Journal of Data Science 14(2016), 553-570

Detecting Financial Fraud Using Data Mining Techniques: A Decade Review from 2004 to 2015

Mousa Albashrawi[1,2]

1Department of Accounting & Management Information Systems

2Department of Operations & Information Systems

Abstract: Objective: Financial fraud has been a big concern for many organizations across industries; billions of dollars are lost yearly because of this fraud. So businesses employ data mining techniques to address this continued and growing problem. This paper aims to review research studies conducted to detect financial fraud using data mining tools within one decade and communicate the current trends to academic scholars and industry practitioners.

Method: Various combinations of keywords were used to identify the pertinent articles. The majority of the articles retrieved from Science Direct but the search spanned other online databases (e.g., Emerald, Elsevier, World Scientific, IEEE, and Routledge - Taylor and Francis Group). Our search yielded a sample of 65 relevant articles (58 peer-reviewed journal articles with 7 conference papers). One-fifth of the articles was found in Expert Systems with Applications (ESA) while about one-tenth found in Decision Support Systems (DSS).

Results: 41 data mining techniques were used to detect fraud across different financial applications such as health insurance and credit card. Logistic regression model appeared to be the leading data mining tool in detecting financial fraud with a 13% of usage.In general, supervised learning tool have been used more frequently than the unsupervised ones. Financial statement fraud and bank fraud are the two largest financial applications being investigated in this area – about 63%, which corresponds to 41 articles out of the 65 reviewed articles. Also, the two primary journal outlets for this topic are ESA and DSS.

Conclusion: This review provides a fast and easy-to-use source for both researchers and professionals, classifies financial fraud applications into a high-level and detailed-level framework, shows the most significant data mining techniques in this domain, and reveals the most countries exposed to financial fraud.

Keywords: Financial fraud, fraud detection, data mining techniques, literature review.

1. Introduction

Financial fraud has been a big concern for many organizations across industries and in different countries since it brings huge devastations to business. Billions of dollars are lost yearly due to financial fraud; Bank of America, for example, agrees to pay $16.5 billion for resolving financial fraud case [49]. Also, IRS (2014) indicates that Mr. Walker, the founder of Bixby Energy Systems, deceived more than 1,800 investors and committed multi-million dollar fraud. His fraudulent actions involve providing false statements of a) his subordinates' salaries and commissions; b) the operational capacity of the firm's core products, and c) an initial public stock offering [30]. Hence, the numbers still indicate this is a growing problem, which needs more attention from professionals and academicians.

Financial fraud detection tools have been brought to scenic in order to address this problem and to provide reliable solutions to business. Financial fraud is normally discovered through outlier detection process [32] enabled by data mining techniques, which also identify valuable information by revealing hidden trends, relationships, patterns found in a large database [25]. Data mining, defined as "a process that uses statistical, mathematical, artificial intelligence, and machine learning techniques to extract and identify useful information and subsequently gain knowledge from a large database" [50], is a major contributor for detecting different types of financial fraud through its diverse methods, such as, logistic regression, decision tree, support vector machine (SVM), neural network (NN) and naïve Bayes. Some of these techniques outperform the others in specific financial contexts. Glancy and Yadav (2011) divide those contexts to three main areas: internal, insurance and credit [22]. Jans et al. (2011) further classify internal fraud into two categories: financial statement fraud and transaction fraud [31]. They define financial statement fraud as "the intentional misstatement of certain financial values to enhance the appearance of profitability and deceive shareholders or creditors" while transaction fraud captures the process of snatching organizational assets.

Although detecting financial fraud is considered a high priority for many organizations, the current literature lacks for an up-to-date, comprehensive and in-depth review that can help firms with their decisions of selecting the appropriate data mining technique. Ngai et al. (2011) provide a well-organized and detailed literature review on detecting financial fraud via data mining methods based on 49 articles ranging from 1997 to 2008 [50]. However, the specified time period is not able to capture the increasing trend of research in this area, specifically in the year of 2011, which is considered as a record year in financial fraud [11]. This has motivated us to extend Ngai et al.'s review and contribute by 1) revealing which context should implement what technique of data mining, 2) unfolding what technique can yield a higher classification accuracy in detecting financial fraud, 3) providing a new classification framework for financial fraud, and 4) expanding the sample of the reviewed articles to make it one of the most comprehensive reviews on this topic. Overall, this paper is an attempt to leverage our knowledge and to increase our understanding of data mining applications in financial fraud.

2. Literature Review

Due to its high importance, financial fraud has been given a considerable attention in prior research. Literature has tapped on different types of financial fraud using different methods of data mining. Table 1 presents the 65 examined articles in chronological order. From the table, we can determine what methods are being frequently implemented for which case of financial fraud and what method can work best across fraud types. For example, the logistic model can help in detecting financial fraud in automobile insurance, corporate insurance, financial statement, and credit card but it can be considered the best-performing method in the context of corporate insurance fraud.

Table 1: Summarized work for detecting financial fraud via data mining techniques (2004-2015)

No.	Fraud Type	Dataset Used	Data Mining Technique Employed	Reference	Best-Performing Technique (Highest Accuracy)[a]

1	Financial statement fraud	158 firm (79 fraud, 79 non-fraud): 1982-1999[b]	Discriminant analysis	[35]	
2	Insurance auto fraud	1,399 personal injury protection (PIP) automobile insurance claims: 1993	Naïve Bayes	[68]	
3	Automobile insurance fraud	4083 cases (245 fraud, 3838 non-fraud)	NN, naïve Bayes and decision tree	[56]	
4	Automobile insurance fraud	1,399 PIP automobile insurance claims: 1993	NN	[70]	
5	Automobile insurance fraud	Spanish automobile insurance claims (half fraud, half legitimate): 1993-1996	Logit model	[6]	
6	Insurance auto fraud	Insurance hypothetical data	Fuzzy logic	[54]	
7	Financial statement fraud	27 firms: 2004-2005[b]	Genetic algorithm	[37]	
8	Corporate insurance fraud	82,807 firms from RMA (US agency under Dept. of Agriculture): 2001	Logit and probit models	[33]	Logit model
9	Fraudulent financial statements	164 Greek firms (41 fraud, 123 non-fraud): 2001-2002	Decision trees, NN, Bayesian network, SVM and nearest neighbour	[40]	Decision tree
10	Health insurance fraud	1812 medical cases (906 fraud, 906 non-fraud)	Process mining	[75]	
11	Accounting fraud	8000 public firms[b]	Logit model, K-means clustering, and decision tree	[71]	
12	Financial statement fraud	Real-world financial data[b]	Genetic algorithm	[8]	
13	Credit card fraud	50 firms based on questionnaire-responded transaction (QRT) data	SVM	[9]	
14	Fraudulent financial statements	76 Greek manufacturing firms (38 fraud, 38 non-fraud)	Decision trees, NN and Bayesian belief networks	[38]	
15	Money laundering	Traditional suspicion data (disgruntled employees, banks, and informants)	Network analysis	[19]	
16	Automobile insurance fraud	2567 suspicious claims from Spanish insurance company	Probit model	[57]	
17	Financial statement fraud	312 service-based computer and technology firms: 1996-2001	Logit model and fuzzy logic	[41]	
18	Financial statement fraud	51 fraudulent firms: 1991-2003[b]	Genetic algorithm	[27]	
19	Automobile insurance fraud	2403 claims (2229 legitimate, 174	Logit model	[69]	

		fraudulent): 2000			
20	Fraudulent financial reporting	1,515 Taiwanese firms (6 fraud, 1,509 non-fraud): 2003-2004	NN, logistic regression and decision tree	[44]	Logistic regression
21	Credit card fraud	41,647 records (15,576 fraud, 26,071 non-fraud)	Artificial immune systems, NN, Bayesian nets, Naïve Bayes and decision tree	[17]	Artificial immune systems
22	Fraudulent credit card transactions	77,345 fraud, 2,943,695 non-fraud transactions	Supervised and unsupervised classification	[34]	
23	Corporate financial fraud	274 firms (137 fraudulent, 137 non-fraudulent: 2002-2004	Logistic regression model	[77]	
24	Credit card fraud	102,000 ATM and POS transactions	Stream clustering	[66]	
25	Insurance auto fraud	10,000 automobile claims (9,899 legitimate, 101 fraudulent): 2000	Bayesian analysis	[3]	
26	Healthcare fraud	60,962 observations from Medicaid payment data	Stepwise multi-stage clustering	[47]	
27	Financial statements fraud	148 firms (24 fraud, 124 non-fraud)	Classification and Regression Tree (CART)	[2]	
28	Credit card fraud	525 credit card transactions	Self-organizing map	[58]	
29	Financial statement fraud	398 Greek firms (199 fraud, 199 non-fraud): 2001-2004	Discriminant analysis, logistic regression, nearest neighbor, NN, SVM, UTilités Additives DIScriminantes (UTADIS), and Multi-group hierarchical discrimination (MHDIS)	[18]	UTADIS
30	Credit card fraud	1,959 clients with 12,107 transactions	Association rules	[62]	
31	Credit card fraud	25,000 payment observations (5,529 fraud, 19,471 non-fraud) from Taiwanese bank: 2005	K-nearest neighbor, logistic model, discriminant analysis, Naïve Bayes, NN, and decision tree	[76]	NN
32	Credit card fraud	2,000 synthetic transactions	Rule-based filtering, Dempster–Shafer adder and Bayesian learning	[53]	Bayesian learning
33	Occupational financial fraud	80 intra-company messages (40 disgruntled and 40 non -disgruntled)	Naïve Bayes	[26]	
34	Financial statement fraud	100 Chinese firms: 1999-2006	Self-organizing map and K-means clustering	[13]	
35	Financial statement fraud	3,319 firms (132 fraud, 3,187 non-fraud): 1999-2006[b]	SVM using custom financial kernel	[7]	
36	Financial	126 Turkish	Three-phase cutting	[15]	

	statement fraud	manufacturing firms (17 fraud, 109 non-fraud)	plane algorithm		
37	Money laundering	20 firms in industrial peer group (IPG) data (5 fraud, 15 non-fraud)	A multiple-criteria index	[74]	
38	Credit card fraud	81,137 observations (67,763 normal, 13,374 rare)	K-means clustering and SVM	[73]	K-means clustering
39	Plastic credit fraud	413,991 transactions (10,484 fraud, 403,507 non-fraud)	Hybrid model(supervised and unsupervised techniques)	[39]	
40	Credit card fraud	70,465 fraud records	Variable binned scatter plot visualization	[24]	
41	Auditing multi-financial fraud	168 fraud firms	NN	[36]	
42	Fraudulent financial reporting	10-K reporting[b]: 2006-2008	Text mining	[22]	
43	Healthcare insurance fraud	Two major US health insurance firms (65 million claims)	Repeated bisections, repeated bisections with global optimization and direct K-way clustering	[21]	Repeated bisection clustering
44	Financial fraud by top management	75 firms from Taiwan's stock market (25 fraud, 50 non-fraud)	SVM	[52]	
45	Transaction fraud in procurement	10,000 process instances from ERP system (SAP)	Process mining	[31]	
46	Financial statement fraud	79,651 firm (293 fraud, 79,358 non-fraud): 1982-2005[b]	Logistic regression	[12]	
47	Credit card fraud	About 5 million transactions (2,420 fraud, the remaining non-fraud)	SVM, random forests and logistic regression	[4]	Random forests
48	Financial statement fraud	Anonymous firm's financial data	Response surface method	[78]	
49	Financial statement fraud	202 companies from Chinese stock exchanges (101 fraud, 101 non-fraud)	SVM, genetic programming, multi-layer feedforward (MLFF), group method of data handling (GMDH), logistic regression, and NN	[59]	NN
50	Financial statement fraud	15,985 firms (51 fraud, 15,934 non-fraud): 1998-2005[b]	Logistic regression, bagging, SVM, NN, C4.5 decision tree and stacking	[55]	Logistic regression and SVM
51	Life insurance fraud	40,080 group insurance claims	K-means clustering	[67]	
52	Fraudulent financial statements	202 firms (101 fraud, 101 non-fraud): 1995-2004[b]	Logistic regression, C 4.5 decision tree, Naïve Bayes, locally weighted learning (LWL), and SVM	[29]	Naïve Bayes and C4.5 decision tree
53	Automotive insurance fraud	98 claims (49 fraud, 49 non-fraud)	Survival analysis, discriminant and logit	[20]	Logit analysis

			analysis, NN and decision tree		
54	Credit card fraud	Online shopping firm's transactions data	Density-based clustering	[14]	
55	Fraudulent bank accounts	10,216 bank accounts (327 fraudulent, 9,889 normal)	Bayesian classification and association rule	[42]	
56	Quarterly and annual financial reports	189 firms with one or more fraud incidents: 1994-2006[b]	SVM	[64]	
57	Financial fraud by management	228 firms (114 fraud, 114 non-fraud): 1998-2002[b]	Probit regression, logistic regression, random forests, stochastic gradient boosting, rule ensemble, and partially adaptive estimators (SGT, EGB2 and HIS)	[72]	Rule ensemble
58	General financial fraud using financial ratios	9,006 firms (815 fraudulent, 8,191 legitimate): 1995-2010[b]	MetaFraud framework	[1]	
59	Credit card fraud	About 22 million credit card transactions (978 fraud, 22 million non-fraud).	Cost-sensitive decision tree	[61]	
60	Tax fraud	532,755 taxpayer enterprises data: 2005-2007	NN, decision tree, and Bayesian networks	[23]	NN
61	Fraudulent financial reporting	116 firms (58 fraud, 58 non-fraud): 1992-2006	Growing hierarchical self-organizing map (GHSOM)	[28]	
62	Credit card fraud	Anonymous 40,918 transactions	Frequent itemset mining, SVM, nearest neighbor, naïve Bayes and random forest	[63]	
63	Credit card fraud	10,000 accounts (100 fraud, 9,900 non-fraud)	Self-organizing map	[51]	
64	Credit card fraud	9,387 transactions from Turkish bank (8,448 legitimate, 9,39 fraudulent)	Fisher discriminant analysis, decision tree, NN and Naïve Bayes	[48]	
65	Financial statement fraud	576 firms (129 fraud, 447 non-fraud): 1998-2010	NN, decision tree, and logistic model	[45]	NN

[a] If many data mining techniques are applied, the best-performing technique is indicated, if reported.
[b] Securities and Exchange Commission (SEC's) Accounting and Auditing Enforcement Releases (AAERs)

3. Method

A number of keywords was used to identify the pertinent articles, for instance, "detecting financial fraud, financial fraud and data mining, financial fraud detection, and detecting financial fraud via data mining". Most of the relevant articles were found in MIS related journals, e.g., Expert Systems with Applications and Decision Support Systems but some were found in finance and economic related journals, e.g., Journal of Risk and Insurance, and Applied

Economics. Table 2 lists thirty-nine titles for both journals and conferences included in our analysis.

Although the majority of the articles retrieved from Science Direct, the search spanned other online databases (e.g., Emerald, Elsevier, World Scientific, IEEE, and Routledge - Taylor and Francis Group). Our search yielded a sample of 65 relevant articles (58 peer-reviewed journal articles with 7 conference papers). One-fifth of the articles was found in Expert Systems with Applications while about one-tenth found in Decision Support Systems (Table 2). Hence, these two journals have been the primary outlet for this topic. However, most of the articles had been conducted in the United States, followed by Taiwan, China and Spain (Table 3).

Table 2: Distribution of articles by journals and conferences (2004–2015)

Journal/Conference Title	**Frequency**	**Percentage (%)**
Expert Systems with Applications	13	20
Decision Support Systems	6	9.23
Managerial Auditing Journal	4	6.15
Knowledge-Based Systems	3	4.62
The Journal of Risk and Insurance	2	3.08
ACM SIGKDD International Conference on Knowledge Discovery and Data mining	2	3.08
International Journal of Intelligent Systems in Accounting and Finance Management	2	3.08
Computational Intelligence	2	3.08
MIS Quarterly	1	1.54
Management Science	1	1.54
Contemporary Accounting Research	1	1.54
Journal of Forecasting	1	1.54
Journal of Data Science	1	1.54
Computers in Human Behavior	1	1.54
Information Fusion	1	1.54
Journal of Practice & Theory	1	1.54
Journal of Economic Policy Reform	1	1.54
IEEE Transaction on Knowledge and Data Engineering	1	1.54
Journal of Money Laundering Control	1	1.54
Journal of Pattern Recognition and Artificial Intelligence	1	1.54
Insurance: Mathematics and Economics	1	1.54
Journal of Information Technology & Decision Making	1	1.54
Applied Economics	1	1.54
European Journal of Operational Research	1	1.54
Data Mining and Knowledge Discovery	1	1.54
International Journal of Computer Applications	1	1.54
Data Mining IX	1	1.54
Journal of Digital Accounting Research	1	1.54
Journal of Emerging Technologies in Accounting	1	1.54
Genetic and Evolutionary Computation Conference	1	1.54
IEEE International Conference on Fuzzy System	1	1.54
Computational Statistics: 18th Symposium (COMPSTAT 2008)	1	1.54
Computational Statistics and Data Analysis	1	1.54
International Journal of Management	1	1.54
SPIE Electronic Imaging Conference	1	1.54
IEEE International Conference on Granular Computing	1	1.54
International Conference on Artificial Immune Systems	1	1.54
The Scientific World Journal	1	1.54
ACM SIGKDD Explorations	1	1.54
Total	**65**	**100**

Table 3: The number of articles for detecting financial fraud by countries

Country	**Frequency**	**Percentage (%)**

United States	23	35.38
Taiwan	8	12.31
China	7	10.77
Spain	4	6.15
Turkey	3	4.62
Greece	3	4.62
India	3	4.62
UK	3	4.62
Canada	2	3.08
Chile	2	3.08
Europe[a]	1	1.54
Poland	1	1.54
France	1	1.54
Cyprus	1	1.54
Brazil	1	1.54
Singapore	1	1.54
Australia	1	1.54
Total	**65**	**100**

[a]European region was only reported

4. Results

This section highlights the most frequent data mining techniques used in financial fraud associated with their usage frequency, description and business application. Also, based on the reviewed different applications of financial fraud, this section provides a new classification scheme at two levels: high and detailed.

4.1. Usage Frequency of Data Mining Techniques

Out of 41 data mining techniques used in the reviewed articles, Table 4 shows the most applied ones in a period ranging from 2004 to 2015. Logistic regression model appears to be the leading data mining technique in detecting financial fraud with a 13%, followed by both of neural network and decision tree, with a 11%. While support vector machine is represented by a 9% and naïve Bayes is represented by a 6%. Besides fraud detection, data mining techniques can address a wide array of business applications, for example, bankruptcy prediction, sales forecasting and scheduling optimization as shown in Table 4.

Table 4: Most used data mining methods, their usage frequency, description and general business application

No.	Method	Frequency	Description	Business Application
1	Logistic regression	17	It is a typical classification method used to generate dichotomous possible values [59].	Prediction of failure probability in selling a specific product
2	Neural network	15	ANN shows better results when testing large sets of data. It consists of neurons or nodes [43].	Credit Rating
3	Decision trees	15	Decision tree or classification tree is a method for assigning and classifying data points into predefined clusters via splitting rules [20].	Stock market prediction
4	Support vector machine	12	SVM is a statistical method that used for linear classification [4].	Bankruptcy prediction
5	Naïve Bayes	8	This tool has the capability of predicting group membership [26].	Sentiment analysis
6	Bayesian	7	“Directed acyclic graph, used to predict the	Tracking performance over

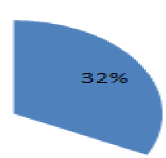
	networks		likelihood of different outcomes, based on a set of facts” [23].	time
7	Discriminant analysis	6	This technique can predict group membership of linearly combined variables [65].	Credit worthiness
8	Nearest neighbor	4	“New data points are classed according to the classes of the points which are closest to them in the training data” [5].	Money laundering analysis
9	K-means clustering	4	K-means is a clustering method that can generate clusters with uniform shapes and it is generally measured by squared Euclidean distance [73].	Market price and cost modeling
10	Self-organizing map	4	This technique, introduced by Kohonen, can identify similarities between objects in multidimensional space [51].	Project prioritization and selection
11	Random forests	3	“A random forest is an ensemble of unpruned classification or regression trees induced from bootstrap samples of the training data, using random feature selection in the tree induction process” [4].	Credit risk prediction
12	Genetic algorithm	3	This tool, an evolutionary computation approach, can handle non-linear functions of multiple variables [27].	Marketing mix strategizing
13	Probit model	3	Probit model uses the assumption of a symmetric distribution with fairly thin tails [72].	Probability of marketing campaign failure
14	Association rules	2	This tool uses “if” and “then” to unfold related items [62].	Market basket analysis
15	Process mining	2	This algorithm gives access to knowledge via mining event logs to analyze system processes [31].	Fraud detection
16	Fuzzy logic	2	This algorithm can deal with human reasoning and decision-making processes.	Models for project risk assessment

This table demonstrates that the supervised learning techniques (e.g., neural network, decision tree, support vector machine, and naïve Bayes) have been used more frequently than the unsupervised ones (e.g., clustering, association rules, and fuzzy logic). Thus, it could be stated that supervised learning techniques are better-performing tools than the unsupervised ones in detecting financial fraud.

4.2. Classification Framework Based on Fraud Type

Based on the analysis of the reviewed articles in this area, it is possible to classify financial fraud at a high-level into four major categories, namely, financial statement fraud, bank fraud, insurance fraud, and other related financial fraud (Table 5). The table shows the number of articles found in each type of financial fraud while the small pieces of pie chart represent those numbers in percentages. It is evident that financial statement fraud and bank fraud constitute the largest portion (63%) – this percentage corresponds to 41 articles out of the 65 reviewed articles.

Table 5: Classification of fraud types examined by data mining methods in one decade

Fraud Type (application)	**Article Count**	**Description**	**Percentage in Chart**
Financial statement fraud	21	This type of fraud is prevalent in today business world and one of the biggest challenges faced by managers and investors. It is basically the act of intentional or irresponsible conducts and conveys deception or misrepresentation; this produces materially misleading	

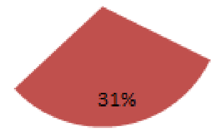
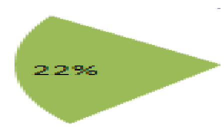
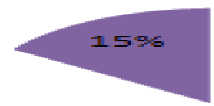
		financial statements [2] and reveals unauthorized benefit.	
Bank fraud	20	"Whoever knowingly executes, or attempts to execute, a scheme or artifice—(1) to defraud a financial institution; or (2) to obtain any of the moneys, funds, credits, assets, securities, or other property owned by, or under the custody or control of, a financial institution, by means of false or fraudulent pretenses, representations, or promises" [10]. Bank fraud is sub-categorized here into credit card fraud, money laundering, and fraudulent bank account.	
Insurance fraud	14	This term is broadly labeled as insurance abuse, especially in practice [69]. Insurance fraud includes here auto insurance fraud, healthcare insurance fraud, and corp insurance fraud.	
Other related financial fraud	10	Other financial fraud category includes general financial fraud, fraudulent financial reporting, financial fraud by top management, tax fraud, and transaction fraud.	
Total	**65**		**100%**

Table 6 further classifies and provides in-depth analysis by indicating the frequency of the sub-categories of financial fraud types. Bank fraud is subcategorized into credit card fraud, money laundering, and fraudulent bank account while insurance fraud is subcategorized into healthcare fraud, auto fraud, and corp fraud.

Table 6: Further break-down for fraud types with corresponding data mining techniques

Techniques	**Fraud Types**							
		Bank fraud			Insurance fraud			
	Financial statement fraud	Credit card fraud	Money laundering	Fraudulent bank account	Healthcare	Auto	Corp	Other related financial fraud
Logistic regression	5	2				3	2	1
Neural network	4	3				3		4
Decision trees	5	4				2		3
Discriminant analysis	2	2				1		2
Bayesian networks	2	3		1		1		1
SVM	3	3						3
Nearest neighbor	2	1						
Association rules		1		1				
Rule-based filtering		1						
Dempster–Shafer adder		1						
Naïve Bayes		4				2		2
Three-phase cutting plane algorithm	1							
A multiple-criteria index			1					
Text mining								1
Process mining					1			1
Random forests		2						1
Response surface method	1							
Genetic programming	1							
MHDIS	1							
GMDH	1							
MLFF	1							
LWL	1							

Bagging and Stacking	1						
Stochastic gradient boosting							1
Rule ensemble							1
MetaFraud framework							1
Network analysis			1				
Self-organizing map	2	2					
Probit model					1	1	
K-means clustering	2	1					
Density-based clustering		1					
Genetic algorithm	3						
Stepwise multi-stage clustering				1			
Fuzzy logic	1				1		
Repeated bisection clustering				1			
Stream clustering		1					
Un/supervised classification		2					
Variable binned scatter plot		1					
Artificial immune systems		1					
Frequent itemset mining		1					
Survival analysis					1		

The proposed classification framework can work as a reference in guiding financial fraud detection research through providing the help to scholars in identifying the demanding areas that need more attention. This framework can also provide industry professionals an index to select the appropriate data mining technique for a specific context of financial fraud. For example, firms that suffer from credit card fraud, they have an option of employing any of the supervised learning tools (i.e., naïve Bayes, decision tree, neural network, and SVM) and it is recommended to go with the most frequent used technique; decision tree. As noted, this selection is based on the fraud context and data mining technique frequency but it can be also based on performance (Table 2).

Table 7 and Chart 1: Yearly distribution of the articles on detecting financial fraud

Year	Amount
2004	3
2005	5
2006	5
2007	6
2008	9
2009	6
2010	7
2011	12
2012	5
2013	2
2014	3
2015	2
Total	**65**

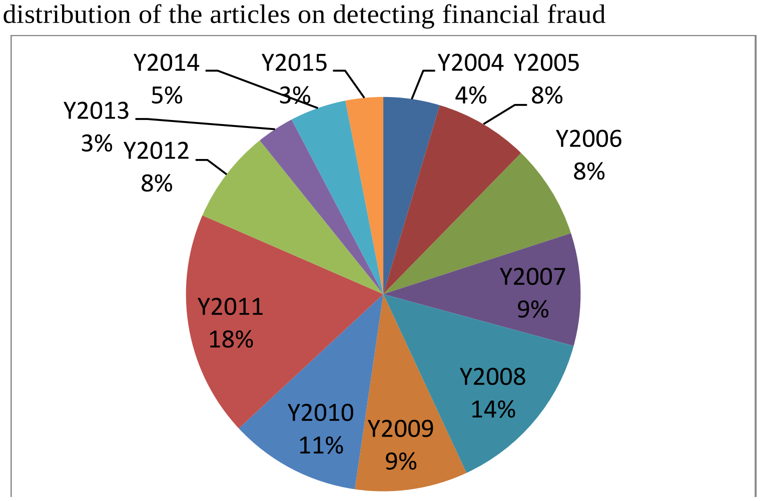


Table 7 and Chart 1 above highlight the yearly distribution of the 65 articles across the 10-year period. The gray highlighted years (2008, 2009, 2010 and 211) account for more than a half of publications in financial fraud detection. This high rate of publications reflects a serious growth in financial fraud across industries during these years. In particular, there had been a dramatic increase of the published papers during 2011. This increase seemed to be a natural response to the surge of fraud activities in that year; a 13% increase of financial fraud in 2011 compared to the previous year [60]. Also, abc NEWS (2012) indicated that the year of 2011 is considered the worst year for financial fraud on record [11].

5. Limitations and Conclusion

This review has some limitations. First, it does not consider all sub-categories of financial fraud, i.e., advanced-fee fraud that targets a very large number of people who looks for "work-from-home" opportunity. This fraud deceives people to pay a fee in advance so that they get the offer but once the fee is collected, they do not realize the expected benefits. Second, a decade review may not be sufficient to address this growing problem as it started when the business started. Third, the 65 articles explored may not reveal the entire story of data mining usage in the domain of financial fraud; several online databases need to be included in the sample for more powerful presentation and analysis.

However, it is crucial to have a wide-ranging review on detecting financial fraud in order to increase the understanding and to expand the knowledge of this area among researchers and professionals. This review sheds light on different valuable aspects of financial fraud detection:

- It provides a fast and easy-to-use source either for scholars or practitioners who are interested in the topic.
- It shows the importance of the investigated data mining techniques in the domain of financial fraud by presenting their frequency, usage percentage, and other general business applications. Although it is notable that logistic regression, decision tree, SVM, NN and Bayesian networks have been widely used (> 50%) to detect financial fraud, they are not always associated with the best classification results.
- This review provides high-level and detailed classification frameworks of financial fraud. The high-level framework includes four major types - financial statement fraud, bank fraud, insurance fraud, and other related financial fraud. The detailed framework sub-classifies bank fraud to credit card fraud, money laundering, and account bank fraud and sub-classifies insurance fraud to healthcare fraud, auto fraud, and corp fraud. Combining the two frameworks into a single integrated catalog scheme can help to classify any new type of financial fraud. However, it is apparent that financial statement fraud has been the most examined type in this area. Thus, it is necessary for business firms to be more cautious when they audit or process their financial statements.
- This paper emphasizes the huge increase of research conducted to address financial fraud in the years of 2008, 2009, 2011 and 2012. These four years account approximately for more than 50% of the publications in the 10-year period. More notably, the amount of research increased by 42% in 2011 compared to the previous year.
- Considering the country distribution table, it is possible to conclude that the countries (United States, Taiwan, China and Spain) that collectively had published 65% of the total articles on this topic, are being more exposed to it. In particular, the United States accounts for more than one-third (35%) of the papers published in this area.

In sum, the highlighted aspects through this review can provide organizations with useful information regarding the various types of financial fraud and data mining techniques available to them. Organizations may be able to select the most suitable technique once considering its particular usage context, frequency, and performance. This could lead to achieving a higher level of accuracy in detecting financial fraud. Besides this benefit, researchers can take advantage of knowing the most frequent used methods and in which context so that they can develop a research project to either investigating such method in a different context or suggesting a new innovative method in a similar context. However, the primary contribution of this paper is twofold; the first is to provide an up-to-date and comprehensive analysis of this crucial topic as an extension to Ngai et al.'s review. The second is to provide scholars and practitioners with an excellent source of data mining applications used in financial fraud for their fast access and use.

References

[1] Abbasi A, Albrecht C, Vance A, Hansen J. Metafraud: A Meta-Learning Framework for Detecting Financial Fraud. *MIS Quarterly* 2012; **36**: 1293-1327.

[2] Bai B, Yen J, Yang X. False Financial Statements: Characteristics of China's Listed Companies and CART Detecting Approach. *International Journal of Information Technology & Decision Making* 2008; **7**: 339–359.

[3] Bermúdez L, Pérez J, Ayuso M, Gómez E, Vázquez F. A Bayesian Dichotomous Model with Asymmetric Link for Fraud in Insurance. *Insurance: Mathematics and Economics* 2008; **42**: 779–786.

[4] Bhattacharyya S, Jha S, Tharakunnel K, Westland, JC. Data Mining for Credit Card Fraud: A Comparative Study. *Decision Support Systems* 2011; **50**: 602–613.

[5] Bidder OR, Campbell HA, Gómez-Laich A, Urgé P, Walker J, Cai Y, Wilson RP. Love Thy Neighbour: Automatic Animal Behavioural Classification of Acceleration Data Using the K-Nearest Neighbour Algorithm. *PLoS ONE* 2014; **9***:* 1-7.

[6] Caudill S, Ayuso M, Guill'en M. Fraud Detection Using A Multinomial Logit Model with Missing Information. *The Journal of Risk and Insurance* 2005; **72**: 539-550.

[7] Cecchini M, Aytug H, Koehler G, Pathak P. Detecting Management Fraud in Public Companies. *Management Science* 2010; **56**: 1146-1160.

[8] Chai W, Hoogs BK, Verschueren BT. Fuzzy Ranking of Financial Statements for Fraud Detection. *In proceeding of IEEE International Conference on Fuzzy System* 2006; 152–158.

[9] Chen R, Chen T, Lin C. A New Binary Support Vector System for Increasing Detection Rate of Credit Card Fraud. *International Journal of Pattern Recognition and Artificial Intelligence* 2006; **20**: 227–239.

[10] Cornell University Law School, White-Collar Crime: An Overview. *Retrieved from* https://www.law.cornell.edu/uscode/text/18/1344, 2015.

[11] Curry C. 2011 Was Worst Year for Suspected Financial Crimes on Record. *abc NEWS* 2012; *Retrieved from* http://abcnews.go.com/Business/record-year-fraud-reports-2011/story?id=15953781

[12] Dechow P, Ge W, Larson C, Sloan R. Predicting Material Accounting Misstatements. *Contemporary Accounting Research* 2011; **28**: 1-16.

[13] Deng Q, Mei G. Combining Self-Organizing Map and K-Means Clustering for Detecting Fraudulent Financial Statements. *In IEEE International Conference on Granular Computing* 2009; 126-131.

[14] Dharwa JN, Patel AR. A Data Mining with Hybrid Approach Based Transaction Risk Score Generation Model (TRSGM) for Fraud Detection of Online Financial Transaction. *International Journal of Computer Applications* 2011; **16**: 18-25.

[15] Dikmen B, Küçükkocaoğlu G. The Detection of Earnings Manipulation: The Three-Phase Cutting Plane Algorithm Using Mathematical Programming. *Journal of Forecasting* 2010; **29**: 442-466.

[16] Dimitras AI. Evaluation of Greek Construction Companies' Securities Using UTADIS Method. *European Research Studies* 2002; 95-107.

[17] Gadi MFA, Wang X, do Lago AP. Credit Card Fraud Detection with Artificial Immune System. I*n ICARIS '08 Proceedings of the 7th International Conference on Artificial Immune Systems* 2008; 119-131

[18] Gaganis C. Classification Techniques for the Identification of Falsified Financial Statements: A Comparative Analysis. *International Journal of Intelligent Systems in Accounting and Finance Management* 2009; **16**: 207-229.

[19] Gao Z, Ye M. A Framework for Data Mining-Based Anti-Money Laundering Research. *Journal of Money Laundering Control* 2007; **10**: 170-179.

[20] Gepp A, Wilson JH, Kumar K, Bhattacharya S. A Comparative Analysis of Decision Trees Vis-a-vis Other Computational Data Mining Techniques in Automotive Insurance Fraud Detection. *Journal of Data Science* 2012; **10**: 537-561.

[21] Ghani R, Kumar M. Interactive Learning for Efficiently Detecting Errors in Insurance Claims. *In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data mining* 2011; 325-333.

[22] Glancy FH, Yadav SB. A Computational Model for Financial Reporting Fraud Detection. *Decision Support Systems* 2011; **50**: 595–601.

[23] González PC, Velásquez JD. Characterization and Detection of Taxpayers with False Invoices Using Data Mining Techniques. *Expert Systems with Applications* 2013; **40**: 1427–1436.

[24] Hao MC, Dayal U, Sharma RK, Keim DA, Janetzko H. Visual Analytics of Large Multidimensional Data Using Variable Binned Scatter Plots. *In IS&T/SPIE Electronic Imaging* 2010.

[25] Hassani H, Gheitanchi S, Yeganegi MR. On the Application of Data Mining to Official Data. *Journal of Data Science* 2010; ***8***: 75-89.

[26] Holton C. Identifying Disgruntled Employee Systems Fraud Risk Through Text Mining: A Simple Solution for a Multi-Billion Dollar Problem. *Decision Support Systems* 2009; **46**: 853–864.

[27] Hoogs B, Kiehl T, Lacomb C, Senturk, D. A Genetic Algorithm Approach to Detecting Temporal Patterns Indicative of Financial Statement Fraud. *Intelligent Systems in Accounting, Finance and Management* 2007; **15**: 41-56.

[28] Huang SY, Tsaih RH, Yu F. Topological Pattern Discovery and Feature Extraction for Fraudulent Financial Reporting. *Expert Systems with Applications* 2014; **41**: 4360–4372.

[29] Humpherys SL, Moffitt KC, Burns MB, Burgoon JK, Felix WF. Identification of Fraudulent Financial Statements Using Linguistic Credibility Analysis. *Decision Support Systems* 2011; **50**: 585–594.

[30] IRS. Examples of Corporate Fraud Investigations. *Retrieved from IRS:* http://www.irs.gov/uac/Examples-of--Corporate-Fraud-Investigations-Fiscal-Year-2014, 2014.

[31] Jans M, Werf JM, Lybaert N, Vanhoof K. A Business Process Mining Application for Internal Transaction Fraud Mitigation. *Expert Systems with Applications* 2011; **38**: 13351–13359.

[32] Jayakumar GDS, Thomas BJ. A New Procedure of Clustering based on Multivariate Outlier Detection. *Journal of Data Science* 2013; **11**: 69-84.

[33] Jin Y, Rejesus R, Little B. Binary Choice Models for Rare Events Data: A Crop Insurance Fraud Application. *Applied Economics* 2005; **37**: 841–848.

[34] Juszczak P, Adams NM, Hand DJ, Whitrow C, Weston DJ. Off-the-Peg and Bespoke Classifiers for Fraud Detection. *Computational Statistics and Data Analysis* 2008; **52**: 4521-4532.

[35] Kaminski KA, Wetzel TS, Guan L. Can Financial Ratios Detect Fraudulent Financial Reporting. *Managerial Auditing Journal* 2004; **19**: 15-28.

[36] Kapardis MK, Christodoulou C, Agathocleous M. Neural Networks: The Panacea in Fraud Detection? *Managerial Auditing Journal* 2010; **25**: 659-678.

[37] Kiehl TR, Hoogs BK, LaComb CA. Evolving Multi-Variate Time-Series Patterns for the Discrimination of Fraudulent Financial Filings. *In Proc. of Genetic and Evolutionary Computation Conference* 2005.

[38] Kirkos E, Spathis C, Manolopoulos Y. Data Mining Techniques for the Detection of Fraudulent Financial Statements. *Expert Systems with Applications* 2007; **32**: 995–1003.

[39] Krivko M. A Hybrid Model for Plastic Card Fraud Detection Systems. Expert Systems with Applications 2010; **37**: 6070-6076.

[40] Kotsiantis S, Koumanakos E, Tzelepis D, Tampakas V. Forecasting Fraudulent Financial Statements Using Data Mining. *International Journal of Computational Intelligence* 2006; **3**: 104–110.

[41] Lenard MJ, Watkins AL, Alam P. Effective Use of Integrated Decision Making: An Advanced Technology Model for Evaluating Fraud in Service-Based Computer and Technology Firms. *The Journal of Emerging Technologies in Accounting* 2007; **4**: 123-137.

[42] Li SH, Yen DC, Lu WH, Wang C. Identifying the Signs of Fraudulent Accounts Using Data Mining Techniques. *Computers in Human Behavior* 2012*;* **28**: 1002–1013.

[43] Liao SH, Chu PH, Hsiao PY. Data Mining Techniques and Applications – A Decade Review from 2000 to 2011. *Expert Systems with Applications* 2012; **39**: 11303–11311.

[44] Liou FM. Fraudulent Financial Reporting Detection and Business Failure Prediction Models: A Comparison. *Managerial Auditing Journal* 2008; **23**: 650-662.

[45] Lin CC, Chiu AA, Huang SY, Yen DC. Detecting the Financial Statement Fraud: The Analysis of the Differences between Data Mining Techniques and Experts' Judgments. *Knowledge-Based Systems* 2015; **89**: 459-470.

[46] Lin JW, Hwang MI, Becker JD. A Fuzzy Neural Network for Assessing the Risk of Fraudulent Financial Reporting. *Managerial Auditing Journal* 2003; **18**: 657-665.

[47] Little B, Rejesus R, Schucking M, Harris R. Benford's Law, Data mining, and Financial Fraud: A Case Study in New York State Medicaid Data. *Data Mining IX: Data Mining, Protection, Detection and Other Security Technologies* 2008; **40**:195-204.

[48] Mahmoudi N, Duman E. Detecting Credit Card Fraud by Modified Fisher Discriminant Analysis. *Expert Systems with Applications* 2015; **42**: 2510–2516.

[49] Fox News. Bank of America pays $16.5 bn to settle financial fraud case. Retrieved from Fox News Latino: http://latino.foxnews.com/latino/news/2014/08/21/bank-america-pays-165-bn-to-settle-financial-fraud-case/, 2014.

[50] Ngai E, Hu Y, Wong Y, Chen Y, Sun X. The Application of Data Mining Techniques in Financial Fraud Detection: A Classification Framework and an Academic Review of Literature. *Decision Support Systems* 2011; **50**: 559–569.

[51] Olszewski D. Fraud Detection Using Self-Organizing Map Visualizing the User Profiles. *Knowledge-Based Systems* 2014; **70**: 324–334.

[52] Pai PF, Hsu MF, Wang MC. A Support Vector Machine-Based Model for Detecting Top Management Fraud. *Knowledge-Based Systems* 2011; **24**: 314–321.

[53] Panigrahi S, Kundu A, Sural S, Majumdar A. Credit Card Fraud Detection: A Fusion Approach Using Dempster–Shafer Theory and Bayesian Learning. *Information Fusion* 2009; **10**: 354–363.

[54] Pathak J, Vidyarthi N, Summers SL. A Fuzzy-Based Algorithm for Auditors to Detect Elements of Fraud in Settled Insurance Claims. *Managerial Auditing Journal* 2005; **20**: 632–644.
[55] Perols J. Financial Statement Fraud Detection: An Analysis of Statistical and Machine Learning Algorithms. *Auditing: A Journal of Practice & Theory* 2011; **30**: 19-50.
[56] Phua C, Alahakoon D, Lee V. Minority Report in Fraud Detection: Classification of Skewed Data. *ACM SIGKDD Explorations Newsletter* 2004; **6**: 50-59.
[57] Pinquet J, Ayuso M, Guill'en M. Selection Bias and Auditing Policies for Insurance Claims. *The Journal of Risk and Insurance* 2007; **74**: 425-440.
[58] Quah JT, Sriganesh M. Real-Time Credit Card Fraud Detection Using Computational Intelligence. *Expert Systems with Applications* 2008; **35**: 1721-1732.
[59] Ravisankar P, Ravi V, Rao GR, Bose I. Detection of Financial Statement Fraud and Feature Selection Using Data Mining Techniques. *Decision Support Systems* 2011; **50**: 491–500.
[60] Sabau AS. Survey of clustering based financial fraud detection research. *Informatica Economica* 2012; **16**: 110.
[61] Sahin Y, Bulkan S, Duman E. A Cost-Sensitive Decision Tree Approach for Fraud Detection. *Expert Systems with Applications* 2013; **40**: 5916–5923.
[62] Sánchez D, Vila M, Cerda L, Serrano J. Association Rules Applied to Credit Card Fraud Detection. *Expert Systems with Applications* 2009; **36**: 3630–3640.
[63] Seeja KR, Zareapoor M. FraudMiner: A Novel Credit Card Fraud Detection Model Based on Frequent Itemset Mining. *The Scientific World Journal* 2014; 1-10.
[64] Skillicorn D, Purda L. Detecting Fraud in Financial Reports. *European Intelligence and Security Informatics Conference* 2012; 7-13.
[65] Stockburger DW. Discriminant Function Analysis. In *Multivariate Statistics: Concepts, Models, and Applications* 1998.
[66] Tasoulis D, Adams NM, Weston DJ, Hand DJ. Mining Information from Plastic Card Transaction Streams. *In Proceedings in Computational Statistics: 18th Symposium (COMPSTAT 2008)* 2008; **2**: 315-322.
[67] Thiprungsri S, Vasarhelyi MA. Cluster Analysis for Anomaly Detection in Accounting Data: An Audit Approach. *The International Journal of Digital Accounting Research* 2011; **11**: 69-84.
[68] Viaene S, Derrig R, Dedene G. A Case Study of Applying Boosting Naive Bayes to Claim Fraud Diagnosis. *IEEE Transactions on Knowledge and Data Engineering* 2004; **16**: 612 - 620.
[69] Viaene S, Ayuso M, Guillen M, Gheel D, Dedene G. Strategies for Detecting Fraudulent Claims in the Automobile Insurance Industry. *European Journal of Operational Research* 2007, **176**: 565–583.
[70] Viaenea S, Dedene G, Derrig R. Auto Claim Fraud Detection Using Bayesian Learning Neural Networks. *Expert Systems with Applications* 2005; **29**: 653–666.
[71] Virdhagriswaran S, Dakin G. Camouflaged Fraud Detection in Domains with Complex Relationships. *In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 2006; 941-947.
[72] Whiting DG, Hansen JV, Mcdonald JB, Albrecht C, Albrecht WS. Machine Learning Methods for Detecting Patterns of Management Fraud. *Computational Intelligence* 2012; **28**: 505-527.
[73] Wu J, Xiong H, Chen J. COG: Local Decomposition for Rare Class Analysis. *Data Mining and Knowledge Discovery* 2010; **20**: 191-220.
[74] Yang S, Wei L. Detecting Money Laundering Using Filtering Techniques: A Multiple-Criteria Index. *Journal of Economic Policy Reform* 2010; **13**: 159–178.
[75] Yang W, Hwang S. A Process-Mining Framework for the Detection of Healthcare Fraud and Abuse. *Expert Systems with Applications* 2006; **31**: 56–68.

[76] Yeh I, Lien C. The Comparisons of Data Mining Techniques for the Predictive Accuracy of Probability of Default of Credit Card Clients. *Expert Systems with Applications* 2009; **36**: 2473–2480.
[77] Yuan J, Yuan C, Deng Y, Yuan C. The Effects of Manager Compensation and Market Competition on Financial Fraud in Public Companies: An Empirical Study in China. *International Journal of Management* 2008; **25**: 322–335.
[78] Zhou W, Kapoor G. Detecting Evolutionary Financial Statement Fraud. *Decision Support Systems* 2011; **50**: 570–575.

Mousa Albashrawi
Department of Accounting & Management Information Systems,
King Fahd University of Petroleum and Minerals,
Dhahran, Saudi Arabia
Department of Operations & Information Systems,
University of Massachusetts Lowell, USA