Detecting anomalies in financial statements using machine learning algorithm

The case of Vietnamese listed firms

Mark Lokanan and Vincent Tran
Faculty of Management, Royal Roads University, Victoria, Canada, and
Nam Hoai Vuong
Faculty of International Economics, Foreign Trade University, Hanoi, Vietnam

Received 22 September 2018
Revised 29 May 2019
Accepted 15 June 2019

Abstract

Purpose – The purpose of this paper is to evaluate the possibility of rating the credit worthiness of a firm's quarterly financial report using a dynamic anomaly detection method.
Design/methodology/approach – The study uses a data set containing financial statements from Quarter 1 – 2001 to Quarter 4 – 2016 of 937 Vietnamese listed firms. In sum, 24 fundamental financial indices are chosen as control variables. The study employs the Mahalanobis distance to measure the proximity of each data point from the centroid of the distribution to point out the extent of the anomaly.
Findings – The finding shows that the model is capable of ranking quarterly financial reports in terms of credit worthiness. The execution of the model on all observations also revealed that most financial statements of Vietnamese listed firms are trustworthy, while almost a quarter of them are highly anomalous and questionable.
Research limitations/implications – The study faces several limitations, including the availability of genuine accounting data from stock exchanges, the strong assumptions of a simple statistical distribution, the restricted timeframe of financial data and the sensitivity of the thresholds for anomaly levels.
Practical implications – The study opens an avenue for ordinary users of financial information to process the data and question the validity of the numbers presented by listed firms. Furthermore, if fraud information is available, similar research can be conducted to examine the tendency for companies with anomalous financial reports to commit fraud.
Originality/value – This is the first paper of its kind that attempts to build an anomaly detection model for Vietnamese listed companies.

Keywords Investors, Fraud

Paper type Research paper

1. Introduction

The vast amount of data and the increasing development in technology in recent years have changed the way in which many industries operate and compete with each other. Millions of bytes, commonly referred to as big data, provide valuable insights for companies to make informed business decisions. Companies that conduct business in the financial service sector employ big data to inform their investment practices and make strategic decisions. The increased use and complexity of big data poses a challenge to users of financial information when analyzing financial statements. This is especially applicable to users who possess fewer financial resources and have inferior knowledge to conduct in-depth analysis of financial statements (Lokanan, 2014). Companies that wants to present a rosy picture

Asian Journal of Accounting Research
Vol. 4 No. 2, 2019
pp. 181-201
Emerald Publishing Limited
2443-4175
DOI 10.1108/AJAR-09-2018-0032

of their financial position may exploit these users' deficiencies through deliberate misleading and omission of financial data in their annual reports (Rezaee, 2002; Albrecht *et al.*, 2006; 2014; Robinson and Lokanan, 2017).

Vietnamese companies were selected because of the high incidence of financial reports manipulation (Tran, 2013). The number of listed companies reported by Hanoi Stock Exchange (HNX) and Ho Chi Minh City Stock Exchange (HOSE) from 2000, when Vietnam's security market was in its infancy stage, has steadily increased till 2016. In 2016, there were more than 1,000 listed companies on these exchanges. Growth and structural development in Vietnam's financial markets comes with intense competition in the marketplace and the possibility of financial statement manipulation of listed companies on the HNX and HOSE (Tran, 2013). Indeed, there has been an increasing number of failed companies and fraudulent reporting in Vietnamese markets in the last few years: to be specific, 6,608 companies in the first seven months of 2017, 12,478 companies in 2016 and 9,467 companies in 2015 (Agency of Business Registration – Ministry of Planning and Investment, 2018). The volume and intensity of fraudulent reporting have made it difficult for humans to process and analyze anomalous transactions (Grace *et al.*, 2017). Even some traditional statistic regression techniques cannot be applied due to the complexity of data set (Fan and Li, 2006). Thus, we need embedded analytical models with highly-automated operating structures to deal with the large volume, variety of features and velocity of the data that the human brain cannot handle.

This is where big data techniques come into play. Big data have brought with it novel techniques, such as machine learning and algorithms, which allow users to conduct in-depth analysis and gain deeper understanding of anomalies in financial statements. The analysis of big data using machine learning techniques can assist users of financial statements to detect unusual patterns and transactions in companies' financials. Big data are massive and can be used by both users and companies to provide data-centric and data-driven insights on financial statement anomalies.

This study is an attempt to use machine learning algorithms to detect anomalies in financial statements in Vietnamese listed firms. As mentioned, the only resources available to ordinary investors are quarterly reports, which may contain misleading financial information. It is not enough just to look at the original state of such financial reports. Much research has proved efficiency by analyzing financial ratios calculated from the values in companies' reports (see Altman, 1968; Kotsiantis *et al.*, 2006). Therefore, we approached the problem by using financial ratios as a series of variables, also known as features. An important point in this paper is that the values of financial ratios are assumed to follow a multivariate distribution, which means each ratio varies around one specific mean value. This assumption will allow us to point out anomalous data by measuring whether the distance of each datum to the "centroid" (which will be explained in Research Methodology) exceeds a certain threshold. Additionally, we will take the concept of distance further by regarding it as the degree or extent of the anomaly. This extension of understanding enables us to rank the credit worthiness of each company in each quarter: the more anomalous a datum, the less credit-worthy it is. Therefore, the central question of this paper is as follows: is it possible to rate the credit worthiness of a firm's financial quarter using an anomaly detection method?

It is also worth noting that, up to this point, we have used the term "anomaly" instead of "fraud" for the main theme of this paper. There is a slight difference between the two terms: "anomaly" generally represents "an unusual and possibly erroneous observation that does not follow the general pattern of a drawn population" (Morozov, 2016, p. 63), while "fraud" is an intentional deceptive action perpetuated against a firm for financial gains (Lokanan, 2015). Because we did not have enough data about fraudulent companies or illegal activities in Vietnam, we chose to use "anomaly" for simplicity and precision. However, we still looked at the fraud detection literature, as it points to a more expansive understanding of existing analytical techniques.

The rest of the paper proceeds according to the following format. We first present a comprehensive review of financial fraud detection research using machine learning methods. In this regard, we provide an analysis of the existing fraud detection literature based on the most important machine learning algorithms and statistical methods employed in the literature to date. Next, we outline the methodology and research design used to collect and analyze the data. This is followed by an analysis of the empirical findings. Finally, we present a conclusion and highlight some of the key issues associated with current practices and highlight areas for future research.

2. Contribution to theory and practice

Theoretically, the paper provides guidance on the machine learning technique and algorithms to use when creating new models for detecting anomalies in financial statements. Statistical methods have been the go-to method to evaluate information from financial statement reports (Agarwal and Taffler, 2008; Altman *et al.*, 2013; Tinoco and Wilson, 2013; Lokanan, 2017). While a statistical method has been a success in detecting anomalies in prior research (Beneish, 1997; Bell and Carcello, 2000; Lyandres and Zhdanov, 2013), machine learning techniques have proven to be just as or even more effective in classification performance (Feroz *et al.*, 2000; Lin *et al.*, 2003; Kotsiantis *et al.*, 2006; Perols, 2011) because they make for easier processing of large data (Fan and Li, 2006) and have proven to be neutral in decision making (West *et al.*, 2005; Kotsiantis *et al.*, 2006). Traditional statistical analysis produces errors terms when there are many observations and the use of longer time series data into the models (Feroz *et al.*, 2000; Lin *et al.*, 2003; Lokanan and Sharma, 2018).

Practically, machine learning, with its many features, can allow users to handle large datasets and improve upon statistical models (Kotsiantis *et al.*, 2006; Perols, 2011). When approaching a problem through traditional statistical methods, the human mind can come up with many hypotheses and generate false prediction toward existing information because hypotheses are never entirely accurate (Farber, 2005; Purda and Skillicorn, 2015). The use of a machine learning tool will reduce the number of hypothesis tests in the calculation by using only primary input data. In this regard, machine learning algometric technique will address the shortcomings of hypothesis bias in traditional statistical data analytics (Perols, 2011).

The results presented in this paper are one step closer to heed the calls for the acceleration of technology in analyzing financial statement data and lay the groundwork for further research on the automation of fraud detection (Kirkos *et al.*, 2007; Song *et al.*, 2014; Lokanan, 2015). It is expected that the models employed in this paper will aid investors and other users of financial information to use financial data, even the most technical ones, to conduct informed analysis of companies' financial performance. The models are built from algorithms and a very large data set, which, together, can inform users' intelligence about red flags of fraud in financial statements.

3. Prior research

3.1 Financial statement irregularities

Over the past three decades, there has been an increased focus on irregularities in corporate accounting reporting in general and financial statement fraud in particular (Beasley, 1996; Beneish, 1997; 1999; Rezaee, 2005; Hogan *et al.*, 2008; Cooper *et al.*, 2013; Lokanan, 2015; Morales *et al.*, 2014). Generally, the literature on financial statement fraud focuses on the individual factors that affect fraudulent behavior in organizations (Albrecht *et al.*, 2004; Bell and Carcello, 2000; Rezaee, 2005; Dellaportas, 2013); the procedures and expertise of auditors to detect "red flags" of fraud (Albrecht and Albrecht, 2004; Rezaee, 2005; Murphy and Dacin, 2011; Murphy, 2012; Power, 2013; Morales *et al.*, 2014); the effects of

fraud risks assessment tools on high risks areas in audit engagements (Johnstone and Bedard, 2001; Rezaee, 2005; Davis and Pesch, 2013; Power, 2013; Lokanan, 2015; Behzadian and Izadi Nia, 2017); and the role of auditing committees to detect red flags associated with fraud (Johnstone and Bedard, 2001; Kranacher *et al.*, 2010; Lokanan, 2014). Together, the academic research offers insights on financial statement fraud and facilitates the development and enhancement of new technologies to detect anomalies in fraud (Hogan *et al.*, 2008; Albrecht *et al.*, 2015; Morales *et al.*, 2014).

3.2 Detecting anomalies in financial statements using machine learning

Financial statement fraud is an issue with far reaching consequences (Rezaee, 2005; Albrecht *et al.*, 2015; Lokanan, 2015). Traditional methods involving manual detection, while successful in certain areas (Eining *et al.*, 1997; Farber, 2005; Beneish *et al.*, 2013; Hajek and Henriques, 2017; Lokanan, 2017), are not only time-consuming and expensive, but, in the age of big data, they are also impractical and unable to deal with large volumes of unstructured data (Perols, 2011). Largely driven by an increase in instrumentation, the financial service industry has turned to automated processes using statistical and computational methods to analyze financial statements data (Anandakrishnan *et al.*, 2017). Machine learning algorithms are not only useful in dealing with big data, they can also mimic how users process unstructured data, text, speech and image, to improve accuracy in interpreting financial statements (Feroz *et al.*, 2000; Beneish and Craig, 2007; Song *et al.*, 2014).

Research that evaluates the effectiveness between machine learning and traditional statistical methods typically compares fraud classification algorithms with regression models (Green and Choi, 1997; Lin *et al.*, 2003; Kirkos *et al.*, 2007; Perols, 2011; Morozov, 2016; Lokanan and Sharma, 2018). This stream of research employed logistic regression, artificial neural networks (ANN), fuzzy logic and ensemble-based methods, and found that the techniques combined are useful for fraud detection even when fraud cases are rare or unavailable. The distinguishing elements that make this stream of research unique, however, is that there are many more companies that prepare accurate financial statements than those that falsify their financial statements (Perols, 2011). The attributes (i.e. financial ratios) used to classify the fraud are noisy, thereby making companies that falsify their financial statement look similar to companies with accurate and clean financial statements (Lin *et al.*, 2003; Kotsiantis *et al.*, 2006; Kirkos *et al.*, 2007; Purda and Skillicorn, 2015).

Research using logistic regression models has found it to be rather unique when comparing fraud and non-fraud firms (Lin *et al.*, 2003; Kirkos *et al.*, 2007; Hajek and Henriques, 2017; Lokanan and Sharma, 2018). Bell and Carcello (2000) employed a logistic regression model that estimates the likelihood of fraudulent financial reporting. Using a sample of 77 fraud engagements and 305 non-fraud engagements, Bell and Carcello (2000) found that their logistic regression model was significantly more accurate than practicing auditors in assessing the risks of fraud for the 77 observations. In another study, Lokanan and Sharma (2018) employed a logistic regression model to test for red flags of fraud in banks that were involved in the LIBOR scandal. Using financial ratios and corporate governance data relating to the sixteen banks that were involved in the LIBOR scandal with a matched sample of non-fraud banks, the authors found supports for using financial ratios and governance data to detect fraud in banks (Lokanan and Sharma, 2018). In a similar study, Skousen *et al.* (2009) employed a logistic regression model to detect financial statement fraud between a set of fraud firms and a matched sample of non-fraud firms. The study revealed that the logistic regression model was effective in predicting which of the sample firms committed fraud vs those that did not. Likewise, Spathis (2002) found that multivariate logistic regression techniques were accurate in detecting false financial statements, using a sample of fraud and non-fraud firms. In another study, Lin *et al.* (2003)

found that fuzzy neural network outperformed logistic regression model and ANN in the prediction of fraud cases. Hajek and Henriques (2017) also found that logistic regression was also useful in detection of financial statement fraud.

Another stream of research focusing on evaluating the classification of machine learning algorithms in detecting fraud in financial statement typically used different variations of ANN (e.g. Eining *et al.*, 1997; Green and Choi, 1997; Fanning and Cogger, 1998; Lin *et al.*, 2003). Green and Choi (1997) showed that there is support for ANN as a fraud-risk assessment tool. Other studies found a high probability of detecting fraudulent financial statements using ANN rather than probit or logit models (Eining *et al.*, 1997; Fanning and Cogger, 1998). Feroz *et al.* (2000) illustrated the application of ANN to test the ability of selected Statement of Auditing Standards No. 53 to predict the targets of the Securities and Exchange Commission's (SEC) investigations and found that an analysis of financial ratios from the trial balance does have predicted value. Harymawan and Nurillah (2017) employed a multiple regression model to test for earnings management in financial reporting and found that corporate reputation has a significant relationship with earnings quality. These studies reinforced the efficiency for using machine learning algorithms as suggested techniques to detect anomalies in financial statements.

Other research looked at classification algorithms to improve fraud classification performance (Kotsiantis *et al.*, 2006; Kirkos *et al.*, 2007; Perols, 2011). These studies explored the effectiveness of machine learning algorithms to detect firms that issue fraudulent financial statements through various algometric classifications. Kotsiantis *et al.* (2006) employed a sample of fraud and non-fraud firms and financial ratios to examine the following classification algorithms: S, K2, C4.5, 3NN, RBF, Ripper, LR and SMO. The findings revealed that the algorithmic classifications performed better than logistic regression and ANN models. Kirkos *et al.* (2007) classified algorithms into Decision Trees, ANN and Bayesian Belief, and examined their usefulness to detect fraud in financial statements. Using financial statement ratios, the study employed a sample of fraud firms with a matched sample of non-fraud firms and found that the Bayesian Belief outperforms Decision Trees and ANN in financial statement fraud detection. In another study, Hoogs *et al.* (2007) presented a genetic algorithm approach to detecting financial statement fraud. Using a sample of fraud companies accused of improper revenue recognition by the SEC and a matched sample of non-fraud companies, the study found that genetic classification of algorithms has many features well-suited for accurate fraud detection. More recently, Dbouk and Zaarour (2017) employed a Bayesian Naïve Classifier (BNC), a supervised machine learning approach, and found that the BNC's approach outperforms conventional audit method in detecting earnings manipulations.

Machine learning research has developed ensembles of predictors other than ANN and logistic regression to generate hypothesis for testing in fraud research (Perols, 2011; Phua *et al.*, 2001). The research in this category used cluster algorithm (i.e. *K*-means clustering) (Li, 2016), bagging (West *et al.*, 2005) and support vector machine (SVM) (Shin *et al.*, 2005) to assess anomalies in financial fraud. Of particular interest with this stream of research is that they used ensemble methodology to show that anomaly detection and predictive analytics for financial risk management bring out the ideas of using some algorithms together. In general, ensemble predictors were found to be superior to single machine learning and statistical models for detecting fraud in financial statements (Phua *et al.*, 2004; West *et al.*, 2005). Ensemble predictors are also able to extract optimal solutions with small and noisy data set (Fan and Palaniswami, 2000; Shin *et al.*, 2005).

The foregoing literature review highlighted the various statistical techniques and machine learning models used to identify anomalies in financial statements. With respect to fraud detection, there is a significant body of research that provides support for machine

learning and, to a certain extent, logistic regression models (Chen and Rezaee, 2012; Albrecht *et al.*, 2015). Overall, ANN outperforms logistic regression models in the literature; however, both were found to be inferior when compared to ensemble-based and classification algorithms methods. Despite this authoritative guidance on these statistical models and machine learning techniques, there remains a significant gap between fraud detection models and their application to large volume of time series and cross-sectional data. This study is an attempt to address these gaps by using machine learning techniques to detect red flags of fraud on a sample of Vietnamese companies.

4. Research methodology

4.1 Data source and collection

In this study, we use financial statement ratios to build the algorithms. The financial ratios, which are divided into seven groups, are obtained from Cophieu68 and Vietstock. These sources contain quarterly and annual financial reports of all listed companies on the Vietnamese stock market from 2011 to 2016. We used data for this period because it was the period when the stock exchange in Vietnam had the largest volume of readily available data (i.e. not too many missing data). The data used in this study contain both audited and unaudited quarterly reports. Quarterly reports are not legally required to be audited in Vietnam. Unaudited raw data have the advantage of showing the earliest anomalous situation in financial statements.

Data were collected from the most reliable sources available in Vietnam: income and cash flow statements from Cophieu68.vn and balance sheets from Vietstock finance. After having excluded banks, financial, insurance companies as well as recently merged or acquired firms, we obtained a total of 937 listed Vietnamese firms. Each document for a company was stored in a matrix-like data structure whose columns are the indices and rows are observations. We chose to conduct anomaly detection on a quarterly basis as we wanted the result to eliminate the large possible time lag. Also, audited information could be booked, thus covering the real financial situation of the companies. The timeframe of the data is from Quarter 1 – 2011 to Quarter 4 – 2016, which spanned 24 quarters. The data values in each quarter represent an observation. We were able to extract data from 1,090 companies listed on Vietnam's stock exchanges. However, 153 financial institutions were eliminated from the sample due to their unique form of financial statements and business operation. The final data set consisted of 937 companies and 22,488 observations.

4.2 Data Pre-processing

4.2.1 Financial indices calculation. We identified that there were 31 essential financial indices. However, we excluded seven indices because of high correlation, whose thresholds were greater than 0.8, to avoid multi- collinearity issues. Thus, we obtained 24 indices that can be considered independent. Every index in the seven categories listed below has different implications for detecting financial anomalies:

(1) liquidity ratios: used to determine how quickly a company can turn its assets into cash if it is experiencing financial distress or impending bankruptcy;

(2) profitability ratios: are ratios that demonstrate how profitable a company is;

(3) activity ratios: are meant to show how well management is managing the company's resources;

(4) solvency ratios: depict how much a company relies upon its debt to fund operations;

(5) market ratios: measure investors' response to owning a company's stock and the cost of issuing stock;

(6) accrued income: is earned in a fund or by a company for providing a service or selling a product that has yet to be received; and

(7) cash flow: is the net amount of cash and cash-equivalents moving into and out of the business.

The general overview of data collection process to obtain financial indices is illustrated in Figure 1.

4.2.2 Data normalization. Due to different sizes of the companies, their financial indices are on various scales. Without normalizing, the model will be biased. In order to proceed, it is necessary to mention vector operations on the matrix-like data structure. In Figure 1, each row is considered a vector. A vector can contain one or many components, and, in this particular case, there are 24 components, which are the financial indices obtained from the data collection process. Thus, when we say we conduct a vector operation on two vectors, such as summing two vectors, we are just summing each corresponding component of two vectors to create a new vector with the same number of components. Vector operations are powerful and essential to perform gigantic calculations on matrix-like data structure.

In the scope of our research, we chose standardization as the method of normalizing our training and testing data because we assumed that, over time, financial indices of a company would stabilize and follow the standard normal distribution, which later will be denoted as "$\sim N(0,1)$." In the standardization method, we compute the values of each financial index to have zero mean and unit variance. First, we calculated the mean and standard deviation of the range of feature values. Next, we subtracted the mean from each function's value, then divided the result by the standard deviation. The process is summarized in the following formula:

$$x' = \frac{x-\mu}{\sigma}. \tag{1}$$

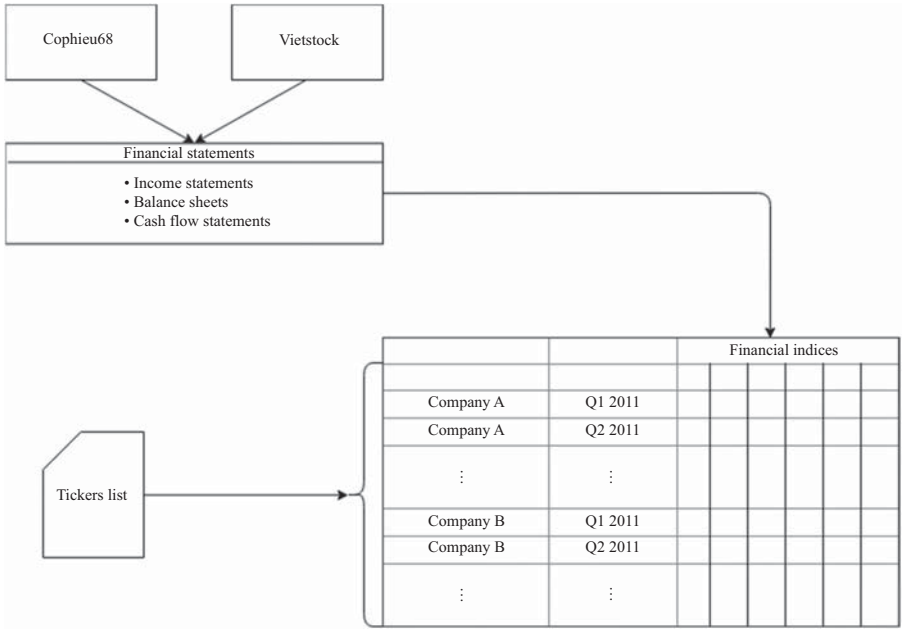

Figure 1.
An abstract overview of data collection process

To implement this, we conduct the following steps:

(1) Obtaining the mean vector (MU) which represents the mean value of all financial indices.

(2) Obtaining the standard deviation vectors (Std) of each company from the training data set, which will be described in detail in the next section.

(3) We conduct a matrix operation according to Equation (1), with x as each observation's vector, μ as the mean vector and as the standard deviation vector of each company. By doing this, we can obtain normalized data values for every observation as x' (Table I).

4.2.3 Pre-implementation. Before implementing the models, we add some fine-tunings for missing data. For a company when there are completely no available data for a financial index, that financial index is removed from the computation. Also, if values of a financial index are partially missing (denoted in the data set as "N/A"), we replace them with values generated from the standard normal distribution. Since we investigate anomalies on a company basis, the procedure for a company does not affect the data values of other companies.

It can be argued that, for every time a missing value is filled with a value from the standard normal distribution, the result will be different. However, we must also be careful not to fill in empty values with a fixed value, such as 0 (a common solution), because the resulting data values will not align with the assumed distribution. A possible solution for this issue is that, for every random value that needs to be filled in, we can set a "random state," which will cause the randomized function to always return the same random number for every run. In doing so, interested readers or researchers can simulate the same implementation to understand the findings and replicate the models in future research. Table II presents a preview of the normalized training data frame.

5. Research design

5.1 Multivariate normal distribution and assumptions

Since the data have 24 independent financial indices, which correspond to 24 features, the multivariate normal distribution (MVN) will be implemented in our model. In general, it is the generalization of the univariate normal distribution to multiple variables (Fan and Palaniswami, 2000; Lokanan, 2017). Although real data may never come from a right MVN, the MVN provides a robust approximation and has many desirable mathematical properties, such as the mean vector and covariance matrix. Furthermore, because of the central limit theorem, many multivariate statistics converge to the MVN distribution as the sample size increases. Overall, MVN has the following properties:

- Joint density.
- Shape: the contours of the joint distribution are n-dimensional ellipsoids.
- Mean, and covariance, specifies the distribution. The MN(μ, Σ) joint distribution is determined by μ and Σ only.

Table I.
A visualization of the mean vectors of the companies

Companies	Financial indices		
	Index 1	Index 2	–
Company A (vector A)	Mean of index 1(A)	Mean of index 2(A)	–
Company B (vector B)	Mean of index 1(B)	Mean of index 2(B)	–

	payables_turnover	cash_ROA	ROA	cash_flow_on_revenue	book_value_per_share	cash_ratio	reinvestment	cash_ROE	net_profit_margin
AAA									
Q1 2011	1.569563	−0.605933	1.465270	−0.532049	0.377707	−1.009985	0.485990	−0.633468	1.589029
Q1 2012	0.403633	0.111815	0.589110	0.143790	1.468297	−0.933268	−0.480826	0.205800	0.486731
Q1 2013	−1.212642	−0.483559	0.668519	−0.405450	−0.120294	0.341740	0.450969	−0.502922	0.983711
Q1 2014	−1.056517	−1.235540	−0.122068	−1.185015	0.063304	−0.552369	1.290759	−1.277581	−0.001672
Q1 2015	−0.854412	−1.615856	−2.420690	−1.920006	−1.299141	0.885708	0.702198	−1.571438	−2.592306
Q2 2011	2.857911	0.130310	−0.135469	0.225825	0.519688	−0.869617	−2.771562	0.230875	−0.022633
Q2 2012	0.555663	0.583030	−0.082492	0.761017	1.445435	−0.763903	−0.319130	0.675693	0.141685
Q2 2013	0.596634	0.005349	0.026183	0.080237	−0.252724	−0.561752	0.377966	−0.065049	0.085655
Q2 2014	−0.565054	0.684313	−0.010287	0.584314	0.122422	−0.485738	−0.110050	0.662314	−0.217191
Q2 2015	−0.369197	−0.538192	−1.328244	−0.468264	−1.269048	1.017459	0.470866	−0.553311	−1.178309
Q3 2011	0.863739	0.260351	1.325117	0.366799	0.803063	−0.539378	0.113013	0.435657	1.526498
Q3 2012	0.229783	1.131976	0.118419	1.066416	1.637285	−0.953882	−1.936275	1.147646	0.012660

Table II.
A preview of normalized training data frame

- Moment generating function: the MN (μ, Σ) distribution:

$$MGFM(t) = \exp\left(\mu^T t + \frac{1}{2} t^T \sum t\right), \tag{2}$$

where t is a real $n \times 1$ vector.

- Characteristic function: the MN (μ, Σ) distribution has:

$$CF\psi(t) = \exp\left(\mu^T t - \frac{1}{2} t^T \sum t\right), \tag{3}$$

where t is a real $n \times 1$ vector.

- Linear combinations.
- Independence.

With MVN, the following assumptions are made regarding our research case:

- Indices after calculation and scaling are Gaussian independently distributed. If not, they are either transformed or omitted from the model.
- Some financial indices are too important to be overlooked. As such, we retained them even if they were highly correlated with the others.
- In sum, 95 percent of "none - anomalous" data points stay within $[\mu-3\sigma; \mu+3\sigma]$.
- Company indices change slowly and can be inferred from history and other indices.

5.2 Mahalanobis distance – statistical distance

One of the simplest methods to filter anomalous observations is the Mahalanobis distance (Thongkam *et al.*, 2008). The Mahalanobis distance will be employed in this study to decide whether an observation is anomalous. Conceptually, the Mahalanobis distance measures the proximity of a data point to the center of the distribution and is a direct generalization of standard deviation. The Mahalanobis distance of a point $x = (x_1, x_2, \ldots, x_n)$ is defined as follows:

$$d(x, \mu) = \sqrt{(x-\mu)^T \sum (x-\mu)^{-1}}, \tag{4}$$

In Equation (2), $\mu = (\mu_1, \mu_2, \ldots, \mu_n)$ is the mean vector of the distribution, and $\sum$ is the covariance matrix of the features in the n-dimensional space. For the application of anomaly detection, we will measure the distance of each test observation to the mean vector (μ) of each company's data in standard deviation unit. The formula is computed as follows:

$$d_M(x, \mu) = \sqrt{(x-\mu)^T \sum (x-\mu)^{-1}}, \tag{5}$$

It is assumed that each feature (financial index) ~$N(0, 1)$, each observation vector, $x \sim N_\rho(\mu, \Sigma)$ with $\Sigma \vee 0$ and ρ is the number of features. Therefore, the random variable $D = (x-\mu)^T \Sigma (x-\mu)^{-1}$ has the χ^2 distribution with ρ degrees of freedom (Bajorski, 2011). From this property, it is inferred that:

$$P(d_M(x, \mu) \leqslant k) = G_\rho\left(k^2\right), \tag{6}$$

where G_ρ is the cumulative distribution function (CDF) of the χ^2 distribution with p degrees of freedom.

By choosing the value of k and referring to the χ^2 distribution table, we can obtain the probability of an observation having its distance to the mean vector less than k standard deviations. This probability is also the proportion of data observations having their distances to the mean vector less than k standard deviations. The application for anomaly detection, which will be discussed in the next part, is mostly based on the conclusion above.

5.3 Anomalies detection with Mahalanobis distance

After collecting data, a correlation matrix was formulated for each company. This approach allows us to measure the correlation between the companies' indices. We also categorized the original data set into two parts: 83 percent for training and 17 percent for testing. Table III shows that the one-dimension (1D) tensors or vectors are represented for calculating Mahalanobis distance:

$$f_x(x_1, x_2, \ldots, x_k) = \frac{1}{\sqrt{(2k)^k |\sum|}} \exp. \tag{7}$$

This function can be generalized for the vector of 1 row × 24 columns (1×24). However, to visualize the multivariate Gaussian distribution, we will use a vector of size 1×2, as shown in Table IV.

The visualization of a 1D tensor with the size of 1×2 example can be seen in Figure 2.

The contour of this bivariate normal distribution is visualized in Figure 3. Importantly, when we squash three-dimension data points into two-dimension ones, the data set will lose

Table III.
Input: 1-D tensors ~ vectors as below for calculating

Company A's current ratio	Company A's quick ratio	–	Company A's total net accruals

Note: [Current ratio, quick ratio, …, total net accrual]

Table IV.
1×2 matrix for better visualization on three-dimension space

Company A's current ratio	Company A's quick ratio

Note: [Current ratio, quick ratio]

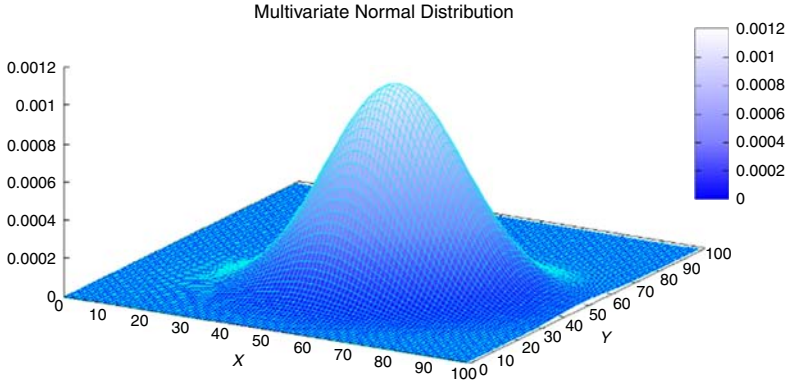

Figure 2.
The visualization of a 1D tensor with size 1×2

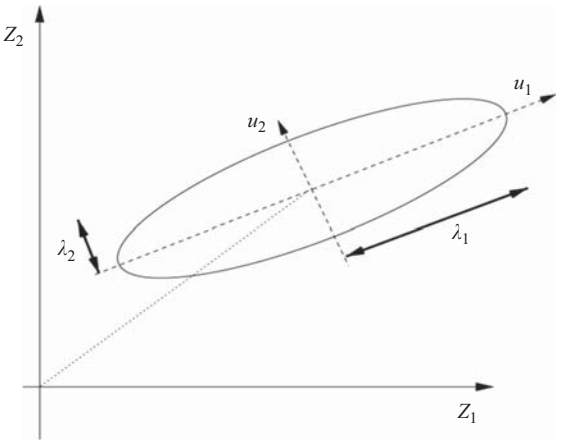

Figure 3.
2D visualization of 1D tensor with size 1×2

valuable information. As we can see in the function and visualization in Figure 3, the output of the MVN is a learned model from the training data set, while $f(X)$ is the probability of whether a data point is part of the normal distribution. We have to set a value of epsilon (ε) to compare with $f(X)$. If $f(X) < \varepsilon$, the data point is anomalous, and vice versa.

However, because of computational complexity, we will not deploy calculations on $f(X)$ directly. Instead, we deal with MVN in another way: given the assumption of the central limit theorem, we will calculate the loss, the Mahalanobis distance, as the product between test data point and the mean. If the loss $\geqslant 3\sigma$, we can conclude the level of anomalies is high, and vice versa. As mentioned above, we can calculate the Mahalanobis distance for each observation. If any of these distances is greater than a certain threshold L_{max}, we consider that observation an anomaly. To calculate L_{max}, we check how likely it is that the most significant Mahalanobis distance is greater than L_{max} using the following equation:

$$Y_i = (X_i - \mu)^T - \sum (X_i - \mu)^{-1},\ i = 1, ..., n. \tag{8}$$

From the section Mahalanobis distance – statistical distance, each Y_i follows the χ^2 distribution with ρ features – ρ degrees of freedom. Now, we can calculate the probability that the largest Y is larger than L^2 using the following equation:

$$\begin{aligned} P\left(\max_{1 \leqslant i \leqslant n} Y_i > L^2\right) &= 1 - P\left(\max_{1 \leqslant i \leqslant n} Y_i \leqslant L^2\right) = 1 - \prod_{i=1}^{n} P\left(Y_i \leqslant L^2\right) \\ &= 1 - \left[G_p\left(L^2\right)\right]^n, \end{aligned} \tag{9}$$

where G_ρ is the CDF of the χ^2 distribution with ρ degrees of freedom.

For this probability to be equal to a small value, we need:

$$L = \sqrt{\chi^2_\rho\left((1-\alpha)^{1/n}\right)}. \tag{10}$$

In this study, we choose the value to be equal to 5 percent, which is a traditionally preferred value when refining significance level in academics (Torbeck, 2010). Any observation whose distance to the mean vector is greater than this value of $L(\alpha = 0.05)$ will be considered an anomaly. As mentioned earlier, we set the fixed L_{max} with $p = 95$ percent and 24 features as 2.64E+15. If the responding Mahalanobis distance of a specific quarter is greater than

the L_{max} value, we can be sure that the quarters of the specific company's financial indices were anomalous. Furthermore, as mentioned in the Introduction, we consolidate the distances into ordered categories to rank the companies' credit worthiness. The projected anomaly ratings are defined in Table V.

6. Empirical results

Training classification was formed to display the data (Perols, 2011). Out of the 22, 488 observations, the training group consists of 18,740 observations, while the test group consists of 3,748 observations (e.g. see Spathis, 2002; Lin *et al.*, 2003; Kirkos *et al.*, 2007; Lokanan and Sharma, 2018). We first employ the Mahalanobis distance and consider each firm in a quarter as a data point. The summary of the Mahalanobis distance of the testing data set is shown in Table VI. The mean and the standard deviations for the distance from one datum to central limit come from the 3,748 observations. The mean distance measuring average length from firm-quarter data point to the center was 5.65E+20. The maximum and minimum Mahalanobis distances from the testing data set are 3.1E+21 and 1.25E+24, which represent observations with the largest and smallest degree of anomaly, respectively. These findings indicate that, for the most part, a significant proportion of the companies were within the normal (vs "anomaly") standard deviation range in their financial statements. A closer look at the results in Table VI shows that most of the observations were closer to the mean and that there was not much variance (i.e. larger standard deviations) among companies.

For more details, the company ratings defined above can be applied to every observation or firm-quarter in the data set. As an example, Tables VII–IX show the Mahalanobis distance values by quarters for Vietnamese listed companies represented by their respective tickers for 2016. As can be seen in Table VII, for the entire 2016 financial year, Type A companies from stock ticker AAA had a small number of anomalies in their financial statements. These results were predicted as it was expected that companies with A ratings will have small variance in their financial statements. On the other hand, with the exception

Table V. Company rating types for all observed companies

Type	Definition	d_M range
A	Good company with no or few anomalies, predictable outcome	[0; 1E+15]
B	Normal company with average anomaly, quite predictable outcome	(1E+15; 5E+15]
U	Unranked company with many anomalies, unpredictable outcome	(5E+15; +infinity)

Table VI. A summary of Mahalanobis distance

Number of observations	Mean	Std	Min
3,748	5.65E+20	2.49E+22	−3.1E+21
Top 25%	Top 50%	Top 75%	Max
−2.4E+15	0.612693	3.69E+15	1.25E+24

Table VII. Financial anomaly analysis result of stock ticker AAA

Time	Mahalanobis distance	Anomaly rating
First quarter 2016	−3.1E+15	A
Second quarter 2016	3.7E+14	A
Third quarter 2016	2.7E+14	A
Fourth quarter 2016	−1.4E+16	A

of the first quarter, the unranked companies in stock tickers AAM (Table VIII) had a higher number of anomalies with very unpredictable outcomes. The companies in the stock ticker DGH (Table IX) had fluctuating results throughout the 2016 fiscal year. For quarters 2 and 4, there was very little anomaly detection and a strong indication that companies were producing reliable financial statements. In quarter three, the results from the Mahalanobis distance reveal the companies had average anomalies and quite predictable outcome. For the unranked companies in the DHG stock ticker, there were many anomalies with very unpredictable outcome in the first quarter of 2106. Overall, it may not be to a company's advantage to manipulate their financial statements, especially when the company already has a good reputation (Harymawan and Nurillah, 2017).

Table X shows the number of rated firm-quarter. A closer look at Table VI shows that 68.89 percent of firm-quarter data is rated A. This means that most of the companies were performing very well with few anomalies in their financial statements. Considering the concerns regarding fraudulent financial statements in Vietnam, these findings are significant for two reasons. First, it shows that there are no overall financial statement level (OFSL) threats and users can feel confident in using these financial statements to make informed financial decision. In other words, the findings reduced OFSL risk, that is, the risk of material misstatement for the financial statement as a whole (see also Behzadian and Izadi Nia, 2017). Second, and partly synonymous with the first, is that the results directly enhance the relevance and reliability of information provided in the financial statements for strategic investment decision making. The rated B companies account for about 7.6 percent of anomalies. This is a minimal amount compared to 23.51 percent of unranked firm-quarters that have significant anomalies with very unpredictable results. This is a serious concern and Vietnamese regulators need to take stock of these results. That fact that 23.51 percent of unranked companies have significant anomalies in their financial statements raises serious questions concerning the efficiency of the audit approach and procedures used to audit these financial statements (Figure 4).

Table VIII. Financial anomaly analysis result of stock AAM

Time	Mahalanobis distance	Anomaly rating
First quarter2016	3.19E+14	A
Second quarter 2016	2.15E+16	U
Third quarter 2016	1.48E+16	U
Fourth quarter 2016	3.44E+16	U

Table IX. Financial anomaly analysis result of stock ticker DHG

Time	Mahalanobis distance	Anomaly rating
First quarter 2016	1.56E+16	U
Second quarter 2016	5.16E+14	A
Third quarter 2016	2.29E+15	B
Fourth quarter 2016	4.92E+14	A

Table X. The summary of company ratings

Type	Number of companies	Percentage
Number of type A – rated firm-quarter	2,582	68.89
Number of type B – rated firm-quarter	285	7.6
Number of type U – rated firm-quarter	881	23.51

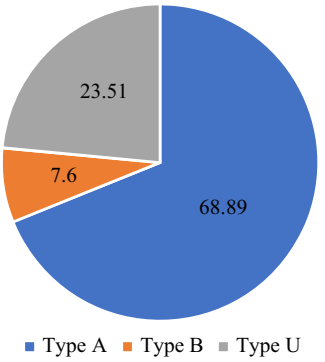

Figure 4.
Visual breakdown of companies in each rating type (in percentages)

7. Conclusion, limitations and areas for future research

The paper produces two meaningful outcomes regarding anomalies detection and anomaly rating of financial statements. First, a significantly high proportion (68.89 percent) of Type A firms had very few anomalies and this is a healthy sign that they are in compliance with legal and ethical standards. For a user's perspective, they can feel confident in using these financial statements to make investment decisions. Second, and more concerning, is that 23.51 percent of the unranked companies had significant anomalies in their financial statements. This is a worrying sign as it indicates that there are risks in the financial statements as a whole and there is a likelihood of fraud or error in the financial statements. In both cases, the first outcome laid a solid foundation for users to analyze and understand financial statements, while the second outcome indicates that, as an aggregate, regulators need to start paying more attention to the audited financial statements of unranked companies for potential fraud or error.

From a practical perspective, the results presented here can be used to provide guidance to audit committees and senior executives concerned about the increased possibility to deter potential accounting fraud or misstatements in financial reporting. Just as there is a need for higher ethics in organizational research, so too is the need for managers to recognize the importance of the fraudulent mechanics of an imperfect system that promotes unethical financial reporting (Lokanan, 2015). If, as seems to be the case, unethical behavior by managers gives a competitive advantage to a company, then monitoring their activities becomes an important strategic objective for firms and external auditors to report them to regulators and publicize the irregularities to all users of their financial information.

People working in the corporate sector regularly spend a lot of time in financial districts, which, despite the best regulatory efforts by regulators, offer regular opportunities to engage in fraud and/or experience sources of friction, which may lead to fraud (Lokanan, 2015; Morales *et al.*, 2014). An accountant, for example, may address the friction caused from not meeting financial targets by manipulating the financial statements (Behzadian and Izadi Nia, 2017; Lokanan and Sharma, 2018). A chief executive officer may address the friction caused from not meeting financial targets by manipulating his compensation bonuses (Perols, 2011). As these individuals repeat their behaviors, questions arise over the materiality of the OFSL risks in these statements. The model employed in this paper is capable of addressing these issues in the specific details of a single company or set of companies. Moreover, the model can detect anomalies in a quarterly format and will give managers and users of the report a more consistent update of the materiality of companies' financial statements. Note also that each company's quarter is ranked with anomaly ratings, which measures how anomalous the company's indices are at a given point in time.

The insights from the anomaly ratings provided in Table V have the potential to play a significant role in understanding companies' financial statements. A financial architecture that can manage OFSL risks will go a long way to protect the public interests and users of financial statements. Dishonest accounting practices could be of interest to some users in struggling companies and, as such, they may entertain the possibility that fraudulent accounting is permissible. However, company managers and regulators needs to be cognizant of the fact that users want accurate and relevant information to make informed strategic decision; as such, anomaly detection models can prove useful in detecting anomalies in financial statements and offset deception, misreporting and the falsification of financial reports. In particular, the anomaly rating can be used to measure the level of accuracy of companies' financials by users of financial statements (e.g. see Fanning and Cogger, 1998; Lyandres and Zhdanov, 2013). These findings provide meaningful insights to financial institutions when determining lending decisions and to investors when evaluating companies' financials to make investment decisions (Bell and Carcello, 2000; Beneish *et al.*, 2013; Behzadian and Izadi Nia, 2017; Lokanan, 2017).

At a more macro level, Vietnam has witnessed accelerated economic changes in its financial regulatory landscape in recent years (Narayan and Zheng, 2010; Phan *et al.*, 2018). With such changes come pressures for companies to compete in one of the strongest growing markets in East Asia (World Bank, 2017; Phan *et al.*, 2018). With such growth, the problem of financial statement manipulation has once again raised its ugly head in financial reporting (Hiep, 2017; Phan *et al.*, 2018).

The results from this study also provide insights for government agencies to control and reduce the degree of financial statement fraud. As mentioned earlier, a very significant 23.51 percent of the testing data set was shown to be anomalous. This result implies that regulators must focus more on enhancing transparency and compliance in financial reporting (Phan *et al.*, 2018). In Vietnam, the Vietnamese Standards on Auditing regulate the procedure for reviewing the quality of financial statements before they are released to the public. From both an audit and a user's point of view, the models employed in this study can provide insights on best practices to improve the accuracy of audited financial statements. As the management may want to manipulate earnings, the use of machine learning algorithms can be employed by regulators, preparers, auditors and users to detect errors in financial reports and mitigate the prevalence of false representations cause by fraudulent reporting (Dbouk and Zaarour, 2017).

Due to the eclectic nature of anomaly detection in financial statements, a general model of outlier detection simply does not exist. As such, the regulatory apparatus must strive to identify models that can offer insights into different types of unethical (and at times fraudulent) behavior in financial reporting. In this regard, the paper advances research in anomaly detection by presenting a model that will assist individuals to analyze complex unethical behavior in accounting manipulation and, at the same time, offer deeper insights into fraud detection in financial statements. More importantly, and especially for the unranked companies, the paper can prove useful to examine managerial intent to act unethically (and sometimes fraudulently) while prioritizing and integrating interests of certain users in the decision-making process.

7.1 Limitations of the model

The machine learning algorithm employed in this paper suffers from several limitations. First, the missing values were treated in the same way in the financial statements. It was biased to do so because the missing values are the result of restriction to private data. In fact, the input data for this research are free and available, which mimics perfectly the insights that can be seen by ordinary stakeholders. This is very important as it not only intensifies human weakness when dealing with big data, but also reveals data imparity and

scarcity in the Vietnamese market. Second, we had to employ statistical assumptions and treat the data as stemming from the normal distribution. The high dimension data require more advanced models, which are far beyond the scope of the data set we are working with (e.g. Fan and Palaniswami, 2000; Spathis, 2002; Shin *et al.*, 2005). Third, the timeframe under study ranged from 2010 to 2016, which is a relatively short interval. Machine learning techniques require data acquired from a longer timeframe to ensure a better fitted model (Hoogs *et al.*, 2007; Perols, 2011). Fourth, the threshold levels of anomalies are difficult to determine. As a result, the model is too sensitive to anomalies. With 23.51 percent of the companies' data points considered to be highly anomalous, it can be said that they were not practical enough when compared to the historical record of the Vietnamese market.

7.2 Future research

The present paper results in two significant findings which can serve as guidelines for further research on machine learning and detecting anomalies in financial statements. First, the data pre-processing methods used in this paper lay a good foundation for future research. Further research can examine how data pre-processing can transform raw data, which will be useful to users of financial information. Second, the result of the anomalous financial analysis can be contextualized in a more meaningful way than just finding out whether they are anomalous or not. One avenue is to pursue further research to see if companies that report anomalous data were cited for fraud or went on to commit fraud in the future.

References

Agarwal, V. and Taffler, R. (2008), "Comparing the performance of market-based and accounting-based bankruptcy prediction models", *Journal of Banking and Finance*, Vol. 32 No. 8, pp. 1541-1551.

Agency of Business Registration – Ministry of Planning and Investment (2018), "Report on Vietnamese market", available at: www.mpi.gov.vn/en/Pages/default.aspx (accessed August 17, 2018).

Albrecht, C., Holland, D., Malagueño, R., Dolan, S. and Tzafrir, S. (2015), "The role of power in financial statement fraud schemes", *Journal of Business Ethics*, Vol. 131 No. 4, pp. 803-813, available at: https://doi.org/doi:10.1007/s10551-013-2019-1

Albrecht, W.S. and Albrecht, C.O. (2004), *Fraud Examination & Prevention*, Thomson/South-Western, OH.

Albrecht, W.S., Albrecht, C.C. and Albrecht, C.O. (2004), "Fraud and corporate executives: agency, stewardship and broken trust", *Journal of Forensic Accounting*, Vol. 5 No. 1, pp. 109-130.

Altman, E. (1968), "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy", *Journal of Finance*, Vol. 23 No. 4, pp. 589-609.

Altman, E., Alessandro, D. and Alberto, F. (2013), "Z-score models' application to Italian companies subject to extraordinary administration", *Journal of Applied Finance*, Vol. 23 No. 1, pp. 1-10.

Anandakrishnan, A., Kumar, S., Statnikov, A., Faruquie, T. and Xu, D. (2017), "Anomaly detection in finance: editors' introduction", *Proceedings of Machine Learning Research*, Vol. 71, pp. 1-7.

Bajorski, P. (2011), "Statistics for imaging, optics, and photonics", available at: www.wiley.com/en-ca/Statistics+for+Imaging%2C+Optics%2C+and+Photonics-p-9780470509456 (accessed September 5, 2018).

Beasley, M.S. (1996), "An empirical analysis of the relation between the board of director composition and financial statement fraud", *The Accounting Review*, Vol. 71 No. 4, pp. 443-465.

Behzadian, F. and Izadi Nia, N. (2017), "An Investigation of expectation gap between independent auditors and users from auditing services related to the quality of auditing services based on their role and professional features", *Asian Journal of Accounting Research*, Vol. 2 No. 2, pp. 36-47.

Bell, T. and Carcello, J. (2000), "A decision aid for assessing the likelihood of fraudulent financial reporting", *Auditing: A Journal of Practice & Theory*, Vol. 19 No. 1, pp. 169-184.

Beneish, M. (1997), "Detecting GAAP violation: implications for assessing earnings management among firms with extreme financial performance", *Journal of Accounting and Public Policy*, Vol. 16 No. 3, pp. 271-309.

Beneish, M. and Craig, N. (2007), "The predictable cost of earnings manipulation", available at: http://dx.doi.org/10.2139/ssrn.1006840 (accessed August 23, 2018).

Beneish, M., Lee, C. and Nichols, D. (2013), "Earnings manipulation and expected returns", *Financial Analysts Journal*, Vol. 69 No. 2, pp. 57-82, available at: https://doi.org/doi.org/10.2469/faj.v69.n2.1

Beneish, M.D. (1999), "The detection of earnings manipulation", *Financial Analysts Journal*, Vol. 55 No. 5, pp. 24-36.

Chen, Y. and Rezaee, Z. (2012), "The role of corporate governance in convergence with IFRS: evidence from China", *International Journal of Accounting & Information Management*, Vol. 20 No. 2, pp. 171-188, available at: https://doi.org/10.1108/18347641211218470

Cooper, D., Dacin, T. and Palmer, D. (2013), "Fraud in accounting, organizations and society: extending the boundaries of research", *Accounting, Organizations and Society*, Vol. 38 Nos 6-7, pp. 440-457.

Davis, J.S. and Pesch, H.L. (2013), "Fraud dynamics and controls in organizations", *Accounting, Organizations and Society*, Vol. 38 No. 6, pp. 469-483.

Dbouk, B. and Zaarour, I. (2017), "Towards a machine learning approach for earnings manipulation detection", *Asian Journal of Business and Accounting*, Vol. 10 No. 2, pp. 215-251, available at: https://pdfs.semanticscholar.org/d9fd/eab8fbd6c697549cc8f84d1284f1ec2d1c9c.pdf (accessed May 22, 2019).

Dellaportas, S. (2013), "Conversations with inmate accountants: motivation, opportunity and the fraud triangle", *Accounting Forum*, Vol. 37 No. 1, pp. 29-39.

Eining, M., Jones, D. and Loebbecke, J. (1997), "Reliance on decision aids: an examination of auditors' assessment of management fraud", *Auditing: A Journal of Practice & Theory*, Vol. 16 No. 2, pp. 1-19.

Fan, A. and Palaniswami, M. (2000), "Selecting bankruptcy predictors using a support vector machine approach", *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000, Neural Computing: New Challenges and Perspectives for the New Millennium*, *Vol. 6*, *Como*, pp. 354-359.

Fan, J. and Li, R. (2006), "Statistical challenges with high dimensionality: feature selection in knowledge discovery", in Sanz-Solé, M., Soria, J., Varona, J.L. and Verdera, J. (Eds), *Proceedings of the International Congress of Mathematicians Madrid*, *August 22–30*, European Mathematical Society Publishing House, Zuerich, pp. 595-622.

Fanning, K.M. and Cogger, K.O. (1998), "Neural network detection of management fraud using published financial data", *Intelligent Systems in Accounting, Finance and Management*, Vol. 7 No. 1, pp. 21-41.

Farber, D. (2005), "Restoring trust after fraud: does corporate governance matter?", *The Accounting Review*, Vol. 80 No. 2, pp. 539-561.

Feroz, E.H., Kwon, T.M., Pastena, V.S. and Park, K. (2000), "The efficacy of red flags in predicting the SEC's targets: an artificial neural networks approach", *Intelligent Systems in Accounting, Finance and Management*, Vol. 9 No. 3, pp. 145-157.

Grace, K., Salvatier, J., Dafoe, A., Zhang, B. and Evans, O. (2017), "When will ai exceed human performance? Evidence from AI Experts", ArXiv:1705.08807 [Cs], available at http://arxiv.org/abs/1705.08807 (accessed September 21, 2018).

Green, B. and Choi, H. (1997), "Assessing the risk of management fraud through neural network technology", *Auditing: A Journal of Practice & Theory*, Vol. 16 No. 1, pp. 14-28.

Hajek, P. and Henriques, R. (2017), "Mining corporate annual reports for intelligent detection of financial statement fraud – a comparative study of machine learning methods", *Knowledge-Based Systems*, Vol. Vol, 125 No. 15, pp. 139-152.

Harymawan, I. and Nurillah, D. (2017), "Do reputable companies produce a high quality of financial statements?", *Asian Journal of Accounting Research*, Vol. 2 No. 2, pp. 1-7, available at: https://doi.org/10.1108/AJAR-2017-02-02-B001

Hiep, N. (2017), "The factors impact on conversion of financial statements from Vietnam's accounting standard (VAS) into international financing reporting standard (IFRS) – experimental research for Vietnamese companies", *International Journal of Science and Research*, Vol. 6 No. 8, pp. 396-406.

Hogan, C.E., Rezaee, Z., Riley, R.A. and Velury, U.K. (2008), "Financial statement fraud: Insights from the academic literature", *Auditing*, Vol. 27 No. 2, pp. 231-252.

Hoogs, B., Kiehl, T., Lacomb, C. and Senturk, D. (2007), "A genetic algorithm approach to detecting temporal patterns indicative of financial statement fraud", *International Journal of Intelligent Systems in Accounting, Finance and Management*, Vol. 15, pp. 41-56.

Investopedia (2017), "Frau", Investopedia, March 21, available at: www.investopedia.com/terms/f/fraud.asp (accessed August 30, 2018).

Johnstone, K.M. and Bedard, J.C. (2001), "Engagement planning, bid pricing, and client response in the market for initial attest engagements", *The Accounting Review*, Vol. 76 No. 2, pp. 199-220.

Kirkos, E., Spathis, C. and Manolopoulos, Y. (2007), "Data mining techniques for the detection of fraudulent financial statements", *Expert Systems with Applications*, Vol. 32 No. 4, pp. 995-1003.

Kotsiantis, S., Koumanakos, E., Tzelepis, D. and Tampakas, V. (2006), "Forecasting fraudulent financial statements using data mining", *International Journal of Computational Intelligence*, Vol. 3 No. 2, pp. 104-110.

Kranacher, M.-J., Riley, R. and Wells, J.T. (2010), *Forensic Accounting and Fraud Examination*, John Wiley & Sons.

Li, Z. (2016), "Anomaly detection and predictive analytics for financial risk management", available at: https://rucore.libraries.rutgers.edu/rutgers-lib/49363/ (accessed September 21, 2018).

Lin, J., Hwang, M. and Becker, J. (2003), "A fuzzy neural network for assessing the risk of fraudulent financial reporting", *Managerial Auditing Journal*, Vol. 18 No. 8, pp. 657-665.

Lokanan, M.E. (2014), "The demographic profile of victims of investment fraud: a Canadian perspective", *The Journal of Financial Crime*, Vol. 21 No. 2, pp. 226-242.

Lokanan, M.E. (2015), "Challenges to the fraud triangle: questions on its usefulness", *Accounting Forum*, Vol. 39 No. 3, pp. 221-224, available at: https://doi.org/10.1016/j.accfor.2015.05.002

Lokanan, M.E. (2017), "Theorizing financial crimes as moral actions", *European Accounting Review*, Vol. 27 No. 5, pp. 1-38.

Lokanan, M.E. and Sharma, S. (2018), "A fraud triangle analysis of the Libor fraud", *Journal of Forensic and Investigative Accounting*, Vol. 10 No. 2, pp. 187-212.

Lyandres, E. and Zhdanov, A. (2013), "Investment opportunities and bankruptcy prediction", *Journal of Financial Markets*, Vol. 16 No. 3, pp. 439-476.

Morales, J., Gendron, Y. and Guénin-Paracini, H. (2014), "The construction of the risky individual and vigilant organization: a genealogy of the fraud triangle", *Accounting, Organizations and Society*, Vol. 39 No. 3, pp. 170-194.

Morozov, I. (2016), "Anomaly detection in financial data by using machine learning methods", available at: https://users.informatik.haw-hamburg.de/~ubicomp/arbeiten/bachelor/morozov.pdf (accessed September 21, 2018).

Murphy, P.R. (2012), "Attitude, machiavellianism and the rationalization of misreporting", *Accounting, Organizations and Society*, Vol. 37 No. 4, pp. 242-259.

Murphy, P.R. and Dacin, M.T. (2011), "Psychological pathways to fraud: understanding and preventing fraud in organizations", *Journal of Business Ethics*, Vol. 101 No. 4, pp. 601-618.

Narayan, P.K. and Zheng, X. (2010), "Market liquidity risk factor and financial market anomalies: evidence from the Chinese stock market", *Pacific-Basin Finance Journal*, Vol. 18 No. 5, pp. 509-520.

Perols, J. (2011), "Financial statement fraud detection: an analysis of statistical and machine learning algorithms", *Auditing: A Journal of Practice & Theory*, Vol. 30 No. 2, pp. 19-50.

Phan, D., Joshi, M. and Mascitelli, B. (2018), "What influences the willingness of Vietnamese accountants to adopt international financial reporting standards (IFRS) by 2025?", *Asian Review of Accounting*, Vol. 26 No. 2, pp. 225-247.

Phua, C., Alahakoon, D. and Lee, V. (2004), "Minority report in fraud detection: classification of skewed data", *SIGKDD Explorations*, Vol. 6 No. 1, pp. 50-59.

Phua, P.K.H., Ming, D. and Lin, W. (2001), "Neural network with genetically evolved algorithms for stocks prediction", Scopus, available at: http://scholarbank.nus.sg/handle/10635/42439 (accessed June 6, 2018).

Power, M. (2013), "The apparatus of fraud risk", *Accounting, Organizations and Society*, Vol. 38 Nos 6-7, pp. 525-543.

Purda, L. and Skillicorn, D. (2015), "Accounting variables, deception, and a bag of words: assessing the tools of fraud detection", *Contemporary Accounting Research*, Vol. 32 No. 3, pp. 1193-1223.

Rezaee, Z. (2002), *Financial Statement Fraud: Prevention and Detection*, John Wiley & Sons, New York, NY.

Rezaee, Z. (2005), "Causes, consequences, and deterrence of financial statement fraud", *Critical Perspectives on Accounting*, Vol. 16 No. 3, pp. 277-298.

Robinson, S. and Lokanan, M.E. (2017), "Testing for impression management in creative accounting: a case of the automobile industry", *Journal of Forensic and Investigative Accounting*, Vol. 9 No. 3, pp. 962-978.

Shin, K.S., Lee, T. and Kim, H.J. (2005), "An application of support vector machines in bankruptcy prediction model", *Expert Systems with Application*, Vol. 28 No. 1, pp. 127-135.

Skousen, C., Smith, K. and Wright, C. (2009), "Detecting and predicting financial statement fraud: the effectiveness of the fraud triangle and SAS No. 99", in Hirschey, M., John, K. and Makhija, A. (Eds), *Corporate Governance and Firm Performance (Advances in Financial Economics)*, Vol. 13, Emerald Group Publishing Limited, Bingley, pp. 53-81, available at: https://doi.org/10.1108/S1569-3732(2009)0000013005

Song, X.P., Hu, Z.H., Du, J.G. and Sheng, Z.H. (2014), "Application of machine learning methods to risk assessment of financial statement fraud: evidence from China", *Journal of Forecasting*, Vol. 33 No. 8, pp. 611-626.

Spathis, C. (2002), "Detecting false financial statements using published data: some evidence from Greece", *Managerial Auditing Journal*, Vol. 17 No. 4, pp. 179-191.

Thongkam, J., Xu, G., Zhang, Y. and Huang, F. (2008), "Support vector machine for outlier detection in breast cancer survivability prediction", *Advanced Web and Network Technologies, and Applications, Presented at the Asia-Pacific Web Conference, Springer, Berlin and Heidelberg*, pp. 99-109.

Tinoco, M. and Wilson, N. (2013), "Financial distress and bankruptcy prediction among listed companies using accounting, market and macroeconomic variables", *International Review of Financial Analysis*, Vol. 30, pp. 394-419, available at: https://doi.org/10.1016/j.irfa.2013.02.013

Torbeck, L.D. (2010), "Statistical solutions: on the verge of significance: why 5%", available at: www.pharmtech.com/statistical-solutions-verge-significance-why-5 (accessed August 5, 2018).

Tran, N.P. (2013), "Analyzing financial ratios to detect frauds and misstatements in financial statements of Vietnamese listed companies", Thesis from the University of Economics, Ho Chi Minh City.

West, D., Dellana, S. and Qian, J. (2005), "Neural network ensemble strategies for decision applications", *Computer and Operations Research*, Vol. 32 No. 10, pp. 2543-2559.

World Bank (2017), "Vietnam: country overview", available at: www.worldbank.org/en/country/ (accessed June 6, 2018).

Further reading

Tran, M.D., Dang, N.H. and Hoang, T.V.H. (2018), "Research on misstatements in financial statements: the case of listed firms on Ho Chi Minh City stock exchange", Vol. 15.

About the authors

Mark Lokanan is Associate Professor in the Faculty of Management at Royal Roads University. He is a graduate from Simon Fraser University, Canada, and is an expert in fraud, forensic and investigative accounting. Mark Lokanan is the corresponding author and can be contacted at: mark.lokanan@royalroads.ca

Vincent Tran received the Bachelor of Business Administration Degree (major) in Sustainability from Royal Roads University. Vincent is currently Research Assistant for Professor Mark Lokanan (Royal Roads University) in forensic accounting studies about Canadian investment industry regulators. Vincent is an expert in data collection, manipulation and analysis; background in Finance and Accounting.

Nam Vuong received the Bachelor of International Economics Degree (major) from Foreign Trade University. Nam is currently Student & researcher. Nam is an expert in information retrieval, simulation, game theory and causal inference.