sustainability

MDPI

Article

Detecting Fraudulent Financial Statements for the Sustainable Development of the Socio-Economy in China: A Multi-Analytic Approach

Jianrong Yao, Yanqin Pan, Shuiqing Yang *, Yuangao Chen and Yixiao Li

School of Information Management and Engineering, Zhejiang University of Finance and Economics, Xiasha Higher Education Zone, Hangzhou 310018, Zhejiang, China; y6310@163.com (J.Y.); panyanqin_9511@163.com (Y.P.); chenyg@zufe.edu.cn (Y.C.); yxli@zufe.edu.cn (Y.L.)

* Correspondence: yangshuiqing@zufe.edu.cn

Received: 21 February 2019; Accepted: 11 March 2019; Published: 15 March 2019

Abstract: Identifying financial statement fraud activities is very important for the sustainable development of a socio-economy, especially in China's emerging capital market. Although many scholars have paid attention to fraud detection in recent years, they have rarely focused on both financial and non-financial predictors by using a multi-analytic approach. The present study detected financial statement fraud activities based on 17 financial and 7 non-financial variables by using six data mining techniques including support vector machine (SVM), classification and regression tree (CART), back propagation neural network (BP-NN), logistic regression (LR), Bayes classifier (Bayes) and K-nearest neighbor (KNN). Specifically, the research period was from 2008 to 2017 and the sample is companies listed on the Shanghai stock exchange and Shenzhen stock exchange, with a total of 536 companies of which 134 companies were allegedly involved in fraud. The stepwise regression and principal component analysis (PCA) were also adopted for reducing variable dimensionality. The experimental results show that the SVM data mining technique has the highest accuracy across all conditions, and after using stepwise regression, 13 significant variables were screened and the classification accuracy of almost all data mining techniques was improved. However, the first 16 principal components transformed by PCA did not yield better classification results. Therefore, the combination of SVM and the stepwise regression dimensionality reduction method was found to be a good model for detecting fraudulent financial statements.

Keywords: fraudulent financial statements; data mining; support vector machine (SVM); dimensionality reduction; stepwise regression; China

1. Introduction

Financial statements are the basic documents that reflect a company's financial status, its operating results and cash flows during a specific accounting period [1]. So, financial statements are the main reference for decision-making for regulators, investors, creditors, and stakeholders. However, in the past few years, financial statement fraud incidents have occurred frequently in China. Examples are Yin Guangxia in 2001, Lantian stock in 2002, Green Land in 2013, and Xin Tai Electric in 2016. Moreover, both the magnitude and pace of financial statement fraud are growing. Increasing high-profile financial statement fraud has not only impeded corporate growth, but has also resulted in great damage to the sustainable development of the socio-economy in China [1]. Therefore, in China's capital market, it is critical to develop effective methods to detect financial statement fraud activities.

With regards to the issue of the financial statement fraud, in practice, auditors have become limited in their ability to detect fraudulent financial statements. On the one hand, they lack experience

in fraud identification. On the other hand, considering the ratio of cost and return it is impossible for auditors to spend the time required to discover all the fraud that occurs. As a result, some accounting firms and companies have begun to use data mining techniques such as cloud auditing to identify fraudulent financial statements (FFS). In academia, research is gradually shifting from using traditional statistical analysis methods to using data mining techniques. The use of data mining techniques can solve the two main shortcomings of the traditional statistical analysis methods: (1) the data used must satisfy strict hypotheses, for example, the data must meet the normal distribution. However, the data used to detect fraudulent financial statements (FFS) usually contain a large number of financial indicators, which are generally not normally distributed [2]; and (2) there are many disadvantages in the detection results, for example, the classification error rate is always very high [3].

Although data mining has generally been shown to be an effective approach in detecting FFS, many problems still remain. First, most of the data mining predictive models of FFS are only trained with financial indicators such as the debt to equity ratio, the sales growth ratio and so on [2,4–7], but non-financial indicators have received relatively little attention. In fact, some empirical studies have unveiled a significant relationship between abnormal non-financial indicators and financial statement fraud. [8,9]. For instance, Basely [8] found that the inclusion of lower proportions of outside members on the board of directors significantly increases the likelihood of financial statement fraud. Kim et al. [9] also discovered that intentionally misstating firms tend to show a higher firm-efficiency ranking. Second, with the continuous development of China's socio-economy, China has become the world's second largest economy, and China's capital market has also developed rapidly. Financial fraud activities have occurred frequently due to lack of effective supervision. Therefore, it is very necessary to establish an effective predictive model of FFS for China's capital market. Unfortunately, almost all of the existing models are developed in the United States [10,11]. Since the US capital market and China's capital market are different in many respects, most of the existing models may not be as well suited for China. Third, detecting FFS is a complex but important task. The number of predictive variables in the models ranges from 20 [12–14] to 100 [6]. When adding more and more predictive variables, dimension disaster may happen, which may cause (1) overfitting of sample data; (2) instability of models; and (3) difficulty in promoting the models. Therefore, it is better to select a subset of the original predictive variables by eliminating variables that are redundant: this is called dimensionality reduction. At present, the most common and relatively mature dimensionality reduction techniques can be summarized into two categories, namely, feature selection and feature transformation. Feature selection techniques select the best indicators from the original variables pool, which retains the raw data and makes it easier to explain. Feature transformation techniques combine many original variables with new variables by means of linear or nonlinear transformation methods. These two techniques have their own advantages and disadvantages. However, studies have rarely adopted both of them or made a comparative analysis to investigate which type of techniques is more suitable to detect fraudulent financial statements [2,7].

Overall, based on previous research, this study makes the following contributions to the literature. First, in order to further improve the performance of the predictive models, we have added Altman's Z-score, a comprehensive financial indicator, and financial variables that reflect a company's profitability, operational capabilities, solvency and growth capabilities. Altman's Z-score is an indicator that predicts whether a company will face financial distress and Kirkos et al. [2] believe that financial distress may be related to financial fraud. Besides, an additional seven non-financial variables that reflect a company's management structure, internal control, shareholders structure and external audit were also included in our study to catch as many predictors as possible. Second, the predictive models for FFS detecting were conducted in the context of China and the Wind database was used to collect the experimental data in our study. Wind is the leading financial data, information and software services company in mainland China, and a large number of domestic and international media, research reports, academic papers, etc. often quote data provided by Wind. The Wind database contains comprehensive historical fraud information for Chinese listed companies. Third, in the

process of dimensionality reduction, stepwise regression and principal component analysis (PCA) were adopted as the feature selection technique and the feature transformation technique, respectively, and we made a comparative analysis to find which type of techniques is more suitable for FFS detection.

In this study, we conducted a comparative study by employing six data mining techniques and two dimensionality reduction methods to detect the FFS of companies listed on the Shanghai and Shenzhen stock exchange during the period 2008–2017. The six classifiers are support vector machine (SVM), classification and regression tree (CART), back propagation neural network (BP-NN), logistic regression (LR), Bayes classifier (Bayes), K-nearest neighbor (KNN). Also, we selected accuracy, recall (sensitivity), precision, F-score and area under the receiver operating characteristic curve (AUC) as metrics to evaluate the classification performance of each classifier.

The remainder of this paper is structured as follows. The next section reviews previous literature on FFS detection with data mining techniques, introduces the main data mining techniques and dimensionality reduction methods in detecting FFS, and also describes the data collected from the Wind database and the experimental setting of this research. Section 3 presents the experimental results. Finally, a detailed discussion appears in Section 4.

2. Materials and Methods

Previous studies have reported the superior classification performance of data mining techniques over traditional statistical methods [1–7,9–20]. The literature related to FFS detection driven by data mining techniques is presented in Table 1.

Table 1. Research on data mining techniques in detecting fraudulent financial statements (FFS).

Sl. No	Author	Publication Year	Academic Journal/Conference
1	Cerullo, M.J. and Cerullo, V.	1999	Computer Fraud & Security
2	Bell, T.B. and Carcello, J.V.	2000	Auditing: A Journal of Practice & Theory
3	Kotsiantis, S. et al.	2006	International Journal of Computational Intelligence
4	Kirkos, E. et al.	2007	Expert Systems with Applications
5	Ravisankar, P. et al.	2011	Decision Support Systems
6	Zhou, W. and Kapoor, G.	2011	Decision Support Systems
7	Glancy, F.H. and Yadav, S.B.	2011	Decision Support Systems
8	Humpherys, S.L. et al.	2011	Decision Support Systems
9	Huang, S.Y.	2013	International Journal of Digital Content Technology and its Applications
10	Dong, W. et al.	2014	PACIS 2014 Proceedings
11	Huang, S.Y. et al.	2014	Expert Systems with Applications
12	Lin, C.C. et al.	2015	Knowledge-Based Systems
13	Liu, C.W. et al.	2015	International Journal of Economics and Finance
14	Kim, Y.J. et al.	2016	Expert Systems with Applications
15	Yeh, C.C.	2016	Cybernetics and Systems: An International Journal
16	Chen, S.	2016	SpringerPlus
17	Hajek, P. and Henriques, R.	2017	Knowledge-Based Systems
18	Dutta, I. et al.	2017	Expert Systems with Applications
19	Jan, C.L.	2018	Sustainability

In this paper, 19 related articles were selected to analyze and summarize the data mining techniques and dimensionality reduction methods frequently used in the field of fraudulent financial statements detection. The reasons for selecting the 19 articles as our review literature are: (1) these articles were published in journals or conferences that are closely related to our research topic so this ensures the representativeness of our review; (2) these articles were published from 1999 to 2018, so the time span is large and it reflects the progress of research in this field in a more comprehensive way.

2.1. Data Mining Techniques in FFS Detection

To determine the main data mining techniques used for detecting FFS, we present a simple analysis of the 19 articles in Table 2. We found that support vector machine, decision trees, neural networks, and logistic regression were the most popular techniques, being used in 71.93% (41 of 57). Also some articles used the Bayesian network, and text mining techniques to identify the FFS. The K-nearest neighbor, rough set theory, genetic programming, and random forest are used relatively infrequently in these 19 selected articles. After comprehensive consideration, we chose support vector machine (SVM), classification and regression tree (CART) (which belongs to decision tree), back propagation neural network (BP-NN), logistic regression (LR), Bayes classifier (Bayes) and K-nearest neighbor (KNN) as the data mining techniques tested in this paper.

Table 2. Data mining techniques in the 19 selected articles.

Sl. No	Techniques	Amount
1	Support Vector Machine	11
2	Decision Tree	10
3	Neural Network	10
4	Logistic Regression	9
5	Bayesian Network	8
6	Text Mining Techniques	3
7	K-Nearest Neighbor	2
8	Rough Set Theory	1
9	Genetic Programming	1
10	Random Forest	1
	Total	56

These selected data mining techniques have varied backgrounds and different theories to support them. In this way, we ensured that the problem at hand was analyzed by disparate models with varying degrees of computational complexity and performance on FFS detection problems.

Support vector machine (SVM) is an artificial intelligence learning method developed by Vapnik [20]. It is a machine learning technique based on statistical learning theory and structural risk minimization. The purpose is to identify the optimal separating hyperplane to divide two or more classes of data with the learning mechanism by training the input data. It is a type of supervised learning to predict and classify items in the field of data mining. SVM is prone to overfitting but performs well on noisy financial fraud data [21].

Classification and regression tree (CART) is a binary decision tree technique used for continuous data or non-parametric data for classification. The decision of dividing conditions is based on the quantity and attributes of the data, as well as the Gini index. Each division separates the data into two subsets, and the process is repeated for each subset to identify the next dividing conditions. Data continue to be divided into two subsets in order to construct a tree structure. The process is finished when data is no longer divisible.

Neural networks derived from the modeling of the human brain and display good performance in signal restoration. Back propagation (BP) is a common method of training artificial neural network to minimize the objective function, which is a supervised learning method. It requires making up the training set for a dataset with many inputs. Figure 1 is a classical back propagation neural network (BP-NN) architecture containing the input layer, hidden layer, and output layer, which has one output, m inputs, and n neurons in the hidden layer.

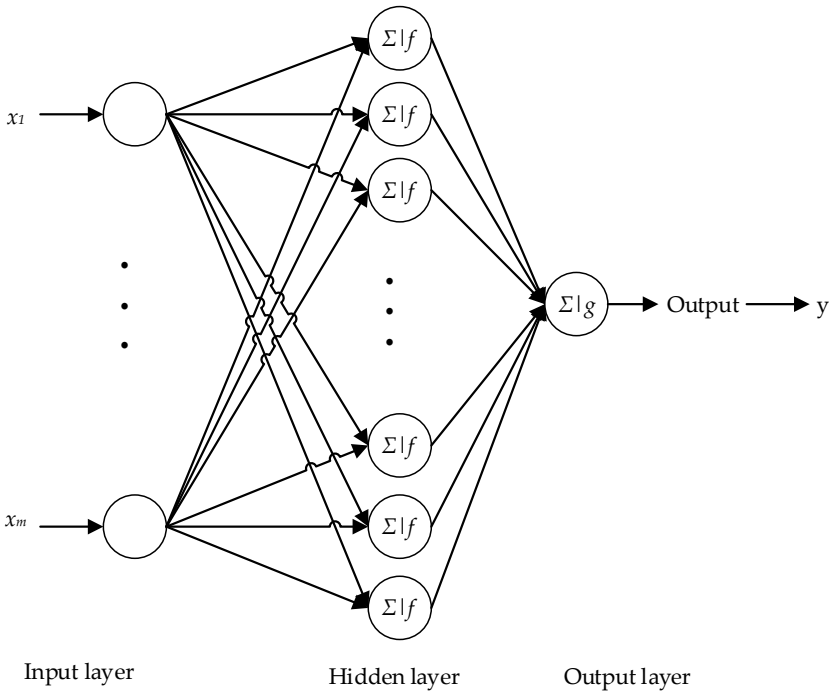

Figure 1. Topology of an m-n-1 back propagation neural network (BP-NN).

Logistic regression (LR) is used to deal with regression problems where the dependent variable is a classification variable. According to Williams et al. [22], LR is a commonly used approach for performing binary classification. The advantage of the LR is its computation which is not expensive and it is easy to understand and implement. However, its outcome is easily under-fitted and the classification accuracy is not always high.

The Bayes classifier is a popular classification algorithm used extensively in financial fraud detection. Its design method is one of the most basic statistical classification methods. The classification principle is to calculate the posterior probability by using the Bayesian formula through the prior probability of an object, that is, the probability that the object belongs to a certain class and selecting the class with the largest posterior probability as the class to which the object belongs.

The K-nearest neighbor (KNN) method computes the class label for the test samples by the labels of the k nearest neighbors of the test samples. The KNN method is only relevant to a very small number of adjacent samples in class decision making. Since the KNN method mainly relies on the surrounding limited samples, rather than relying on the discriminant domain method to determine the category, the KNN method is more suitable for the cross-over or overlapping sample sets of the class domain.

2.2. Dimensionality Reduction Methods in FFS Detection

The accuracy can be improved significantly by screening out those variables which have the greater effect in detecting FFS [1]. In the field of detection FFS, dimensionality reduction methods are often used to screen out those variables. So, we summarized the dimensionality reduction methods used in these 19 articles. First of all, we found that 14 articles adopted the dimensionality reduction methods and the other five articles did not use these techniques. It can be seen that the use of dimensionality reduction methods is the mainstream in detecting FFS. Secondly, we further summarized the specific dimensionality reduction methods used in these articles, which are shown in Table 3. Obviously, every article either selected a feature selection technique or a feature transformation technique. Few articles used two categories of dimensionality reduction methods at the same time to make a comparative analysis. In order to address this research gap, we selected the stepwise regression

method and PCA method, which are good representatives of the two categories of dimensionality reduction methods, to analyze which type of method is more suitable for fraud detection.

Table 3. Dimensionality reduction methods in the 14 related articles.

Sl. No	Dimensionality Reduction Method(s)	Amount	Category
1	T-Statistic	2	feature selection
2	Principal Component Analysis (PCA)	2	feature transformation
3	Relief	2	feature selection
4	ANOVA	1	feature selection
5	Factor Analysis	1	feature transformation
6	Random Forest	1	feature selection
7	Multinomial Logistic Regressions (MLR)	1	feature selection
8	Rough Set Theory	1	feature selection
9	Classification and Regression Trees (CART) & Chi Squared Automatic Interaction Detector (CHAID)	1	feature selection
10	Stepwise Forward Selection	1	feature selection
11	Artificial Neural Network (ANN) & Support Vector Machine (SVM)	1	feature selection

The stepwise regression method is based on the assumption that under linear conditions, the variable combinations that can account for more dependent variable variation are retained. There are three specific methods of stepwise regression: (1) Forward selection. First, there is only one independent variable that can explain the largest variation of the dependent variable. Next, another independent variable is added to see if the variation in the dependent variable can be significantly explained. This process is iterative until there is no independent variable that meets the conditions for joining the model; (2) Backward elimination. First, all independent variables are put into the model. Then, one of the independent variables is removed to see if the entire model still interprets the variation in the dependent variable significantly. This process is iterative until there is no independent variable that meets the culling conditions; and (3) Bi-directional elimination. This method is equivalent to combining the methods of forward selection and backward elimination. This method does not blindly increase independent variables or eliminate independent variables. Instead, after adding an independent variable, it tests all the independent variables in the whole model and then eliminates the independent variables to get an optimal combination of variables. So, we selected the bi-directional elimination method.

Principal component analysis (PCA) is a data dimensionality reduction method for continuous attributes. It constructs an orthogonal transformation of the original data. The base of the new space removes the correlation of the data under the original spatial base and only a few new variables can be used to explain most of the variation in the raw data. These new variables are called principal components. The calculation steps of PCA are as follows:

Step 1. Let n times observed data matrix of the original variables $X_1, X_2, \ldots, X_p$ be:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = (X_1, X_2, \cdots, X_p) \quad (1)$$

Step 2. The data matrix is centrally standardized by column. For convenience, the standardized data is still recorded as X.

Step 3. Calculate the correlation coefficient matrix R, $R = (r_{ij})_{p*p}$, the definition of r_{ij} is as follows: $(r_{ij} = r_{ji})$:

$$r_{ij} = \sum_{k=1}^{n} (x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j) / \sqrt{\sum_{k=1}^{n} (x_{ki} - \overline{x}_i)^2 \sum_{k=1}^{n} (x_{kj} - \overline{x}_j)^2} \quad (2)$$

Step 4. Calculate eigenvalue: $\det(R - \lambda E) = 0, \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.

Step 5. Determine the number of principal components m:

$$\frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \geq \alpha \quad (3)$$

α is determined according to the actual problem, generally taking 80%.

Step 6. Calculate corresponding unit eigenvectors:

$$\beta_1 = \begin{bmatrix} \beta_{11} \\ \beta_{21} \\ \vdots \\ \beta_{p1} \end{bmatrix}, \beta_2 = \begin{bmatrix} \beta_{12} \\ \beta_{22} \\ \vdots \\ \beta_{p2} \end{bmatrix}, \cdots, \beta_m = \begin{bmatrix} \beta_{1m} \\ \beta_{2m} \\ \vdots \\ \beta_{pm} \end{bmatrix}. \quad (4)$$

Step 7. Calculate principal components:

$$Z_i = \beta_{1i} X_1 + \beta_{2i} X_2 + \cdots + \beta_{pi} X_p, i = 1, 2, \cdots, m \quad (5)$$

2.3. Data and Experimental Setting

2.3.1. Data Sources

The academic community has developed two main definitions of financial statement fraud: first, the Securities Regulatory Commission (SRC) administrative penalty announcements; and second, the audit opinion in the audit report. Among these, the method of defining fraud according to the SRC's administrative penalty announcements has been accepted by most scholars. The advantage of this method is that the fraud samples obtained have true fraud. However, the drawback is that the number of fraud samples that can be obtained is often relatively small. Meanwhile, according to the audit opinion in the audit report, the number of fraud samples that can be obtained is often relatively large, however, this method of definition is not reliable. Due to the existence of commercial buyouts and collusion, there are few audit reports that clearly indicate the existence of fraud in financial statements. At the same time, the financial statements that get "standard unqualified" audit opinions are not necessarily true.

In view of the fact that the purpose of this paper is to identify truly fraudulent financial statements (FFS), we chose the method of defining fraud according to the SRC's administrative penalty announcements. Therefore, the fraudulent companies in this paper are the companies punished by the China Securities Regulatory Commission (CSRC) for violating financial statement disclosure standards.

We identified a total of 134 public companies involved in alleged instances of fraudulent financial statements (FFS) during the period 2008–2017 from the Wind database. They are all listed companies in the Shanghai Stock Exchange and Shenzhen Stock Exchange. To obtain a matched sample of non-fraudulent firms, we identified 402 firms with the same scale of total assets (±10%), same industry (industry standard of CSRC) and the same year with a matching ratio of 1:3 [7]. The industry and year distribution of these 536 public companies are shown in Table 4 and Figure 2.

Table 4. Industry distribution of the 536 public companies.

Industry	Fraud Count	Non-Fraud Count	Proportion	Cumulative Proportion
Manufacturing	66	198	49.25	49.25
Information transmission, software and information technology services	15	45	11.19	60.45
Wholesale and retail	10	30	7.46	67.91
Real estate	8	24	5.97	73.88
Mining	6	18	4.48	78.36
Power, heat, gas and water production and supply	6	18	4.48	82.84
Agriculture, forestry, animal husbandry and fishery	5	15	3.73	86.57
Construction	4	12	2.99	89.55
Transportation, warehousing and postal services	4	12	2.99	92.54
Management of water conservancy, environment and public facilities	2	6	1.49	94.03
Rental and business services	2	6	1.49	95.52
Comprehensive	2	6	1.49	97.01
Scientific research and technical services	1	3	0.75	97.76
Health and social work	1	3	0.75	98.51
Culture, sports and entertainment	1	3	0.75	99.25
Accommodation and catering	1	3	0.75	100.00
Total	134	402	100.00	

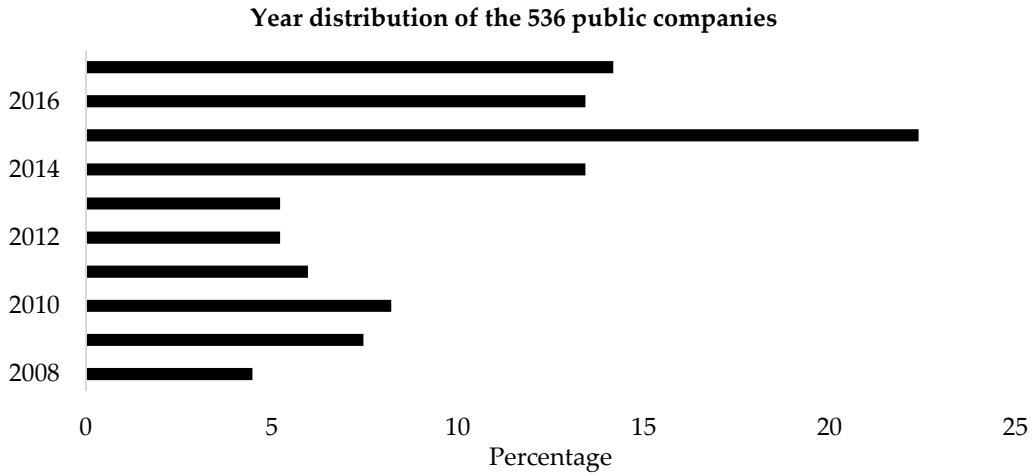

Figure 2. Year distribution of the 536 public companies.

From Table 4, we can see that nearly half of the companies are in the manufacturing industry, which may result from the fact that the proportion of manufacturing companies to all public companies itself is large. From Figure 2, we see that the year the fraud occurred is concentrated in 2014–2017, which accounts for more than 50 percent.

2.3.2. Variables Selection

According to existing research, both financial variables and non-financial variables have a relationship with fraudulent financial statements (FFS). Thus, we selected financial variables and non-financial variables from the Wind database to detect the FFS. All variables were extracted from the annual consolidated financial statements.

- Financial Variables

In this paper, financial variables were selected to detect FFS and the chosen set of financial variables cover most aspects of a firm's financial performance in order to detect various kinds of FFS. Previous literature has provided strong theoretical evidence for the use of financial variables [23]. Our selection of the financial variables listed in Table 5 was therefore influenced by previous FFS

detection studies. The financial variables selected can be divided into five categories: (1) Z-score, a comprehensive financial indicator used to predict the financial distress of a company proposed by Altman [2]; (2) profitability indicators; (3) solvency indicators; (4) operation indicators; and (5) growth indicators.

Table 5. Financial variables used for the detection of FFS.

	Variables	Variables Description	Formula	Category
X1	Z-score	a comprehensive financial indicator used to predict whether a company will go bankrupt	1.2 * net working capital/total assets + 1.4 * retained earnings/total assets + 3.3 * EBIT/total assets + 0.6 * stock value /total liabilities + 0.999 * sales/total assets	a comprehensive financial indicator
X2	ROE	return on equity	net profit/total equity	Profitability indicators
X3	ROA	return on total assets	total profit/total assets	
X4	EPS	earnings per share	net profit/total equity	
X5	NPS	net profit margin on sales	net profit/sales	
X6	CR_1	current ratio	current assets/current liabilities	Solvency indicators
X7	QR	quick ratio	quick assets/current liabilities	
X8	DAR	debts assets ratio	total liabilities/total assets	
X9	CR_2	cash ratio	monetary fund/current liabilities	
X10	TAT	total assets turnover	net operating income/total assets	Operation indicators
X11	FAT	fixed assets turnover	net operating income/fixed assets	
X12	CAT	current assets turnover	net operating income/current assets	
X13	ART	accounts receivable turnover	net sales on credit/accounts receivable	
X14	OI-growth	growth rate of operating income	growth in operating income/total operating income of the previous year	Growth indicators
X15	EPS-growth	growth rate of earnings per share	growth in earnings per share/earnings per share of the previous year	
X16	NP-growth	growth rate of net profit	growth in net profit/net profit of the previous year	
X17	TA-growth	growth rate of total assets	growth in total assets/total assets of the previous year	

- Non-Financial Variables

In order to effectively detect FFS, it is necessary to use not just financial variables but also non-financial variables that have been found to have some predictability in detecting FFS. Researchers have found that non-financial variables (e.g., low proportions of outside members on the board of directors) may signal the presence of possible accounting manipulation [8]. So, in this paper, seven non-financial variables were selected that reflect a company's management structure, internal control, shareholders structure and external audit. Our selection of the non-financial variables is listed in Table 6.

Table 6. Non-financial variables used for the detection of FFS.

No.	Variables	Variables Description
X18	Num_dir	number of directors
X19	Num_ind_dir	number of independent directors
X20	Rat_ext_dir	ratio of external directors
X21	Top1	the shareholding ratio of the largest shareholder
X22	Top10	total shareholding ratio of the top ten shareholders
X23	Big4	whether is audited by Big 4 accounting firms ("0" represents "not", "1" represents "yes") (Big 4 accounting firms refer to PWC, KPMG, DTT and EY)
X24	Aud_rep	Type of audit report ("0" represents "unqualified opinion", "1" represents "qualified opinion")

2.3.3. Descriptive Statistics of the Data

In this paper, we collected the financial and non-financial variables from the Wind database. For the imputation of the missing values, we used the "mean value" to estimate the missing values, and we also dealt with the outliers by using the "mean value". Table 7 shows the basic descriptive statistics of the samples. The most striking result to emerge from the data is that fraudulent firms showed a higher debt-asset ratio (DAR), a lower total asset turnover (TAT), a lower current asset turnover (CAT) and a lower shareholding ratio for the largest shareholder (Top1). In Table 7, the Big4 line represents the number of firms audited by the Big 4 accounting firms or not (the number of companies not audited by the Big4 accounting firms is on the left of the colon and the number of companies audited by the Big4 is on the right of the colon). The Aud_rep line represents the number of firms that got a qualified opinion or not (the number of companies who got an unqualified opinion is on the left of the colon and the number of companies who got a qualified opinion is on the right of the colon).

Table 7. Descriptive statistics of the samples.

	FFS			**Non-FFS**		
	mean	**std**	**cv**	**mean**	**std**	**cv**
Z-score	8.288	15.7935	1.9056	9.284	14.0568	1.5141
ROE	−17.14	136.3167	−7.9531	5.475	27.3674	4.9986
ROA	−0.0163	17.7052	−1088.211	6.303	10.7819	1.7106
EPS	-0.1031	1.3776	−13.3621	0.2593	0.4375	1.6872
NPS	9.939	171.3122	17.2364	4.467	60.118	13.458
CR_1	2.135	3.5507	1.6631	2.44	2.4651	1.0103
QR	1.702	3.5199	2.0681	2.0173	2.4058	1.1926
DAR	**58.47**	**40.499**	**0.6926**	**43.56**	**41.8474**	**0.9607**
CR_2	0.9292	2.8763	3.0954	1.142	1.7453	1.5283
TAT	**0.5319**	**0.7732**	**1.4537**	**0.7635**	**0.8666**	**1.135**
FAT	9.393	40.4364	4.305	40.38	377.6604	9.3527
CAT	**1.034**	**1.1308**	**1.0936**	**1.4355**	**1.2727**	**0.8866**
ART	18.683	48.0661	2.5727	503.05	5135.771	10.209
OI-growth	15.319	124.4192	8.1219	20.55	143.5552	6.9857
EPS-growth	−332.03	1573.886	−4.7402	−77.883	844.9721	−10.849
NP-growth	−296.81	1288.859	−4.3424	−17.35	429.8532	−24.775
TA-growth	22.367	245.9319	10.9953	20.89	107.0459	5.1243
Num_dir	8.127	1.509	0.1857	8.251	2.0612	0.2498
Num_ind_dir	3.022	0.433	0.1433	3.032	0.7108	0.2344
Rat_ext_dir	37.84	5.6409	0.1491	37.23	5.5483	0.149
Top1	**28.71**	**15.0326**	**0.5236**	**32.79**	**14.7375**	**0.4495**
Top10	49.8	16.8402	0.3382	54.15	15.9625	0.2948
Big4		131:3			391:11	
Aud_rep		126:8			399:3	

2.3.4. Experimental Setting

The experimental setting used in this study is shown in Figure 3. In order to eliminate the influence of the dimensional differences between different variables on the experiment, we implemented the "zero-mean" normalization process and normalized each of the independent variables from the original dataset during the data preprocessing stage. Also, we randomly created 10 stratified samples from the normalized dataset in which a 7:3 ratio was followed for training (376 firms in total, 188 fraudulent and 188 non-fraudulent) and testing data (160 firms in total, 80 fraudulent and 80 non-fraudulent). Furthermore, ten-fold cross-validation is performed to improve the reliability of the results. Then, we analyzed the normalized dataset using six data mining techniques including SVM, CART, BP-NN, LR, Bayes, KNN. We chose these six techniques not only because they were used frequently in prior studies,

but also because they have varied backgrounds and different theories to support them. So, we ensured that the problem of detecting FFS was studied and analyzed comprehensively from all perspectives.

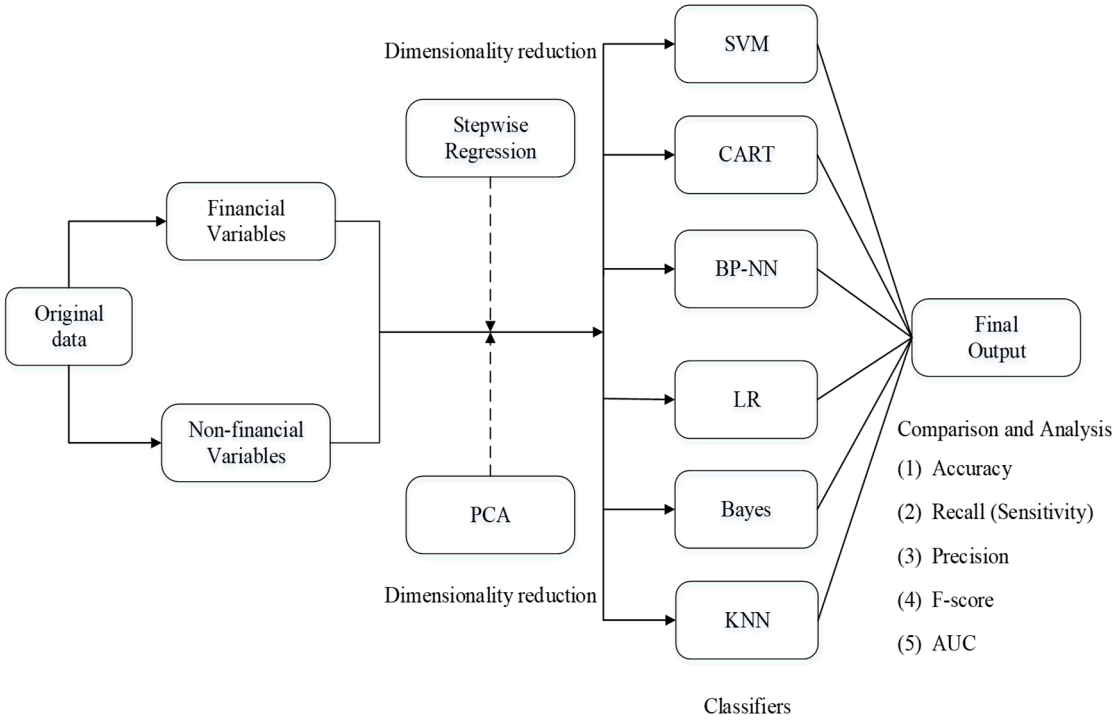

Figure 3. Experimental setting.

Next, we performed dimensionality reduction on the normalized dataset to identify the most significant variables that could detect the presence of financial statement fraud. We used the stepwise regression method and the PCA method to reduce the dimensionality of the dataset. The stepwise regression method is a feature selection technique and PCA is a feature transformation technique. This process resulted in new combinations such as stepwise regression-SVM, stepwise regression-CART, stepwise regression-BP-NN, stepwise regression-LR, stepwise regression-Bayes, stepwise regression-KNN, PCA-SVM, PCA-CART, PCA-BP-NN, PCA-LR, PCA-Bayes, and PCA-KNN.

Finally, we compared and analyzed the performance of the six data mining techniques with: (1) an all normalized dataset without dimensionality reduction; (2) a selected dataset by stepwise regression; and (3) a transformed dataset by PCA. The performance metrics are accuracy, recall (sensitivity), precision, F-score and area under the receiver operating characteristic curve (AUC).

- Performance Metrics

Financial statement fraud detection represents a binary classification problem with four possible classification outcomes [24]: (1) true positive (a non-fraud firm correctly classified as a non-fraud firm); (2) false negative (a non-fraud firm incorrectly classified as a fraud firm); (3) true negative (a fraud firm correctly classified as a fraud firm); and (4) false positive (a fraud firm incorrectly classified as a non-fraud firm).

Classification performance can be measured in many different ways: absolute ability, performance relative to other factors, probability of success, and others [25]. In this paper we adopted accuracy, recall (sensitivity), precision, F-score and the area under the receiver operating characteristic curve (AUC). They are described in Table 8.

Table 8. Definition of our performance metrics in terms of confusion matrix entities. Here in our example, positive and negative instances refer to the instances of non-fraud and fraud, respectively.

Metric	Governed Equation	Definition
Accuracy	$\frac{TP+TN}{P+N}$	Proportion of the total number of predictions that are correct
Precision	$\frac{TP}{TP+FP}$	Proportion of the predicted positive cases that are correct
Recall/Sensitivity	$\frac{TP}{TP+FN}$	Proportion of positive cases that are correctly identified
Specificity	$\frac{TN}{FP+TN}$	Proportion of negative cases that are correctly identified
F-score	$\frac{2 \times Precision \times Recall}{Precision + Recall}$	Weighted harmonic mean of precision and recall
AUC	_____	Area under the receiver operating characteristic curve

3. Results

The data mining tool used in this paper is R, the six classifiers we used are basically all directly calling the relevant analysis packages in R, and most of the parameters of the functions are chosen as default values.

3.1. Classification without Dimensionality Reduction

Table 9 summarizes the results of the experiment for all variables without dimensionality reduction. From Table 9, we can see that the performance of the six classifiers we selected is satisfactory, all the classification accuracy is over 0.73, and the accuracy of SVM is more than 0.80; The trend in the F-score performance metric is similar to the accuracy, and the SVM classifier also gets the highest F-score. After SVM, LR also performed excellently.

Table 9. Classification performance of the six classifiers using all variables.

	Accuracy	Recall/Sensitivity	Precision	F-Score	Specificity	AUC
SVM	**0.8063**	0.9417	0.8248	**0.8794**	0.4000	0.6708
CART	0.7438	0.7833	0.8624	0.8210	0.6250	0.7818
BP-NN	0.7313	0.8333	0.8130	0.8230	0.4250	0.7369
LR	0.8000	0.9333	0.8235	0.8750	0.4000	0.6667
Bayes	0.7313	0.7083	0.9140	0.7981	0.8000	0.7542
KNN	0.7813	0.9667	0.7891	0.8689	0.2250	0.5958

3.2. Classification with Stepwise Regression

After the 24 selected independent variables were normalized, the correlation matrix between them was calculated and this is shown in Figure 4. We can see that there is a correlation between multidimensional data under the original space base. Therefore, we used stepwise regression to select a few features that can explain most of the original features. Thus, redundant information can be eliminated and the performance of the model may be improved. Finally, we selected 13 significant original variables. They are ROE, ROA, NPS, DAR, CR_2, FAT, CAT, ART, EPS-growth, NP-growth, Num_dir, Top1 and Top10. Table 10 summarizes the six classifiers' classification results using 13 significant original variables selected by the stepwise regression. From Table 10, we can see that SVM is still more suitable for detecting fraudulent financial statements in this study in terms of accuracy (0.8250) and F-score (0.8915), and LR is second to SVM. Besides, almost all classifiers' performance was improved with the stepwise regression, except for the Bayes classifier. In terms of accuracy, SVM's accuracy changes from 0.8063 to 0.8250, CART's accuracy changes from 0.7438 to 0.8063, BP-NN's accuracy gets up to 0.7375 from 0.7313 and the accuracy of LR was upgraded to 0.8125 from 0.8000. KNN's accuracy has been improved 0.0187.

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20	X21	X22	X23	X24
X1	1.00	0.10	0.12	0.15	0.20	0.56	0.58	-0.62	0.49	-0.06	0.10	-0.13	-0.04	0.04	0.02	0.03	0.05	-0.10	-0.01	0.12	0.01	0.15	-0.05	-0.04
X2		1.00	0.71	0.62	0.37	0.22	0.22	-0.22	0.24	0.33	0.21	0.21	0.08	0.35	0.39	0.40	0.36	-0.08	-0.11	-0.01	0.20	0.24	0.05	-0.19
X3			1.00	0.68	0.49	0.21	0.24	-0.19	0.27	0.30	0.17	0.18	0.00	0.24	0.33	0.32	0.26	-0.06	-0.08	0.00	0.18	0.21	0.04	-0.12
X4				1.00	0.46	0.32	0.35	-0.28	0.32	0.32	0.23	0.16	0.04	0.32	0.27	0.26	0.39	-0.09	-0.10	0.02	0.22	0.28	0.04	-0.14
X5					1.00	0.26	0.25	-0.29	0.27	-0.22	0.05	-0.23	-0.06	0.05	0.24	0.24	0.19	-0.09	-0.06	0.07	0.06	0.16	0.02	0.02
X6						1.00	0.89	-0.67	0.73	0.01	0.19	-0.25	-0.11	0.12	0.04	0.06	0.14	-0.17	-0.13	0.07	0.06	0.19	-0.02	-0.07
X7							1.00	-0.67	0.79	0.03	0.17	-0.20	-0.16	0.09	0.01	0.03	0.16	-0.20	-0.16	0.09	0.06	0.21	0.00	-0.06
X8								1.00	-0.62	0.07	-0.02	0.13	0.05	-0.07	-0.05	-0.07	-0.10	0.09	0.01	-0.12	-0.08	-0.21	-0.03	0.09
X9									1.00	0.02	0.14	-0.15	-0.02	0.07	0.02	0.02	0.12	-0.17	-0.14	0.06	0.05	0.18	0.04	-0.10
X10										1.00	0.18	0.68	0.08	0.19	0.08	0.08	0.15	-0.02	-0.11	-0.12	0.02	0.07	-0.04	-0.05
X11											1.00	0.05	0.13	0.22	-0.01	0.03	0.22	-0.12	-0.14	0.01	0.09	0.10	-0.02	0.00
X12												1.00	0.16	0.11	0.09	0.09	0.06	0.01	-0.06	-0.11	0.12	0.05	0.07	-0.06
X13													1.00	0.01	0.00	0.02	-0.05	0.05	-0.03	-0.12	0.02	-0.02	0.07	-0.02
X14														1.00	0.27	0.28	0.34	-0.05	-0.06	0.01	-0.06	-0.01	-0.04	-0.12
X15															1.00	0.91	0.18	0.03	0.06	0.02	0.01	-0.09	0.01	-0.12
X16																1.00	0.19	0.00	0.01	0.01	0.00	-0.05	-0.02	-0.10
X17																	1.00	-0.07	-0.06	0.04	0.08	0.19	-0.01	-0.04
X18																		1.00	0.83	-0.47	-0.05	-0.03	0.05	-0.06
X19																			1.00	0.05	-0.06	-0.03	0.05	-0.05
X20																				1.00	0.00	0.00	0.00	0.02
X21																					1.00	0.65	0.13	-0.12
X22																						1.00	0.14	-0.08
X23																							1.00	-0.02
X24																								1.00

Figure 4. The correlation matrix of the original data.

Table 10. Classification performance of the six classifiers using 13 significant original variables.

	Accuracy	Recall/Sensitivity	Precision	F-Score	Specificity	AUC
SVM	**0.8250**	0.9583	0.8333	**0.8915**	0.4250	0.6917
CART	0.8063	0.8833	0.8618	0.8724	0.5750	0.7616
BP-NN	0.7375	0.8500	0.8095	0.8293	0.4000	0.6985
LR	0.8125	0.9417	0.8309	0.8828	0.4250	0.6833
Bayes	0.6750	0.6083	0.9359	0.7374	0.8750	0.7417
KNN	0.8000	0.9583	0.8099	0.8779	0.3250	0.5958

3.3. Classification with PCA

In the third part of our experiment, we transformed the 24 independent features and selected a few new features that can explain most of the original features to learn with PCA. Finally, we selected the first 16 principal components, which can explain 95.67% of the information in the original data.

Based on the PCA for dimensionality reduction, we used the 16 new principal components to compare the performance of the selected six classifiers. The classification results are shown in Table 11. We can see that the highest classification accuracy is still in the SVM and the corresponding F-score is also the highest. Also, the LR ranks second. In addition, after using PCA, the classification accuracy of almost all classifiers is slightly inferior to the classifiers using stepwise regression, except for the Bayes classifier. Moreover, compared to the experimental results without dimensionality reduction, we found that the classification accuracy of CART, BP-NN, and LR classifiers decreased. Figure 5 presents the comparison of the accuracy of the classification results of the three parts of the experiment. It can be clearly seen from Figure 5 that the classification accuracy after using stepwise regression is overall superior to using PCA, and the classification accuracy of the SVM classifier is always the highest. Therefore, the combination of SVM and the stepwise regression is the most suitable model for detecting fraudulent financial statements in our experiment.

Table 11. Classification performance of the six classifiers using 16 principal components.

	Accuracy	Recall/Sensitivity	Precision	F-Score	Specificity	AUC
SVM	**0.8188**	0.9833	0.8138	**0.8906**	0.3250	0.6542
CART	0.6563	0.7417	0.7876	0.7639	0.4000	0.6379
BP-NN	0.7000	0.7917	0.8051	0.7983	0.4250	0.6789
LR	0.7938	0.9250	0.8222	0.8706	0.4000	0.6625
Bayes	0.7625	0.9167	0.7971	0.8527	0.3000	0.6083
KNN	0.7813	0.9583	0.7931	0.8679	0.2500	0.5958

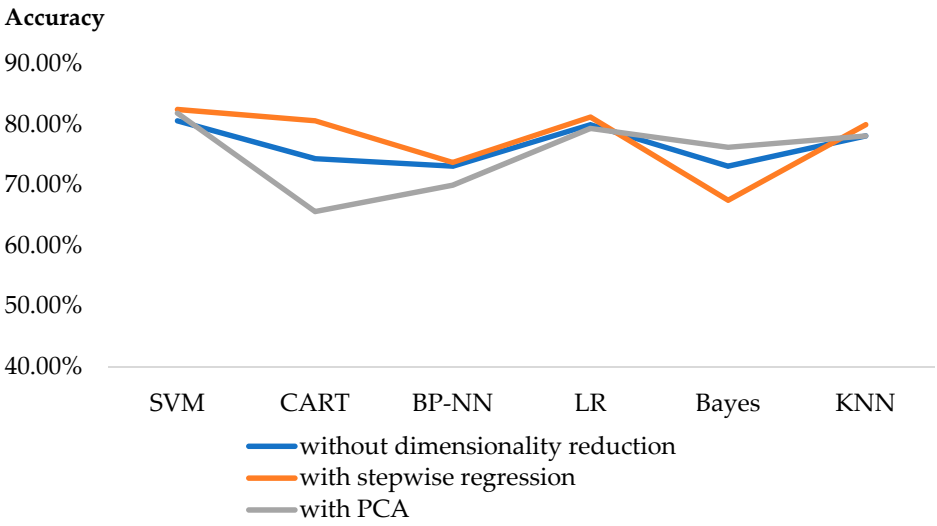

Figure 5. Comparison of the accuracy of the classification results.

4. Discussion

Financial statement fraud not only causes huge losses for investors, but also creates a crisis of distrust in accounting firms. Furthermore, it puts the company's financial situation in a vicious cycle that reduces the long-term sustainable development of the whole socio-economy [1]. Therefore, detection of fraudulent financial statements plays an important role in enhancing sustainability of the socio-economy in China. This study adopted a multi-analytic method and integrated multi-source variables to detect fraudulent financial statements of listed companies in China. Some interesting findings are provided in this study.

Our experimental results can be divided into three parts. Part 1: We found that the SVM classifier performs most outstandingly when there are 24 input variables and the accuracy is 0.8063. The accuracy of the other classifiers is 0.7438, 0.7313, 0.8000, 0.7313 and 0.7813. Part 2: Thirteen significant original variables were selected by using the stepwise regression feature selection technique. They are ROE, ROA, NPS, DAR, CR_2, FAT, CAT, ART, EPS-growth, NP-growth, Num_dir, Top1 and Top10. We found that these variables cover the four aspects of financial variables and non-financial variables. In addition, the accuracy of the classification of financial statements can be further improved by using the 13 input variables selected by the stepwise regression method compared to the 24 input variables used. Moreover, the performance of the SVM classifier is the most outstanding. Part 3: After using the PCA method to transform the original 24 input variables, we selected the first 16 principal components based on previous studies. They can explain 95.67% of the original input variables. These principal components do not have a linear relationship. However, we found that the results of using these 16 principal components to classify financial statements are not satisfactory. The accuracy of almost all classifiers decreased compared to classifiers that use 24 input variables. Overall, the experimental results based on accuracy and F-score indicated that SVM is the top performer and the classification performance is improved more by stepwise regression rather than PCA.

This study has several theoretical implications. First, traditional statistical analysis methods for the detection of fraudulent financial statements are mostly flawed in terms of strict hypotheses and detection results [2,3]. Many previous studies using data mining techniques also have problems, such as the insufficient selection of predictors. In addition, most of the past fraud detection studies were not conducted in China [10,11]. Therefore, the results obtained from these studies cannot be directly used in China because of different cultural and socio-economic contexts. This paper adopted a multi-analytic method to detect fraudulent financial statements from listed companies in China, which is a study specifically for China's emerging capital market. Second, we made a comparative analysis of six data mining techniques in our study. The dataset consisting of 536 Chinese companies was analyzed using stand-alone techniques like SVM, CART, BP-NN, LR, Bayes and KNN. Moreover, 17 financial variables and 7 non-financial variables constitute the original feature set. The selected 24 variables comprehensively reflect the various aspects of a company, including financial distress, profitability, solvency, operational capability, development capability, management structure, internal control, shareholder structure and external audit. This is the second main strength that should be highlighted. Then, stepwise regression and PCA were used for dimensionality reduction, in which stepwise regression represents the feature selection technique and PCA represents the feature transformation technique. Then 13 significant original variables were selected with stepwise regression and 16 principal components were transformed with PCA. With the reduced feature subset, the classifiers SVM, CART, BP-NN, LR, Bayes and KNN were invoked again. We are not aware of much research that has also compared two different types of dimensionality reduction methods in financial statement fraud detection, however, such comparative analysis will help us to more accurately select the dimensionality reduction method for fraud detection in the future, thereby improving the efficiency of detection. This is the third main strength in this paper. Finally, we compared and analyzed the classification performance of six data mining techniques with different input variables based on five performance metrics, which are accuracy, recall (sensitivity), precision, F-score and area under the receiver operating characteristic curve (AUC).

This study highlights some practical implications for auditors, investors, and so on. First, detection of fraudulent financial statements is extremely important as it can save huge amounts of money from being embezzled. Our study is an important step in that direction and it highlights the use of data mining for solving this serious problem In our study, the combination of SVM and stepwise regression was found to be a good model for detecting FFS, with a classification of 0.8250. This optimal model could be of assistance to auditors, both internal and external, by saving a lot of audit time. Second, the use of the proposed multi-analytic approach also could be applied to the tax authorities or other government regulatory agencies, individual and institutional investors, stock exchange markets, law firms, credit scoring agencies and banking systems.

There are still shortcomings in this study. First, this study did not cover overall companies that listed in the Shanghai Stock Exchange and Shenzhen Stock Exchange. Future studies are thus encouraged to collect more sample to retest our model. In addition, the proposed multi-analytic approach in this paper may also require some necessary modification when it is applied to other countries or regions. Finally, there are more classification algorithms in the field of data mining than the six mentioned in the article. So, future studies could use other ensemble learning methods such as the random forest to predict FFS.

Author Contributions: Conceptualization, J.Y.; Methodology, Y.P.; Formal analysis, Y.P.; data Curation, Y.P.; Writing—original draft preparation, Y.P.; Writing—review and editing, S.Y., Y.C. and Y.L.; Supervision, J.Y.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 61502414. The APC was funded by Zhejiang Provincial Natural Science Foundation of China, grant number LY18G020013 and LY18G020014.

Acknowledgments: The authors would like to thank the editors and the anonymous reviewers of this journal. What is more, the authors express great thanks for the financial support from the National Natural Science Foundation of China and Zhejiang Provincial Natural Science Foundation of China.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jan, C.L. An effective financial statements fraud detection model for the sustainable development of financial markets: Evidence from Taiwan. *Sustainability* **2018**, *10*, 513. [CrossRef]
2. Kirkos, E.; Spathis, C.; Manolopoulos, Y. Data mining techniques for the detection of fraudulent financial statements. *Expert Syst. Appl.* **2007**, *32*, 995–1003. [CrossRef]
3. Chen, S. Detection of fraudulent financial statements using the hybrid data mining approach. *SpringerPlus* **2016**, *5*, 89–104. [CrossRef]
4. Ravisankar, P.; Ravi, V.; Raghava Rao, G.; Bose, I. Detection of financial statement fraud and feature selection using data mining techniques. *Decis. Support Syst.* **2011**, *50*, 491–500. [CrossRef]
5. Dutta, I.; Dutta, S.; Raahemi, B. Detecting financial restatements using data mining techniques. *Expert Syst. Appl.* **2017**, *90*, 374–393. [CrossRef]
6. Liu, C.W.; Chan, Y.X.; Alam Kazmi, S.H.; Fu, H. Financial fraud detection model: Based on random forest. *Int. J. Econ. Financ.* **2015**, *7*, 178–188. [CrossRef]
7. Kotsiantis, S.; Koumanakos, E.; Tzelepis, D.; Tampakas, V. Forecasting fraudulent financial statements using data mining. *Int. J. Comput. Int.* **2006**, *3*, 104–110.
8. Beasley, M.S. An empirical analysis of the relation between the board of director composition and financial statement fraud. *Account. Rev.* **1996**, *71*, 443–465.
9. Kim, Y.J.; Baik, B.; Cho, S. Detecting financial misstatements with fraud intention using multi-class cost-sensitive learning. *Expert Syst. Appl.* **2016**, *62*, 32–43. [CrossRef]
10. Glancy, F.H.; Yadav, S.B. A computational model for financial reporting fraud detection. *Decis. Support Syst.* **2011**, *50*, 595–601. [CrossRef]
11. Hajek, P.; Henriques, R. Mining corporate annual reports for intelligent detection of financial statement fraud—A comparative study of machine learning methods. *Knowl. Base Syst.* **2017**, *128*, 139–152. [CrossRef]
12. Huang, S.Y.; Tsaih, R.H.; Yu, F. Topological pattern discovery and feature extraction for fraudulent financial reporting. *Expert Syst. Appl.* **2014**, *41*, 4360–4372. [CrossRef]
13. Humpherys, S.L.; Moffitt, K.C.; Burns, M.B.; Burgoon, J.K.; Felix, W.F. Identification of fraudulent financial statements using linguistic credibility analysis. *Decis. Support Syst.* **2011**, *50*, 585–594. [CrossRef]
14. Huang, S.Y. Fraud detection model by using support vector machine techniques. *JDCTA* **2013**, *7*, 32–42.
15. Cerullo, M.J.; Cerullo, V. Using neural networks to predict financial reporting fraud: Part 1. *Comput. Fraud Secur.* **1999**, *1999*, 14–17.
16. Bell, T.B.; Carcello, J.V. A decision aid for assessing the likelihood of fraudulent financial reporting. *Audit. J. Pract. Theory* **2000**, *19*, 169–184. [CrossRef]
17. Zhou, W.; Kapoor, G. Detecting evolutionary financial statement fraud. *Decis. Support Syst.* **2011**, *50*, 570–575. [CrossRef]
18. Dong, W.; Liao, S.Y.; Fang, B.; Cheng, X.; Chen, Z.; Fan, W.J. The detection of fraudulent financial statements: An integrated language model approach. In Proceedings of the Pacific Asia Conference on Information Systems (PACIS 2014), Chengdu, China, 26–28 June 2014.
19. Lin, C.C.; Chiu, A.A.; Huang, S.Y.; Yen, D.C. Detecting the financial statement fraud: The analysis of the differences between data mining techniques and experts' judgments. *Knowl. Base Syst.* **2015**, *89*, 459–470. [CrossRef]
20. Yeh, C.C.; Chi, D.J.; Lin, T.Y.; Chiu, S.H. A hybrid detecting fraudulent financial statements model using rough set theory and support vector machines. *Cybernet. Syst.* **2016**, *47*, 261–276. [CrossRef]
21. Pai, P.F.; Hsu, M.F.; Wang, M.C. A support vector machine-based model for detecting top management fraud. *Knowl. Base Syst.* **2011**, *24*, 314–321. [CrossRef]
22. Williams, D.P.; Myers, V.; Silvious, M.S. Mine classification with imbalanced data. *IEEE Geosci. Remote. Sens. Lett.* **2009**, *6*, 528–532. [CrossRef]
23. Abbasi, A.; Albrecht, C.; Vance, A.; Hansen, J. Metafraud: A meta-learning framework for detecting financial fraud. *MIS Q.* **2012**, *36*, 1293–1327. [CrossRef]

24. Perols, J. Financial statement fraud detection: An analysis of statistical and machine learning algorithms. *Audit. J. Pract. Theory* **2011**, *30*, 19–50. [CrossRef]
25. West, J.; Bhattacharya, M. Mining financial statement fraud: An analysis of some experimental issues. In Proceedings of the 2015 IEEE 10th Conference on Industrial Electronics and Applications (ICIEA), Auckland, New Zealand, 15–17 June 2015.