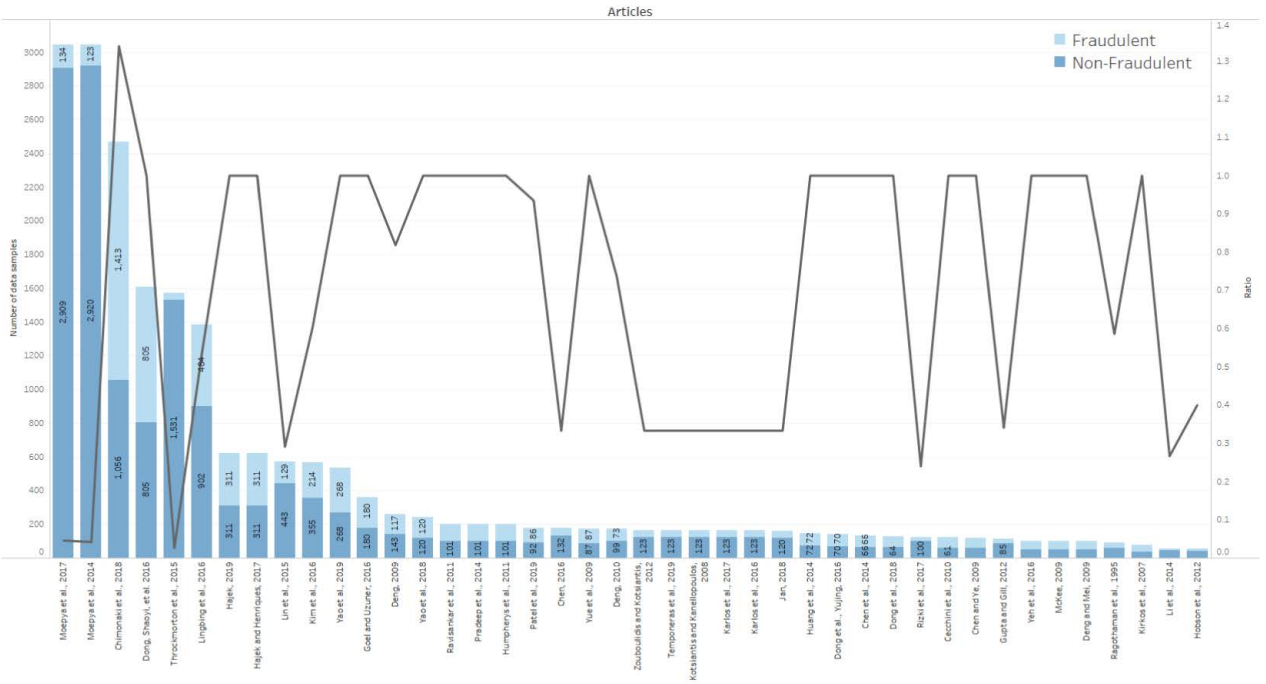
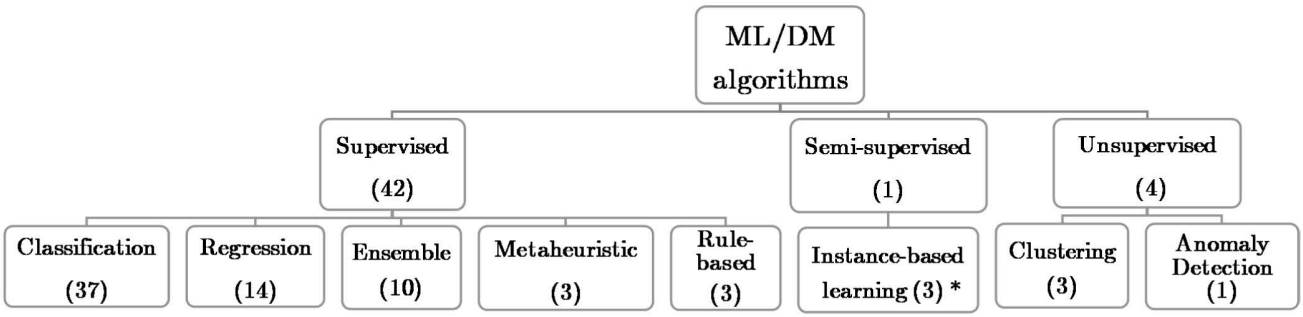


FIGURE 8. The dataset size and the number of fraudulent and non-fraudulent samples, as well as the dataset ratio (#Fraudulent#Non-Fraudulent) for each article.



* Instance based learning algorithms are implemented as either supervised (2) or semi-supervised (1) techniques.

FIGURE 9. ML/DM algorithms.

following, we explain each category, as well as the ML/DM methods in these categories.

a: CLASSIFICATION

Classification methods are the most commonly applied approach used to identify fraudulent financial statements. Different classification techniques are applied in the literature for fraud detection. Support vector machines (SVM) is a supervised classification approach seeking a maximum margin hyper plane that categorizes input samples into two classes. In particular, it classifies new data points based on a labeled training dataset for each category. SVM is the most generally used classification technique in the reviewed articles (31 articles). The kernel of SVM refers to a set of mathematical functions that take the input data and map it to high-dimensional space. Therefore, SVM is able to perform both linear and non-linear (using kernel function) classification tasks. Sequential Minimal Optimization (SMO), Radial Basis Function (RBF), linear, Gaussian, polynomial, and class-weighted kernel are examples of kernel function exploited in the studies.

In the SVM category, the most popular kernel was SMO, which was used in prior fraud detection literature to train an SVM classifier [22], [71], and it was also employed six times among our reviewed articles. One of the earlier studies in 2011 showed that logistic regression and SMO performed well compared to bagging, stacking, artificial neural network, and C4.5. Their experiments revealed that logistic regression and SVM outperformed or performed as well as a relatively representative set of classification algorithms [55]. Furthermore, relatively new studies examined the performance of the SMO kernel on their input dataset as well. Huang *et al.* [31]

TABLE 6. SVM kernels.

SVM Kernel	Frequency
Not mentioned	11
SMO	6
Linear	6
RBF	4
Gaussian	1
Polynomial kernel	1
Class-weighted	1
MI SVM	1
Sum	31

TABLE 7. Decision Tree extensions.

Decision Tree	Frequency
C4.5	9
CART	7
CHAID	2
QUEST	1
JRip	1
Logistic Model Trees	1
Tree Induction	1
Not Mentioned	1
Sum	24

created an ensemble Adaboost classifier by attractively running a given SMO classifier on various distributions across the training data. They compared the results with many other classifiers, including random forest, decision tree, neural networks, and Naive Bayes. They observed that the rate of classifying fraudulent firms as fraudulent (true positive rate) is better for ensemble methods. RBF kernels were used four times among the studies. Chen [15] concluded from the experimental results that SVM model with RBF kernel is effective in detection of fraudulent statements. Table 6 presents the distribution of the SVM kernels over the articles. Eleven articles did not mention the employed kernel of the SVM in their research.

The family of the decision tree (DT) algorithms has attracted significant attention from researchers (24 articles). There are different ways to create a decision tree for classification. Table 7 shows the distribution of these methods over the reviewed articles. The traditional decision tree algorithms have been around for decades. Furthermore, random forest is a modern variation of decision trees and is among the most powerful techniques. Classification and Regression Trees (CART) refer to decision tree algorithms that might be applied for regression or classification predictive modeling problems. CART is a binary decision tree method dealing with classifying non-parametric or continuous data. The conditions are subdivided based upon attributes of the data, the quantity and, the Gini index. The data is divided into two branches repeatedly to specify the next dividing conditions. This division of data is continued to build a tree structure. This procedure is terminated when data is not splittable anymore. As illustrated from Table 7, CART is the most commonly implemented decision tree approach. Hajek and Henriques [29] stated that CART succeeded to classify 95% cases. In particular, among 29 fraudulent cases, 25 firms were correctly classified as fraudulent. Consequently, even early studies (in 2012) examined CART for the FSFD problem and performed relatively well for this problem [29]. Furthermore, a most recent article used six different classifiers that one of them was the CART algorithm [77]. C4.5 algorithm operates as a decision tree classifier among ML/DM techniques. It takes a decision relying on particular instances of data (univariate or multivariate predictors). C4.5 decision tree algorithm measures each attribute's information gain and divides the data using the attribute with the greatest information gain. Six articles applied C4.5 tree as their fraud detection classifier [29], [34], [37], [47], [55], [81]. There was an article published in 2011 reported that C4.5 managed to achieve the highest classification accuracy [34]. The other frequent decision tree algorithms are Chi-Squared Automatic Interaction Detector (CHAID), Quick Unbiased and Efficient Statistical Tree (QUEST), JRip, Logistic Model Trees, and Tree Induction approaches (Table 7).

Another family of classification methods is Bayesian methods. We explored 19 articles that used one of two types of Bayesian methods, namely, Naive Bayes (13 papers), and Bayesian Belief Networks (6 papers). Naive Bayes is a kind of machine learning techniques which is based on the Bayes theorem and used for classification purposes. It predicts membership probabilities per class. In particular, this classifier predicts a given data point label based on the probability that it belongs to a specific class. The most likely class refers to the class with the highest probability. The first presumption of a Naive Bayes classifier is that the mutual dependencies and the correlations between the data points are disregarded conveniently. A particular attribute's value is independent of any other attribute's value. Bayesian methods are good in supervised settings. In many practical applications, the input is the training data, and the output is the required parameters for the classifier in the training phase. Naive Bayes models utilize Maximum Likelihood Estimation (MLE), optimization of loss criterion, and Bayesian Estimation (Maximum a posteriori) to select parameters for the classifier from the training data. This indicates that it is not necessary to rely on the Bayesian theorem to work with the naive Bayes method.

Bayesian Belief Networks are statistical probabilistic graphical models that illustrate a set of variables and their conditional correlations through a directed acyclic graph (DAG). In the direction of predicting the probability that which of several potential recognized causes was the contributing component, Bayesian networks are an excellent choice. An earlier study in 2007 investigated the advantage of Bayesian Belief Networks over Decision Trees and Neural Networks in identifying fraudulent financial statements in which the Bayesian Belief Network method obtained the highest performance (90.3%) on the validation set [40].

The input vector was formed from ratios that originated from financial statements. Various studies were used Bayesian networks as well to predict the fraudulent financial

TABLE 8. Artificial neural networks.

ANN Approach	Frequency
Multi-layer feedforward neural network (MLFF)	11
Not Mentioned	4
Probabilistic neural network (PNN)	1
Self-OrganizingMap (SOM)	1
Group method of data handling (GMDH)	1
Sum	18

statements [16], [39]. However, the most recent research was done in 2017 [31].

Artificial Neural Networks (ANN) are another category of classification algorithms. ANNs are information-processing models inspired by biological nervous systems behavior. Neural Networks are very powerful when a massive volume of data is available. Although the learning phase of ANNs could be supervised, semi-supervised, or unsupervised, there was only one extension of ANNs across literature that employed an unsupervised learning known as Self Organizing Map (SOM) [33]. The other types of ANNs were implemented using supervised approaches. ANNs could be categorized into various classes in terms of their learning strategy. Multi-layer feed-forward neural network (MLFF), probabilistic neural network (PNN), SOM are the most popular ANNs in our study, in terms of usage frequency respectively. Furthermore, we placed the group method of data handling (GMDH) under neural network models since it is somewhat inspired by research in perceptrons and learning filters. It is a supervised approach that automatically self-organizes (like SOM) the models and determines the best number of nodes and network structure. Table 8 illustrates the frequency of various ANNs across the reviewed studies. Eleven articles were attempted to detect FFS by implementing an MLFF neural network [21], [31], [31], [38], [47], [48], [55], [69], [77], [81]. However, there was only one article which employed PNN [59] and SOM [17] neural network approaches. Unfortunately, three articles did not mention the learning way of their neural network approach [16], [61], [76].

PNNs was proposed for the first time by Specht are based on the Bayes theorem [68]. They are feedforward neural networks that are suitable for classification problems and data mapping. In particular, PNNs seek to minimize the probability of misclassification. Although these types of neural networks can work with a smaller size of training data, they require large amounts of memory by increasing the size of training data. [59] stated that PNN was outperformed the other machine learning approaches. They examined this approach in two scenarios with and without feature selection and in both cases it obtained a better accuracy rather than other techniques.

SOMs, also known as Kohonen networks, are the only neural network that is trained using unsupervised neural networks employed in the literature. Dong [17] developed a model combining a clustering method and SOM to detect fraudulent financial reports.

Karlos *et al.* [35] introduced a kind of inductive learning algorithms called GMDH, which attempts to automatically discover interdependence in data to increase the accuracy of existing algorithms through choosing the optimum structure for the network. It is a self-organizing method that builds and evaluates polynomial models by applying some external conditions on distinct sections of data. This model leads to the minimum error between the predicted value and expected output. The majority of GMDH networks apply regression analysis to solve the problem. There was only one article across the studies that employed GMDH as the classification approach [59].

Backpropagation is the primary and most well-known and commonly used algorithm for training a neural network model in a supervised learning scheme. It is just a method of propagating the total error backward in the neural network. It uses a gradient descent approach to calculate the error in terms of the weights of neurons. Given an error function, the purpose is to update the weights to minimize the error by giving the nodes with lower error rates higher weights and vice versa.

K-nearest neighbors (KNN) algorithm is a handy, straightforward supervised machine learning method that attempts to address both regression and classification problems. KNN method decides the class label based on a very restricted number of nearest samples. In particular, it determines the label for the test instances according to the k adjacent neighbors' labels of that instance. Consequently, this method is more appropriate for the overlapping or crossover sample sets of samples, since it largely relies on the surrounding limited samples, instead of the discriminant domain method to decide the category [77]. Several articles used KNN as classifiers in their works with a various number of nearest neighbors (k). For example, Kim *et al.* [37] supposed k to be three, and Zouboulidis and Kotsiantis [81] considered only one nearest neighbor in the algorithm. Cressey [12] reported that the KNN had higher average accuracy (89,11%), relative to other classifiers. The KNN classifier is non-parametric and hence does not consider a model. KNN classifier is not a model-based or distribution based classifier. While it performed well on several datasets, the performance of this machine learning method likely to be undermined by an unbalanced dataset.

The latest investigated classification method is Distance Weighted Discrimination (DWD) model [47]. DWD model has an excellent performance in high-dimension low-sample-size (HDLSS) settings and was applied to detect fraudulent financial statements. Researchers combined textual and financial features to enhance financial statement fraud detection by employing the DWD model. It achieved a reasonable performance in high-dimension low-sample-size (HDLSS) settings to detect fraudulent financial statements. Likewise, they proposed a DWD method based on a Genetic algorithm to perform parameter optimization and feature selection, which achieved reasonably good classification performance using a limited set of features [47].

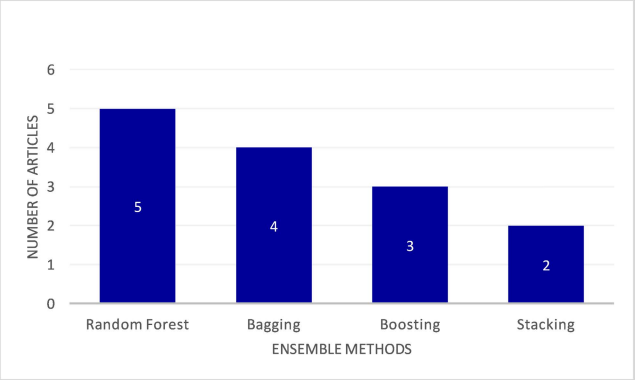


FIGURE 10. Distribution of Ensemble methods.

b: REGRESSION

The studies in this SLR utilized Logistic Regression (LR) method frequently in their researches. In particular, 16 studies employed logistic regression [10], [11], [31], [32], [34], [39], [48], [50], [54], [55], [59], [69], [71], [76], [77], [80]. Logistic regression models are applicable in multi-class and binary classifications. Another reason for the popularity of the logistic regression models in related accounting literature is that they are statistically interpretable. The basic logistic regression is a statistical method that applies a logistic function to model a binary dependent variable; however, several other intricate extensions are available. Kirkos and Manolopoulos [39] employed a multinomial logistic regression. They compared the results from multinomial logistics with SVM and BayesNet classifiers and found that the logistic method performed better. They stated that the reason behind selecting the logistic regression model is that it is used widely among the previous studies. Furthermore, they mentioned that according to a review article by Abbasi *et al.* [1], the most popular ML model used in developing financial misstatement detection is the logistic regression model. Based on that review article, two-thirds of the studies employed logistic regression models for FSFD.

c: ENSEMBLE METHODS

Ensemble methods are meta algorithms that aggregate multiple intelligent models into one predictive model. The general purpose of ensemble algorithms is to convert individual weak learners to stronger learners. Every ensemble method follows different goals; for instance, bagging tries to decrease variance; boosting manages to decrease bias, and stacking wants to improve predictions. Moreover, Random forest (RF) or random decision forest is another ensemble learning technique for regression, classification, and other problems. We identified these four types (bagging, boosting, stacking, and random forest) of ensemble models in the literature (14 articles). Figure 10 presents the distribution of ensemble models per category. Some articles employed several (two or three) ensemble approaches in their studies [31], [55], [81].

RF model is the most common ensemble approach among the examined literature [31], [38], [52], [54], [76]. It generally tends to achieve higher performance against an individual decision tree by generating a bunch of decision trees over the training and outputting the class based on the task. In particular, it outputs the mode of the classes of the single trees in classification tasks and the median prediction for regression problems. While RF is an approach proposed in 1995 -like many other ML/DM methods- it is still practical. A study performed recently in 2018 indicated that RF outperformed SVM, DT, ANN, and logistic regression models [76].

Bagging, which is known as Bootstrap Aggregating, generates numerous samples from the training set with replacement. Decision trees are susceptible to small changes in the training set. Accordingly, by randomly splitting the training dataset into two subsets and fitting a decision tree to each of them, the outcomes can be fairly different. In the interest of decreasing the variance of a decision tree classifier, the bagging method can be applied. This method aims to build several subsets of data from the training sample picked randomly with replacement. Four articles used bagging methods across the reviewed articles [31], [37], [55], [81]. The base learners are different in various bagging approaches. Thereby, Huang *et al.* [31] adopted Reduced Error Pruning Trees (REPTs) that create decision trees using information gain and reducing the variance for base learners of bagging.

In bagging, each model is trained independently. However, the goal of boosting is to train weak learners sequentially by adjusting the distribution of the training dataset according to the predecessor's accuracy adaptively. There are various types of boosting, including AdaBoost, which is the most common approach [24]. AdaboostMI is a multi-instance Adaboost that was employed by Huang *et al.* [31]. In this study, AdaboostMI repeatedly executes different distributions of SVM (as the weak classifier) throughout the training dataset and then merged the classifiers into an individual hybrid classifier [31].

Stacking is another ensemble learning technique that combines several classifications or regression models. There are notable differences between stacking and the former ensemble algorithm. First, unlike bagging, the models are generally different, and instead of working on samples of the training data, they operate on the same dataset. Second, in contrast to boosting, an individual model is used to properly combine the predictions from the contributing models, rather than a sequence of models that improve the predictions of preceding models. Two articles applied this approach in order to learn a classifier for FSFD [50], [55], namely, the primary purpose of Moepya *et al.* [50] was to determine whether the stacking form of meta-learning performed better in the FSFD problem compared to the individual methods.

d: METAHEURISTIC ALGORITHMS

In addition to the previous category of ML/DM methods, we figured out three metaheuristic algorithms in the articles [29], [57], [59]. Metaheuristic approaches are applicable in both supervised and unsupervised learning. However, since the articles in our study employed only the supervised version, we categorized it under supervised approaches.

Genetic Programming (GP) belongs to the family of evolutionary algorithms which extends the genetic algorithms (GA) to support the exploration of the space of computer programs [45]. Like GA, GP randomly generates an initial population of solutions for the problem and manipulates them, subsequently using genetic operations like crossover, mutation, and reproduction. This novel population is known as a generation. Serrano-Cinca [59] and Hajek and Henriques [29] applied GP alongside other data mining methods to identify companies that committed financial statement fraud. Serrano-Cinca [59] observed that GP outperformed the other classifiers.

e: RULE-BASED APPROACHES

Rule-based methods are another popular class of techniques in machine learning and data mining. They share the goal of finding regularities in data that can be expressed in the form of an IF-THEN rule. Particularly, rule-based classification can be applied to refer to any classification scheme that utilizes IF-THEN rules for the prediction task. There were three studies in the articles that used rule-based methods to predict the fraudulent financial statements [30], [57], [58]. Ravisankar *et al.* [57] proposed rule-based classifiers based on firefly (FF) and threshold accepting (TA) algorithms to distinguish the fraudulent and non-fraudulent firms according to their financial statements. Their approach is a combination of rule-based and evolutionary algorithms. The firefly algorithm is a metaheuristic inspired by the flashing behavior of fireflies and introduced by Yue *et al.* [75].

Reurink [58] developed a prototype expert system, which evaluates flaws in financial statements to discover potential fraud. This model supports auditors during the development of substantive tests when violations in the financial statements are presumable. Their work was the earliest work that was included in our study. However, it does not mean that the rule-based approaches are deprecated today. Hobson *et al.* [30] incorporated a genetic feature selection and rule extraction to propose a fuzzy rule-based fraud detection model. Eliminating irrelevant attributes through feature selection and performing a Fuzzy Unordered Rule Induction Algorithm (FURIA) enabled this model to achieve a reasonable performance.

f: INSTANCE-BASED LEARNING

This term refers to a family of learning algorithms that substitute performing explicit generalization by comparing unseen samples to samples seen in training, which have been stored in memory. Three articles applied this type of method to discover fraudulent statements as either supervised [34], [44] or semi-supervised [37] techniques. Kumar and Ravi [44] examined the efficiency of multi-instance learning techniques for FSFD. For this purpose, several experiments have been performed utilizing representative learning algorithms, which were trained using a dataset of fraudulent and non-fraudulent Greek companies.

Locally Weighted Learning (LWL) is a function approximation technique and another investigated method in the family of instance-based learning. In these methods, the aim is to reveal the relationship between input and output. Accordingly, an approximated local model around the current point was used to perform a prediction task. Karlos *et al.* [34] exploited LWL methodology to investigate FSFD by performing a natural language processing method on the MD&A section textual data. The main idea was to identify some textual cues from the MD&A section, which are the most informative indicators of the author's intent.

The majority of proposed ML/DM approaches for FSFD aimed to detect the fraudulent statements regardless of the shortage of the available labeled data (fraudulent labeled statements) and the imbalanced datasets. Accordingly, most of the introduced approaches employed supervised learning. Kim *et al.* [37] compared the performance of seven semi-supervised schemes that employ instance learning, decision trees, and SVM with the corresponding supervised algorithms for FSFD using data from Greek firms. The advantage of using semi-supervised approaches for FSFD is that they require just a limited number of labeled samples for obtaining robust learning behaviors to discover appropriate patterns from a greater quantity of unlabeled samples. They concluded that half of the evaluated semi-supervised learning algorithms outperformed the supervised approaches.

g: CLUSTERING

Clustering is the technique of organizing identical instances into the same groups. This unsupervised machine learning technique aims to look for similarities in the instances and to group similar instances. Clustering methods were implemented considerably less than classification approaches in the reviewed studies. There are only three articles utilizing clustering approaches to detect fraudulent financial statements [17], [25], [33]. Gozman and Currie [25] employed a text mining hierarchical clustering on MD&A textual content to create an FSFD model. While preserving the maximum amount of information, a basic scheme for dimension reduction is creating a singular value decomposition vector (SVD) [3]. Gozman and Currie [25] created SVDs to reduce the text dimension. Then the documents are clustered, and the clustering results are reviewed and evaluated. Afterward, the SVDs and clustering algorithm are modified if necessary.

A growing hierarchical self-organizing map (GHSOM) is an extension of SOM. It is an unsupervised neural network for clustering, which was applied for the financial problems [63], [66]. Ivakhnenko [33] extended its property of topology to systematically identify the spatial relationships of dichotomous cases. The dual GHSOM approach is introduced to evaluate the non-fraud-central spatial hypothesis [33]. They proposed a GHSOM method to detect the topological patterns of fraudulent financial statements. In particular, the proposed approach comprised two main mechanisms. The first mechanism is the classification mechanism. Two

TABLE 9. ML/DM technique types and descriptions.

Technique Type	Technique Name	Frequency
Classification	SVM	31
	Decision Tree	24
	Bayesian Classifiers	19
	Artificial Neural Network	15
	KNN	8
	Group method of data handling (GMDH)	1
	Generalized Likelihood RatioTest (GLRT)	1
Regression	Logistic Regression	14
Ensemble Methods	Random Forest	5
	Bagging	4
	Boosting	3
	Stacking	2
Metaheuristic Algorithms	Genetic programming (GP)	1
	Hybridized Firefly-Threshold Accepting (FFTA)	1
	Evolutionary Computation Based Rule Miners	1
Rule based	Rule Induction	1
	Evolutionary Computation Based Rule Miners	1
	Fuzzy rule-based system	1
Instance based learning	Locally Weighted Learning (LWL)	1
	Multi-instance learning	1
	Instance learning	1
Clustering	K-means	1
	SVD Clustering	1
	Self Organizing Map	1
Anomaly Detection	Multivariate anomaly detection Mahalanobis distance	1

dual GHSOMs were trained with identical parameters for each of the fraudulent and non-fraudulent samples during this phase. In this way, two trees were generated for each class label called fraud tree (FT) and non-fraud tree (NFT). The second mechanism was the feature-extraction, wherein the relationship between the subgroups was evaluated. This stage exploited the benefits of GHSOMs and unsupervised learning. Taking advantage of the duality of two GHSOMs to detect topological patterns in the equivalent leaf nodes of each tree discriminates this study.

Dong *et al.* [17] combined SOM and K-means clustering to design a clustering model. To overcome one of the disadvantages of SOM, which is the uncertain clustering borders of nodes, this model applied k-means clustering to the outcomes of SOM. Additionally, since the inputting order of samples and the initial values of the nodes affects the clustering results, this study employed the Silhouette cluster validity measure to evaluate the robustness of various clustering results.

h: ANOMALY DETECTION

There is another unsupervised approach among the literature that we categorized it as an anomaly detection approach. Mahalanobis distance is one of the most straightforward methods to filter abnormal data points [70]. Yao *et al.* [72] employed a data mining approach in which the Mahalanobis distance is used to calculate the closeness of each data point from the centroid of the distribution to declare the amplitude of the anomaly. The suggested model was able to effectively arrange financial statements in the sense of trustworthiness. Implementing the model on all observations indicated that almost a quarter of them were highly anomalous and questionable, and the majority of Vietnamese listed companies' financial statements are credible [72].

Table 9 summarizes the ML/DM categories, the algorithms in each group, and the usage frequency in the literature.

C. PERFORMANCE

In this section, the performance results reported in the reviewed articles are presented and compared. In addition to the accuracy, several other performance measures have been used in the literature, including sensitivity and specificity and AUC. However, since the most prevalent reported performance criterion was accuracy, we only included the articles in which the accuracy is reported in our analysis.

To compare the ML/DM techniques in terms of performance, we divided them into three categories based on the model types: ensemble, hybrid, and single models.

Both hybrid and ensemble methods take advantage of the information fusion concept but in a slightly different way. In the case of hybrid approaches, completely different heterogeneous machine learning models are combined, while ensemble classifiers, in turn, combine several homogeneous, weak classifiers typically at the individual output level, using various merging methods. Ensemble models are discussed in Section III-B2c.

Furthermore, a dataset ratio is calculated for each dataset. The dataset ratio is the number of fraudulent samples divided

by the number of non-fraudulent samples, and it is used as a parameter in our analysis.

Figure 11 presents the maximum of accuracy (%) for each technique across ensemble, hybrid, and single models. The average ratio is presented by colors and labels for each model. In terms of the maximum accuracy for each model, the single model overall had a better performance. The best performance (93.7%) was obtained using a deep, dense multilayer perceptron by [69]. This study aimed to examine a deep neural network architecture in predicting FSFD in a binary classification task using Greek data. Their dataset comprised 164 records of data, including 41 fraudulent and 123 non-fraudulent records. The next well-performed model is a logistic regression (93.33%) employed by [11] making use of 44 fraudulent and 44 non-fraudulent financial statements.

From the ensemble category, a boosting method (with an accuracy of 91.8%) by [44] performed better than random forest (87.5%), bagging (87.9%), and stacking (83%), respectively. Among hybrid approaches, a combination of self-organizing map and k-means clustering algorithms (91%) overperformed the others [17]. Furthermore, combining neural networks with various decision tree approaches, especially CART (90.83%) and CHAID (90.37%) models, led to promising results.

In terms of the dataset ratio, from Figure 11, it can be concluded that generally, for models with higher performance, this ratio is higher (the color is roughly darker for the upper part of the upper histograms in Figure 11, and it is brighter for the lower ones). The higher the ratio, the closer number of fraudulent and non-fraudulent samples. This is not surprising as most of the employed techniques are binary classification models and perform better with a balanced dataset.

RQ2: What are the gaps, trends of research, and future research directions in this area?

D. TREND OF RESEARCH OVER THE YEARS

To address this question, investigate the trend of the articles, and compare the datasets and techniques developed over the years, we divided the articles into two categories; the articles published between 1995 and the end of 2014 (21 articles), and the articles between 2015 and 2019 (26 articles). Our goal was to compare the results over these two periods, to determine which dataset types, data categories, and techniques have gained attention recently. This also helped us to find the gaps in this area of study to lead future researchers to consider these gaps in future works.

First, we looked at the distribution of the structured and unstructured data used in the studies during each period of 1995-2014 and 2015-2019. Figure 12 demonstrates this comparison. There was no significant difference between the two groups of prior and former studies. In other words, the unstructured data have gained less attention in the studies conducted in both periods.

Textual contents and non-financial variables from the financial statements became popular during the second period (Figure 5). Nevertheless, financial variables were used widely in both periods more frequently than the others. It is worth mentioning that some of the studies made use of several indicators in their research. Accordingly, in the illustrated diagrams, some articles might be considered more than once. It could be figured out from Figure 5 that vocal speech data did not gain popularity. We can conclude that considering unstructured datasets might open new roads to future researchers to contribute to this area.

The diversity of the countries that attempted to use FSFD techniques is increasing over time. As stated in Figure 6, six new countries made an effort to use ML/DM techniques for FFS detection. These countries were Indonesia, Taipei, Lebanon, Vietnam, India, and Korea. United States, Greece, and China are the pioneer countries that started using ML/DM approach for the FFSD task.

In terms of ML/DM techniques, supervised algorithms have always been of interest 13. However, it is interesting that among four unsupervised learning algorithms in the literature, three of them were used before 2015 [17], [25], [33]. It seems that the popularity of these algorithms decreased-however, the only semi-supervised approach employed in a study in 2016 [37]. Figure 13 shows the yearly distribution of the articles in terms of learning type.

In addition, Figure 14 reveals more details regarding the eight aforementioned technique types within the two time periods (2014 and before, or 2015 and after). As illustrated, classification and regression techniques were always of interest. An increasing trend in FSFD was the use of ensemble methods that utilize the advantages of multiple algorithms to classify samples. Among ten articles, seven articles were published in years from 2015 to 2019 [31], [37], [38], [49], [52], [54], [76]. The first article that used an ensemble method was published in 2009 [7]. In total, the other categories of algorithms, such as meta-heuristic, clustering, instance-based learning, rule-based methods, and anomaly detection, were used far less than classification, regression, and ensemble methods. There is only one article that employed an unsupervised anomaly detection approach that was published in 2019. In the future, the number of studies attempt to perform anomaly detection in this area of study might be increased. The other interesting insight from Figure 12 is that while regression and ensemble methods were used evenly (14 times among literature), ensemble methods gathered more attention from 2015 to 2019.

E. GAP ANALYSIS AND FUTURE DIRECTION

We synthesized the articles to determine their limitations and explore the gaps in this area of study and future work opportunities. In the following, we explain the limitations (gaps) in the reviewed literature and elaborate on the opportunities for future studies.

1) ML/DM TECHNIQUES

Categorizing the ML/DM algorithms applied in FSFD is an effective way to determine the appropriate methods for this

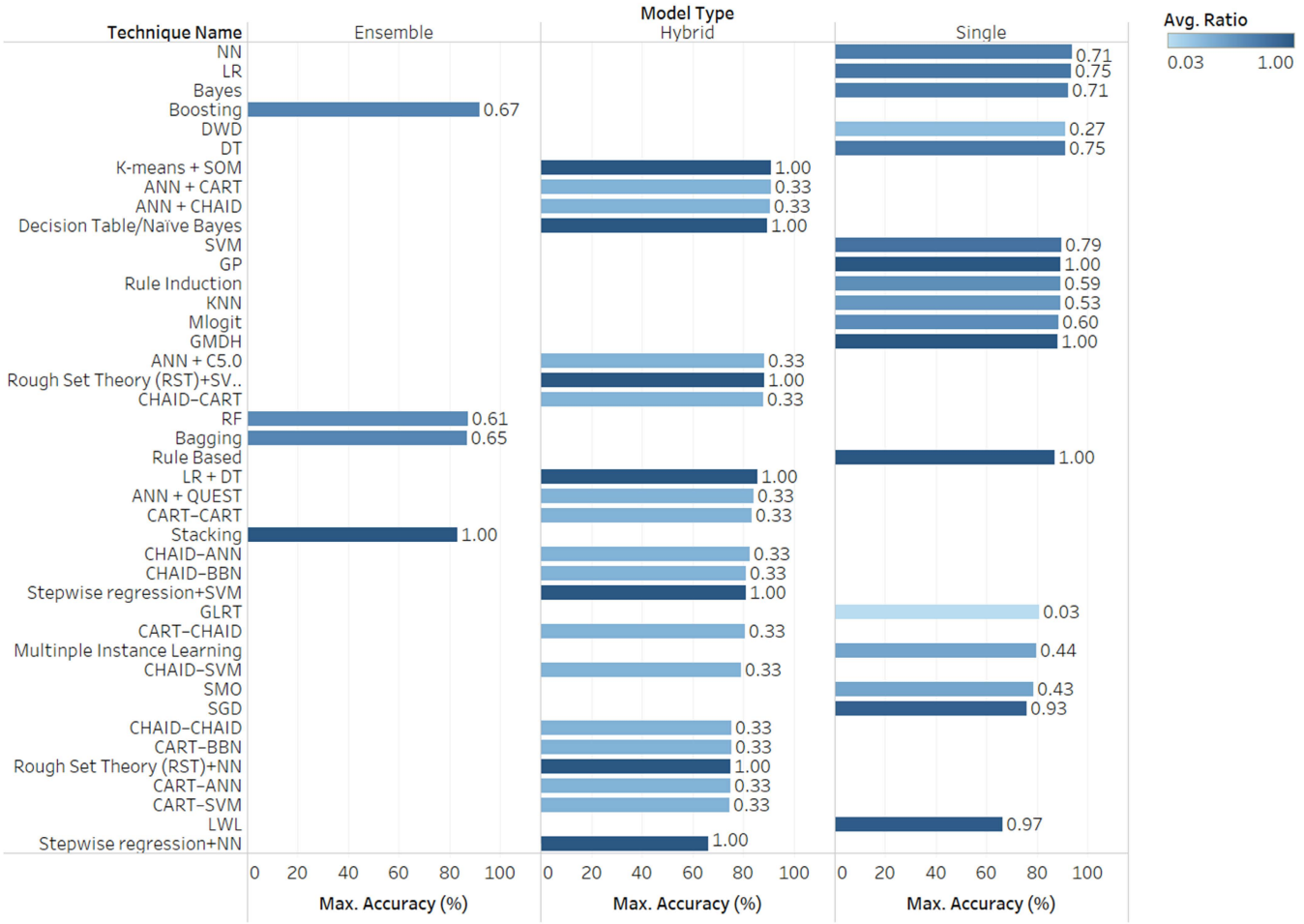


FIGURE 11. Maximum of Accuracy (%) for each Technique Name broken down by Model Type. Dataset Ratio is the number of fraudulent divided by the number of non-fraudulent data samples. Colors and marks show average Dataset Ratio.

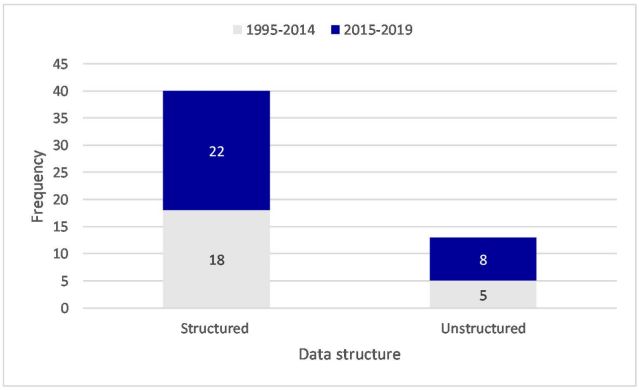


FIGURE 12. The data type used in the studies during the two time periods.

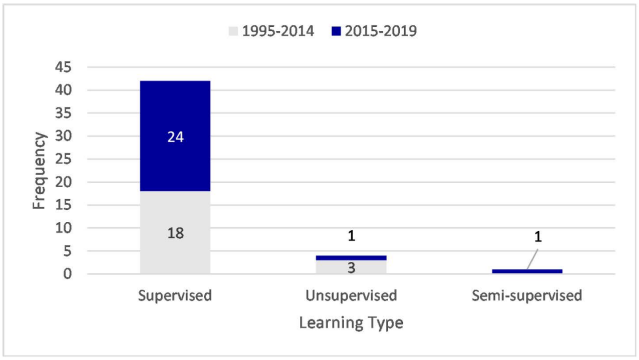


FIGURE 13. Yearly distribution of the articles, in terms of learning type.

research domain. It is also beneficial to decide why specific approaches were selected. Furthermore, we can identify gaps in the research by investigating the methods that have not been received much attention. We observed that ensemble methods were considered more frequently in the past five years. Moreover, among unsupervised approaches, clustering methods have not been employed recently. Consequently, anomaly detection approaches, as an unsupervised approach, would be an interesting area for future research. Additionally, heuristics and meta-heuristics algorithms combined with bio-inspired algorithms such as artificial immune systems (AIS), and genetic algorithms (GA) can be studied further for fraud detection in financial statements.

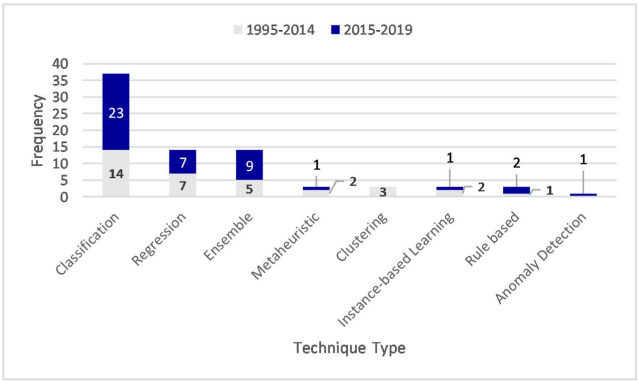


FIGURE 14. Yearly distribution of the articles, in terms of ML/DM technique type.

2) ENRICHING FEATURE VECTORS

The majority of the articles in which only financial ratios are considered as the input data stated that the performance of the detection method could be improved by enhancing the input vector. Further qualitative information, such as previous auditors' reports or the combination of the administrative board, could be beneficial for this purpose. There is an opportunity for future works to integrate multiple sources of data such as financial social media like SeekingAlpha, both textual (MD&A section) and numeric data from financial statements, as well as the earning calls transcripts to create a more informative feature vector. Additionally, if textual data is going to be included in building the models, then exploring emerging techniques to convert textual contents into vectors of features, such as Word2Vec, Doc2Vec, and BERT algorithms, merit further investigation.

3) DATA SIZE

Some articles mentioned that the size of the dataset is the shortcoming of their work. For example, Jan [36] stated that the size and scope of the financial market in Taiwan are smaller than the European Union, China, the U.S., Japan, or the U.K. Furthermore, the number of listed companies in Taiwan is relatively small in scale. Consequently, the data size is one of the major issues in other countries. If there is a general ML/DM FSFD model that could detect fraudulent financial statements across countries, limitations on the size of the data could be resolved.

4) IMBALANCED DATASET

The number of fraudulent financial statements is far fewer than the non-fraudulent ones. For instance, Karlos *et al.* [36] stated that Taiwan's competent authority rigorously administrates the financial market and listed companies. Accordingly, the majority of the listed companies were non-fraudulent in their report [36]. Therefore, one of the issues with the datasets is that they are highly unbalanced. Some articles made use of oversampling methods to balance the datasets [20], [21]. The others tried to propose a model that could work fine with highly imbalanced data. Lingbing *et al.* [47] and Pradeep *et al.* [55] used imbalanced data and left balancing dataset using the oversampling technique as future work. Also, McKee [48] indicated that they avoid oversampling since it might result in choice-based sample biases. There were no articles that used under-sampling techniques among the reviewed literature. In addition, the only applied oversampling technique was the SMOTE method (Synthetic Minority Oversampling Technique). With this in mind, future studies could take into account employing other oversampling methods, as well as under-sampling methods.

5) SOCIAL MEDIA DATA ANALYSIS

Including the comments and opinions from financial social media opens new avenues for future research in this area. Discussions and comments on financial social media platforms disclose informative data about the financial violations and ethical behavior of a company. Two recent research in 2018 and 2019 made use of social media data to perform detection tasks. However, future research can benefit from the advantage of sentiment and semantic analysis approaches using emerging natural language processing techniques on financial statement reports.

6) UNSTRUCTURED DATA

Recently, researchers focused on various kinds of unstructured data such as textual and vocal inputs. However, exploring the unstructured data in the financial fraud detection domain deserves more attention to attain remarkable results. Future studies are expected to investigate the text sources like the MD&A section from the financial statements, as well as earnings calls transcripts and CEO conferences. Additionally, another avenue for future research would be to use emerging text mining techniques, and in particular, word embedding techniques (Word2Vec, Doc2Vec, BERT) to transform the financial texts into vectors of features which will then be used to build machine learning models.

IV. THREATS TO VALIDITY AND LIMITATION

We identified ML/DM techniques, the data sources, and the research trends in this systematic literature review. We designed our protocol to promote internal and external validity as much as possible while answering the research questions. However, there are still some limitations and validity threats that we mention in the following.

A. SEARCH BIAS

Although we have consulted major digital libraries to explore the articles, there might be additional digital libraries with relevant studies that might not be considered. To mitigate this limitation, we used the snowballing technique to include related articles that were not included during automatic searching. Additionally, we examined the search terms and keywords against a well-known list of research studies. Nevertheless, there is a possibility that we have missed some synonyms in searching for the keywords. To mitigate this

TABLE 10. The list of all queries explored in five digital libraries.

Digital Library	Query
Scopus	TITLE-ABS-KEY (("fraud*") AND ("financial *statement*") AND ("machine learning" OR "data mining" OR "artificial intelligence"))
ACM Digital Library	"query": { Title:((fraud*) AND (financial *statement*) AND (machine learning OR data mining OR artificial intelligence))) AND Abstract:((fraud*) AND (financial *statement*) AND (machine learning OR data mining OR artificial intelligence))) AND Keyword:((fraud*) AND (financial *statement*) AND (machine learning OR data mining OR artificial intelligence))) }
Science Direct	("fraud*") AND ("financial *statement*") AND ("machine learning" OR "data mining" OR "artificial intelligence")
IEEE Xplore	(("Document Title":fraud* OR "Abstract":fraud* OR "Index Terms":fraud) AND ("Document Title":financial statement* OR "Abstract":financial statement* OR "Index Terms":financial statement* OR "Document Title":financial restatement) AND ("Document Title":machine learning OR "Document Title":data mining OR "Document Title":artificial intelligence OR "Abstract":machine learning OR "Abstract":data mining OR "Abstract":artificial intelligence OR "Index Terms":machine learning OR "Index Terms":data mining OR "Index Terms":artificial intelligence))
Web of Science	TS=((fraud*) AND (financial *statement*) AND (machine learning OR data mining OR artificial intelligence))

issue, the SLR protocol is revised by a peer student and two university professors to ensure no essential terms are missed.

B. LANGUAGE BIAS

We only searched for the articles written in English. This causes language bias since some related articles might exist in this area of study in other languages. However, fortunately, all articles that we collected in this study were in English (i.e., the search engine did not identify any article published in other languages). As such, we have no language bias.

V. CONCLUSION

Financial statement fraud detection (FSFD) is a developing area in which it is advantageous to outrun the fraudsters. Besides, there are still aspects of intelligent FSFD that have not been investigated thoroughly. We performed a systematic literature review using Kitchenham methodology to analyze the FSFD problem in terms of machine learning/data mining approaches and datasets used in the studies. Out of 148 retrieved articles, 47 articles merited our review protocol criteria were synthesized in this research. We presented some of the key issues, gaps, and limitations of FSFD and suggested future research areas. Our study presented various ML/DM algorithms employed in the existing literature. Categorizing FSFD methods by the ML/DM fraud detection technique is an efficient method to identify the promising practices for this area of research. Accordingly, we categorized these techniques into eight classes, which enabled determining why some techniques favored others.

Furthermore, we identified gaps in the research by examining unexplored or less studied algorithms. Early researchers in FSFD focused on supervised classification and regression methods, such as SVM, neural networks, and logistic regression. The use of ensemble methods that take advantage of multiple algorithms to classify samples is a rising trend in FSFD. Interestingly, we find that unsupervised learning approaches, such as clustering, were only employed four times in the present literature. Clustering is beneficial for investigating latent relations and resemblances. Besides, since there are quite a small number of fraud cases to be identified, clustering could be effective. Future studies recommended paying more attention to unsupervised practices such as anomaly detection, which can uncover new insights.

In addition to exploring the ML/DM techniques, this study focuses on analyzing the datasets used for financial fraud detection in the existing literature. We investigated the data structure and classified them based on their data types into four classes: financial ratio, non-financial ratio, textual, or vocal data. Additionally, we determined the extraction sources of financial ratios, textual content, and other data types. Exploring the unstructured data in the financial fraud detection domain deserves more attention to attain remarkable results. We recommend future studies to examine the text sources like the MD&A section from the financial statements and earnings calls transcripts and CEO conferences. Additionally, another avenue for future research would be to use emerging text mining techniques and word embedding techniques (Word2Vec, Doc2Vec, BERT) to transform the financial texts into vectors of features, which will then be used to build machine learning models.

APPENDIX

SEARCH QUERY

Table 10 presents the employed search string for each digital library.

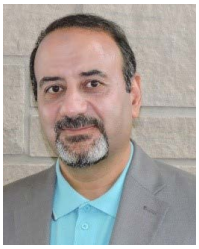
REFERENCES

[1] A. Abbasi, C. Albrecht, A. Vance, and J. Hansen, "MetaFraud: A meta-learning framework for detecting financial fraud," *Mis Quart.*, vol. 36, pp. 1293–1327, Dec. 2012.

[2] A. Albizri, D. Appelbaum, and N. Rizzotto, "Evaluation of financial statements fraud detection research: A multi-disciplinary analysis," *Int. J. Discl. Governance*, vol. 16, no. 4, pp. 206–241, Dec. 2019.
[3] R. Albright, "Taming text with the SVD. SAS institute white paper," SAS Inst., Cary, NC, USA, White Paper 10.1.1.395.4666, 2004.
[4] M. S. Beasley, "An empirical analysis of the relation between the board of director composition and financial statement fraud," *Accounting Rev.*, vol. 71, pp. 443–465, Oct. 1996.
[5] T. B. Bell and J. V. Carcello, "A decision aid for assessing the likelihood of fraudulent financial reporting," *Auditing A, J. Pract. Theory*, vol. 19, no. 1, pp. 169–184, Mar. 2000.
[6] M. D. Beneish and C. Nichols, "The predictable cost of earnings manipulation," Dept. Accounting, Kelley School Bus., Indiana Univ., Bloomington, IN, USA, Tech. Rep. 1006840, 2007.
[7] R. J. Bolton and D. J. Hand, "Statistical fraud detection: A review," *Stat. Sci.*, vol. 17, no. 3, pp. 235–249, Aug. 2002.
[8] M. Cecchini, H. Aytug, G. J. Koehler, and P. Pathak, "Making words work: Using financial text as a predictor of financial events," *Decis. Support Syst.*, vol. 50, no. 1, pp. 164–175, 2010.
[9] Q. Deng, "Detection of fraudulent financial statements based on naïve Bayes classifier," in *Proc. 5th Int. Conf. Comput. Sci. Educ.*, 2010, pp. 1032–1035.
[10] S. Chen, Y.-J.-J. Goo, and Z.-D. Shen, "A hybrid approach of stepwise regression, logistic regression, support vector machine, and decision tree for forecasting fraudulent financial statements," *Sci. World J.*, vol. 2014, pp. 1–9, Aug. 2014.
[11] X. Chen and R. Ye, "Identification model of logistic regression analysis on listed Firms' frauds in China," in *Proc. 2nd Int. Workshop Knowl. Discovery Data Mining*, Jan. 2009, pp. 385–388.
[12] C. Chimonaki, S. Papadakis, K. Vergos, and A. Shahgholian, "Identification of financial statement fraud in greece by using computational intelligence techniques," in *Proc. Int. Workshop Enterprise Appl., Markets Services Finance Ind.* Cham, Switzerland: Springer, 2018, pp. 39–51.
[13] D. R. Cressey, "Other people's money; a study of the social psychology of embezzlement," *Amer. J. Sociol.*, vol. 59, no. 6, May 1954, doi: 10.1086/221475.
[14] B. Dbouk and I. Zaarour, "Towards a machine learning approach for earnings manipulation detection," *Asian J. Bus. Accounting*, vol. 10, no. 2, pp. 215–251, 2017.
[15] Q. Deng, "Application of support vector machine in the detection of fraudulent financial statements," in *Proc. 4th Int. Conf. Comput. Sci. Educ.*, Jul. 2009, pp. 1056–1059.
[16] S. Chen, "Detection of fraudulent financial statements using the hybrid data mining approach," *SpringerPlus*, vol. 5, no. 1, p. 89, Dec. 2016.
[17] Q. Deng and G. Mei, "Combining self-organizing map and K-means clustering for detecting fraudulent financial statements," in *Proc. IEEE Int. Conf. Granular Comput.*, Aug. 2009, pp. 126–131.
[18] W. Dong, S. Liao, and L. Liang, "Financial statement fraud detection using text mining: A systemic functional linguistics theory perspective," in *Proc. Pacific Asia Conf. Inf. Syst. (PACIS)*. Chiayi City, Taiwan: Association For Information System, 2016.
[19] W. Dong, S. Liao, Y. Xu, and X. Feng, "Leading effect of social media for financial fraud disclosure: A text mining based analytics," in *Proc. AMCIS*, 2016.
[20] W. Dong, S. Liao, and Z. Zhang, "Leveraging financial social media data for corporate fraud detection," *J. Manage. Inf. Syst.*, vol. 35, no. 2, pp. 461–487, Apr. 2018.
[21] I. Dutta, S. Dutta, and B. Raahemi, "Detecting financial restatements using data mining techniques," *Expert Syst. Appl.*, vol. 90, pp. 374–393, Dec. 2017.
[22] K. Fanning, K. O. Cogger, and R. Srivastava, "Detection of management fraud: A neural network approach," *Intell. Syst. Accounting, Finance Manage.*, vol. 4, no. 2, pp. 113–126, Jun. 1995.
[23] E. H. Feroz, T. M. Kwon, V. S. Pastena, and K. Park, "The efficacy of red flags in predicting the SEC's targets: An artificial neural networks approach," *Int. J. Intell. Syst. Accounting, Finance Manage.*, vol. 9, no. 3, pp. 145–157, 2000.
[24] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *Proc. Int. Conf. Mach. Learn.*, vol. 96, 1996, pp. 148–156.
[25] F. H. Glancy and S. B. Yadav, "A computational model for financial reporting fraud detection," *Decision Support Syst.*, vol. 50, no. 3, pp. 595–601, 2011.
[26] S. Goel and O. Uzuner, "Do sentiments matter in fraud detection? Estimating semantic orientation of annual reports," *Intell. Syst. Accounting, Finance Manage.*, vol. 23, no. 3, pp. 215–239, Jul. 2016.
[27] D. Gozman and W. Currie, "The role of investment management systems in regulatory compliance: A post-financial crisis study of displacement mechanisms," *J. Inf. Technol.*, vol. 29, no. 1, pp. 44–58, Mar. 2014.
[28] G. L. Gray and R. S. Debreceny, "A taxonomy to guide research on the application of data mining to fraud detection in financial statement audits," *Int. J. Accounting Inf. Syst.*, vol. 15, no. 4, pp. 357–380, Dec. 2014.
[29] R. Gupta and N. Singh, "Prevention and detection of financial statement fraud—An implementation of data mining framework," *Int. J. Adv. Comput. Sci. Appl.*, vol. 3, no. 8, pp. 150–160, 2012.
[30] P. Hajek, "Interpretable fuzzy rule-based systems for detecting financial statement fraud," in *Proc. IFIP Int. Conf. Artif. Intell. Appl. Innov.* Cham, Switzerland: Springer, 2019, pp. 425–436.
[31] P. Hajek and R. Henriques, "Mining corporate annual reports for intelligent detection of financial statement fraud—A comparative study of machine learning methods," *Knowl.-Based Syst.*, vol. 128, pp. 139–152, Jul. 2017.
[32] J. L. Hobson, W. J. Mayew, and M. Venkatachalam, "Analyzing speech to detect financial misreporting," *J. Accounting Res.*, vol. 50, no. 2, pp. 349–392, May 2012.
[33] S.-Y. Huang, R.-H. Tsaih, and F. Yu, "Topological pattern discovery and feature extraction for fraudulent financial reporting," *Expert Syst. Appl.*, vol. 41, no. 9, pp. 4360–4372, Jul. 2014.
[34] S. L. Humpherys, K. C. Moffitt, M. B. Burns, J. K. Burgoon, and W. F. Felix, "Identification of fraudulent financial statements using linguistic credibility analysis," *Decis. Support Syst.*, vol. 50, no. 3, pp. 585–594, Feb. 2011.
[35] A. G. Ivakhnenko, "The group method of data of handling; a rival of the method of stochastic approximation," *Sov. Autom. Control*, vol. 13, no. 3, pp. 43–55, 1968.
[36] C.-L. Jan, "An effective financial statements fraud detection model for the sustainable development of financial markets: Evidence from Taiwan," *Sustainability*, vol. 10, no. 2, p. 513, 2018.
[37] S. Karlos, N. Fazakis, S. Kotsiantis, and K. Sgarbas, "Semi-supervised forecasting of fraudulent financial statements," in *Proc. 20th Pan-Hellenic Conf. Informat.*, Nov. 2016, pp. 1–6.
[38] S. Karlos, G. Kostopoulos, S. Kotsiantis, and V. Tampakas, "Using active learning methods for predicting fraudulent financial statements," in *Proc. Int. Conf. Eng. Appl. Neural Netw.* Cham, Switzerland: Springer, 2017, pp. 351–362.
[39] Y. J. Kim, B. Baik, and S. Cho, "Detecting financial misstatements with fraud intention using multi-class cost-sensitive learning," *Expert Syst. Appl.*, vol. 62, pp. 32–43, Nov. 2016.
[40] E. Kirkos, C. Spathis, and Y. Manolopoulos, "Data mining techniques for the detection of fraudulent financial statements," *Expert Syst. Appl.*, vol. 32, no. 4, pp. 995–1003, 2007.
[41] S. Kirkos and Y. Manolopoulos, "Data mining in finance and accounting: A review of current research trends," in *Proc. 1st Int. Conf. Enterprise Syst. Accounting (ICESAcc)*, 2004, pp. 63–78.
[42] B. Kitchenham, "Procedures for performing systematic reviews," *Keele, UK, Keele Univ.*, vol. 33, no. 2004, pp. 1–26, 2004.
[43] H. Klein. (2020). *Consumer Fraud Complaints Hit Record High*. [Online]. Available: https://www.eisneramper.com/occupational-fraud-financial-cost-1116/
[44] S. Kotsiantis and D. Kanellopoulos, "Multi-instance learning for predicting fraudulent financial statements," in *Proc. 3rd Int. Conf. Converg. Hybrid Inf. Technol.*, vol. 1, Nov. 2008, pp. 448–452.
[45] J. R. Koza and R. Poli, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, vol. 1. Cambridge, MA, USA: MIT Press, 1992.
[46] B. S. Kumar and V. Ravi, "A survey of the applications of text mining in financial domain," *Knowl.-Based Syst.*, vol. 114, pp. 128–147, Dec. 2016.
[47] X. Li, W. Xu, and X. Tian, "How to protect investors? A GA-based DWD approach for financial statement fraud detection," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2014, pp. 3548–3554.
[48] C.-C. Lin, A.-A. Chiu, S. Y. Huang, and D. C. Yen, "Detecting the financial statement fraud: The analysis of the differences between data mining techniques and experts' judgments," *Knowl.-Based Syst.*, vol. 89, pp. 459–470, Nov. 2015.
[49] T. Lingbing, P. Pin, and L. Changqing, "Financial statement fraud detection through multiple instance learning," *Sci. Bull. Nat. Mining Univ.*, vol. 3, pp. 146–155, Jan. 2016.

[50] T. E. McKee, "A meta-learning approach to predicting financial statement fraud," *J. Emerg. Technol. Accounting*, vol. 6, no. 1, pp. 5–26, Jan. 2009.
[51] S. O. Moepya, S. S. Akhoury, and F. V. Nelwamondo, "Applying cost-sensitive classification for financial fraud detection under high class-imbalance," in *Proc. IEEE Int. Conf. Data Mining Workshop*, Dec. 2014, pp. 183–192.
[52] S. O. Moepya, F. V. Nelwamondo, and B. Twala, "Increasing the detection of minority class instances in financial statement fraud," in *Proc. Asian Conf. Intell. Inf. Database Syst.* Cham, Switzerland: Springer, 2017, pp. 33–43.
[53] M. Y. M. Narekar and M. S. K. Chavan, "A review on credit card fraud detection using BLAST-SSAHA method," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 4, no. 11, pp. 425–433, 2015.
[54] H. Patel, S. Parikh, A. Patel, and A. Parikh, "An application of ensemble random forest classifier for detecting financial statement manipulation of indian listed companies," in *Recent Developments in Machine Learning and Data Analytics*. Singapore: Springer, 2019, pp. 349–360.
[55] J. Perols, "Financial statement fraud detection: An analysis of statistical and machine learning algorithms," *Auditing A, J. Pract. Theory*, vol. 30, no. 2, pp. 19–50, May 2011.
[56] A. Pinkasovitch. (2019). *Detecting Financial Statement Fraud*. [Online]. Available: https://www.investopedia.com/articles/financial-theory/11/detecting-financial-fraud.asp
[57] G. Pradeep, V. Ravi, K. Nandan, B. Deekshatulu, I. Bose, and A. Aditya, "Fraud detection in financial statements using evolutionary computation based rule miners," in *Proc. Int. Conf. Swarm, Evol., Memetic Comput.* Cham, Switzerland: Springer, 2014, pp. 239–250.
[58] S. Ragothaman, J. Carpenter, and T. Buttars, "Using rule induction for knowledge acquisition: An expert systems approach to evaluating material errors and irregularities," *Expert Syst. Appl.*, vol. 9, no. 4, pp. 483–490, Jan. 1995.
[59] P. Ravisankar, V. Ravi, G. R. Rao, and I. Bose, "Detection of financial statement fraud and feature selection using data mining techniques," *Decision Support Syst.*, vol. 50, no. 2, pp. 491–500, 2011.
[60] A. Reurink, "Financial fraud: A literature review," *J. Econ. Surveys*, vol. 32, no. 5, pp. 1292–1325, Dec. 2018, doi: 10.1111/joes.12294.
[61] A. A. Rizki, I. Surjandari, and R. A. Wayasti, "Data mining application to detect financial fraud in Indonesia's public companies," in *Proc. 3rd Int. Conf. Sci. Inf. Technol. (ICSITech)*, Oct. 2017, pp. 206–211.
[62] E. Sengupta, N. Jain, D. Garg, and T. Choudhury, "A review of payment card fraud detection methods using artificial intelligence," in *Proc. Int. Conf. Comput. Techn., Electron. Mech. Syst. (CTEMS)*, Dec. 2018, pp. 494–498.
[63] C. Serrano-Cinca, "Self organizing neural networks for financial diagnosis," *Decis. Support Syst.*, vol. 17, no. 3, pp. 227–238, Jul. 1996.
[64] V. Shah, P. Shah, H. Shetty, and K. Mistry, "Review of credit card fraud detection techniques," in *Proc. IEEE Int. Conf. Syst., Comput., Autom. Netw. (ICSCAN)*, Mar. 2019, pp. 1–7.
[65] A. Sheshasayee and S. S. Thomas, "Implementation of data mining techniques in upcoding fraud detection in the monetary domains," in *Proc. Int. Conf. Innov. Mech. Ind. Appl. (ICIMIA)*, Feb. 2017, pp. 730–734.
[66] J.-Y. Shih, "Using self-organizing maps for analyzing credit rating and financial ratio data," in *Proc. IEEE Int. Summer Conf. Asia Pacific Bus. Innov. Technol. Manage.*, Jul. 2011, pp. 109–112.
[67] X.-P. Song, Z.-H. Hu, J.-G. Du, and Z.-H. Sheng, "Application of machine learning methods to risk assessment of financial statement fraud: Evidence from China," *J. Forecasting*, vol. 33, no. 8, pp. 611–626, Dec. 2014.
[68] D. F. Specht, "Probabilistic neural networks," *Neural Netw.*, vol. 3, no. 1, pp. 109–118, 1990.
[69] G. S. Temponeras, S.-A.-N. Alexandropoulos, S. B. Kotsiantis, and M. N. Vrahatis, "Financial fraudulent statements detection through a deep dense artificial neural network," in *Proc. 10th Int. Conf. Inf., Intell., Syst. Appl. (IISA)*, Jul. 2019, pp. 1–5.
[70] J. Thongkam, G. Xu, Y. Zhang, and F. Huang, "Support vector machine for outlier detection in breast cancer survivability prediction," in *Proc. Asia–Pacific Web Conf.* Berlin, Germany: Springer, 2008, pp. 99–109.
[71] C. S. Throckmorton, W. J. Mayew, M. Venkatachalam, and L. M. Collins, "Financial fraud detection using vocal, linguistic and financial cues," *Decis. Support Syst.*, vol. 74, pp. 78–87, Jun. 2015.
[72] M. Lokanan, V. Tran, and N. H. Vuong, "Detecting anomalies in financial statements using machine learning algorithm: The case of vietnamese listed firms," *Asian J. Accounting Res.*, vol. 4, no. 2, pp. 181–201, Oct. 2019.
[73] J. West and M. Bhattacharya, "Intelligent financial fraud detection: A comprehensive review," *Comput. Secur.*, vol. 57, pp. 47–66, Mar. 2016.
[74] C. Wohlin and R. Prikladnicki, "Systematic literature reviews in software engineering," *Inf. Softw. Technol.*, vol. 55, no. 6, pp. 919–920, 2013.
[75] X.-S. Yang, "Firefly algorithms for multimodal optimization," in *Proc. Int. Symp. Stochastic Algorithms*. Berlin, Germany: Springer, 2009, pp. 169–178.
[76] J. Yao, J. Zhang, and L. Wang, "A financial statement fraud detection model based on hybrid data mining methods," in *Proc. Int. Conf. Artif. Intell. Big Data (ICAIBD)*, May 2018, pp. 57–61.
[77] J. Yao, Y. Pan, S. Yang, Y. Chen, and Y. Li, "Detecting fraudulent financial statements for the sustainable development of the socio-economy in China: A multi-analytic approach," *Sustainability*, vol. 11, no. 6, p. 1579, Mar. 2019.
[78] C.-C. Yeh, D.-J. Chi, T.-Y. Lin, and S.-H. Chiu, "A hybrid detecting fraudulent financial statements model using rough set theory and support vector machines," *Cybern. Syst.*, vol. 47, no. 4, pp. 261–276, May 2016.
[79] D. Yue, X. Wu, Y. Wang, Y. Li, and C.-H. Chu, "A review of data mining-based financial fraud detection research," in *Proc. Int. Conf. Wireless Commun., Netw. Mobile Comput.*, Sep. 2007, pp. 5519–5522.
[80] D. Yue, X. Wu, N. Shen, and C.-H. Chu, "Logistic regression for detecting fraudulent financial statement of listed companies in China," in *Proc. Int. Conf. Artif. Intell. Comput. Intell.*, 2009, pp. 104–108.
[81] E. Zouboulidis and S. Kotsiantis, "Forecasting fraudulent financial statements with committee of cost-sensitive decision tree classifiers," in *Proc. Hellenic Conf. Artif. Intell.* Berlin, Germany: Springer, 2012, pp. 57–64.

MATIN N. ASHTIANI received the B.Sc. degree in computer science from the Sharif University of Technology, and the M.Sc. degree from the University of Tehran, Iran, in 2016. She is currently pursuing the Ph.D. degree in digital transformation and innovation with the University of Ottawa, Canada. She has done researches in the areas of machine learning, data mining, natural language processing, and fintech. She is also a member of the Knowledge Discovery and Data Mining (KDD) Laboratory, Telfer School of Management, University of Ottawa, where she is participating in various academic and industrial projects. Her current research interests include design and development of machine learning and data mining techniques and algorithms for intelligent financial fraud detection and intelligent financial market forecasting.



BIJAN RAAHEMI received the Ph.D. degree in electrical and computer engineering from the University of Waterloo, in 1997. He was a Senior Researcher in telecommunications industry focusing on computer networks architectures, data, and multimedia communications and services. He is currently a Professor of information systems and analytics, and the Founder and the Director of the Knowledge Discovery and Data mining (KDD) Laboratory, University of Ottawa, Canada. His research focus with the KDD Laboratory is on the novel algorithms in data analytics and machine learning, including big data analytics, anomaly detection in high-dimensional data, as well as emerging applications of machine learning and data mining in engineering, business, and healthcare. His work has appeared in more than 70 peer-reviewed journals and conference proceedings. He also holds eight patents in data communications. He is a Registered Member of the Professional Engineers of Ontario (PEO), and a member of the Association for Computing Machinery (ACM). He is a Co-Editor of the *Handbook of Research on Data Science in Healthcare*.