Hindawi
Discrete Dynamics in Nature and Society
Volume 2020, Article ID 8865489, 15 pages
https://doi.org/10.1155/2020/8865489

Research Article

Detecting Falsified Financial Statements Using a Hybrid SM-UTADIS Approach : Empirical Analysis of Listed Traditional Chinese Medicine Companies in China

Ruicheng Yang and Qi Jiang

School of Finance, Inner Mongolia University of Finance and Economics, Hohhot 010070, China

Correspondence should be addressed to Ruicheng Yang; yang-ruicheng@163.com

Received 8 August 2020; Revised 10 October 2020; Accepted 27 October 2020; Published 21 November 2020

Academic Editor: Dehua Shen

By combining the similarity matching (SM) method with the utilities additives discriminates (UTADIS) method, we propose a hybrid SM-UTADIS approach to detect falsified financial statements (FFS) of listed companies. To evaluate the performance of this hybrid approach, we conduct experiments using the annual financial ratios of listed traditional Chinese medicine (TCM) companies in China. There are three stages in the detection procedure. First, we use the cosine similarity matching method to select matched companies for each considered company, derive the deviation data of each considered company as a sample dataset to capture the intrinsic law of the financial data, and further divide these into training and testing datasets for the next two stages. Second, we put the training dataset into the UTADIS to train the SM-UTADIS model. Finally, we use the trained SM-UTADIS model to classify the testing dataset and evaluate the performance of the proposed method. Furthermore, we use other approaches, such as single UTADIS and logistic and SM-logistic regression models, to detect FFS. By comparing these results to those of the hybrid SM-UTADIS approach, we find that the proposed hybrid approach greatly improves the accuracy of FFS detection.

1. Introduction

Falsified financial statements (FFS) are deliberate misstatements of material facts by management in a company's accounts with the aim of deceiving investors and creditors. FFS primarily consist of overstating profit, sales, or assets or understating liabilities, expenses, or losses [1,2]. Such illegitimate behaviours have a severe effect on the global economy because they significantly undermine the confidence of investors and creditors. Falsified financial statements have become a serious problem worldwide, especially in some fast-growing countries like China, where FFS often cause investor failure, such as huge losses.

With the current upsurge in FFS, there is an increasing demand for greater transparency and consistency and for more information to be incorporated in financial statements. Detecting FFS has attracted considerable attention from investors, creditors, regulators, academic researchers, etc. FFS detection has always been an important but complex task for accounting professionals, and this problem has been difficult for traditional internal audits to solve effectively. In fact, detecting FFS is a classification problem because we can classify FFS as a group and non-FFS as another group. Hence, there are many studies in the literature regarding FFS detection which introduce advanced techniques or construct formal models, such as statistical models, data mining techniques, and multicriteria decision models. The classic statistical models mainly include logistic regression models, discriminant analysis, and probit models. Among these models, logistic regression is the most widely used approach to detect FFS, and it was developed by statistician Cox [3]. Beasley [4] applies logistic regression to analyze 75 fraud and 75 nonfraud firms and derives that nonfraud firms have boards with significantly higher percentages of outside members than fraud firms. Ines et al. [5] explore fraud in financial statements using logistic regression and find that

performance pressure on managers is a factor leading to fraud in the financial statements. Hansen et al. [6] introduce a powerful generalized qualitative-response model, EGB2, to predict management fraud based on data developed by an international public accounting company; therefore, the EGB2 model mainly consists of Probit and logistic techniques. The results indicate a good predictive ability for both symmetric and asymmetric cost assumptions. In addition, Persons [7] uses logistic regression to predict fraudulent financial reporting. Spathis [8] uses logistic regression analysis estimated using financial ratios from companies to determine which ratios are related to FFS. Chen et al. [9] screen important variables using stepwise regression, and then they match logistic regression, support vector machine, and decision trees to construct classification models for comparison. Ye et al. [10] adopt a random forest approach to detect FFS by learning imbalanced data.

With the development of artificial intelligence, neural networks are developed rapidly and used in economic prediction problems. For example, Zhang et al. [11] use Long Short-Term Memory (LSTM) networks to predict stock price movement. The results show that the LSTM model outperforms other models with the best prediction accuracy. Also, neural networks have a better performance in FFS detection. Green and Choi [12] develop a neural network fraud classification model using endogenous financial data. A classification model from the learned behaviour pattern was applied to a test sample. During the preliminary stage of an audit, a financial statement classified as fraudulent signals an auditor to increase substantive testing. By combining feature selection and machine learning classification, Yao et al. [13] propose an optimized financial fraud detection model. Jan [14] finds that variables screened with an artificial neural network (ANN) and processed by CART yield the best classification results in the detection of financial statements fraud. Fanning and Cogger [15] use ANN to develop a model for detecting management fraud. Using publicly available predictors of fraudulent financial statements, they develop a model using eight variables with a high probability for detection. Pazarskis et al. [16] apply 30 financial ratios and several statistical tests to create a model that uses ratios as predictors in the analysis of financial statements for fraud. Temponeras et al. [17] present a new predictive model for fraud detection using a deep dense artificial neural network. Kirkos et al. [18] explore the effectiveness of data mining classification techniques in detecting companies that issue FFS. To identify factors associated with FFS, they investigate the performances of decision trees, neural networks, and Bayesian belief networks in the identification of fraud financial statements. Gupta and Gill [19] implement three data mining methodologies, a decision tree, naïve Bayesian classifier, and genetic programming, to detect FFS. The three data mining methods for the detection of financial statement fraud were compared on the basis of two important evaluation criteria: sensitivity and specificity.

Different from natural world data, financial statement data are often irregular and it is hard to capture their intrinsic law. To date, the statistical models and data mining techniques have not derived ideal results. Hence, many researchers borrowed multiple-criteria decision-making models to identify FFS. Multiple-criteria decision-making (MCDM) or multiple-criteria decision analysis (MCDA) is a subdiscipline of operations research that explicitly evaluates multiple conflicting criteria in decision-making (both in daily life and in settings such as business, government, and medicine); Zionts [20] popularized the acronym. The approach was first summarized comprehensively in a book by Roy Bernard [21]. The significant approach in MCDA is utilities additives (UTA) method, which is based on preference disaggregation that aims at the estimation of an additive utility function through the analysis of global judgments (ranking or grouping of alternatives) of decision-makers. Lagrèze and Siskos [22] assess the additive utility functions that aggregate multiple criteria in a composite criterion, using linear programming to estimate the parameters of the utility function. Siskos et al. [23] analyze the UTA method and its variants to summarize the progress made in this field. The UTA method is a well-known preference disaggregation method applied in many sorting problems. Furthermore, Corrente et al. [24] integrate the multiple-criteria hierarchy process and UTA method for dealing with MCDA in case of a hierarchical structure of the family of evaluation criteria. Mousseau et al. [25] consider the inverse multiple-criteria sorting problem (IMCSP) with UTA and other sorting methods for determining which actions to implement to provide guarantees on object classification. Mota [26] uses the approach to support project managers to focus on the main tasks of a project network.

Zopounidis and Doumpos [27] propose the UTADIS method based on the preference disaggregation approach and estimate a set of additive utility functions and utility profiles using linear programming techniques to minimize misclassification errors in sorting problems. They present the application of the UTADIS method in two real-world classification problems concerning the field of financial distress. Kosmidou et al. [28] use UTADIS to investigate the performance of small and large UK banks over multiple criteria, such as asset quality, capital adequacy, liquidity, and efficiency/profitability. The results determine the key factors that classify a bank as small or large and provide us with responsible banking decision-makers for future readjustments. Mehregan et al. [29] use the UTADIS method to classify securities and to form a profitable investment portfolio. Doumpos et al. [30] propose a robust multicriteria approach that can be used to provide early warning signals for possible future capital shortfalls that banks may face. These research results show that the proposed MCDA approach provides models with strong discriminative power. Recently, Spathis et al. [31] apply UTADIS classification method to detect factors associated with FFS; a jackknife procedure approach is used for model validation and comparison with multivariate statistical techniques, namely, discriminant and logistic analysis. The results indicate that the UTADIS methodology achieves relatively good results in detecting FFS.

Based on this, we borrow the UTADIS idea to detect the FFS of companies. The sample data are chosen from the

financial ratios of listed traditional Chinese medicine (TCM) companies in China, which is a historical and prosperous industry. The reasons why we choose this industry as our research sample are the following: (1) There is a necessary sample size of FFS for our research in this sector. (2) There are few mixed businesses in TCM industry, and the main business of this sector is relatively concentrated. This will ensure that the selected samples have the homogeneous feature in their main business. Accordingly, this also can reduce the interference of unrelated noises.

As we know, in the real world, the data of each financial ratio may change drastically over time. For example, the outbreak of an epidemic will raise the income of almost all of the companies in the TCM sector, whereas an increase in material costs will result in a decline in that sector. Accordingly, related financial ratios will change sharply, inducing FFS misjudgments. However, we observe that the operating performance of a company is usually similar to other companies in the same sector; therefore, such companies should have similar changes in their financial ratios. In view of this, we introduce the cosine similarity algorithm to help us select companies most similar to the matched companies and use their financial data to compute the deviation of the considered company. Then, the deviation data are used for UTADIS classification (more details in Section 2). The merit of the financial deviation data is that they reflect the intrinsic law of a considered company, making it easier to detect FFS with UTADIS. This is the main contribution of this paper, that is, based on the UTADIS method, we combine the similarity algorithm with UTADIS and formulate an integrated method, SM-UTADIS, for detecting FFS.

The remainder of this paper is organized as follows. In Section 2, we describe the proposed SM-UTADIS methodology, including the similarity computation, UTADIS method, and classification procedure. Section 3 provides the results obtained using the SM-UTADIS classification method and reports the comparisons with the single UTADIS and logistic regression approaches. Concluding remarks and opportunities for future research are presented in Section 4.

2. Model Description

For convenience, we first introduce the following notations that are used throughout the paper:

(1) Considered company A_c ($c = 1, 2, \ldots, C$): the c-th considered company whose annual financial data we classify into FFS and non-FFS groups.

(2) Candidate matching company $\#m$ ($m = 1, 2, \ldots, L$): the m-th candidate matching company.

(3) Matched company $\#M_q$ ($q = 1, 2, 3, \ldots, Q$): the q-th matched company with $Q \le L$; these matched companies are selected from the above candidate matching companies.

(4) Variables or financial ratios R_i^l ($i = 1, 2, \ldots, J$; $l = 1, 2, \ldots, C$ or L) represent the i-th variable or financial ratio of company $\#l$. In this paper, we choose the same financial ratios for all companies (including considered and candidate matching companies), and each different right superscript represents a different company. In total, there are J financial ratios for each company.

We propose a hybrid classification model that combines the SM and UTADIS methods, as illustrated in Figure 1. As shown in Figure 1, the procedure has three stages. In the SM stage, the cosine similarity matching algorithm is applied to select matched companies for each considered company. Using the initial financial data, we compute the deviation data of each considered company and gather all of the deviation financial data into the research sample. We further divide the research sample into training and testing datasets for the next two stages. In the second stage, we put the training dataset into the UTADIS to train the model. In the last stage, we put the testing data into the well-trained UTADIS model to predict the testing data and evaluate the classification performance of the proposed method.

In the detection procedure, the key algorithms are the cosine SM method in Stage 1 and UTADIS in Stages 2 and 3. Thus, we provide more explicit descriptions of the algorithms in the two following subsections.

2.1. Similarity Matching Method. The operating performance of a company is usually similar to other companies in the same industry. Therefore, there should be similar changes in the financial ratios of these companies. In view of this, for each considered company, we use the cosine similarity matching algorithm to select the most similar matched companies and to obtain the deviation data of each considered company. Without loss of generality, the algorithm for considered company A_c is as follows:

Step 1—Computation of cosine similarity: Cosine similarity is a measure of the similarity between two vectors of an inner product space that measures the cosine of the angle between them [32]. Here, we give the cosine similarity between considered company A_c and the candidate matching company $\#m$ ($m = 1, 2, \ldots, L$) as follows:

$$\begin{aligned} S_m^c &= \cos\left(\theta_m^c\right) \\ &= \frac{\vec{R^c} \cdot \vec{R^m}}{\left\|\vec{R^c}\right\|\left\|\vec{R^m}\right\|} \\ &= \frac{\sum_{j=1}^{J} R_j^c \times R_j^m}{\sqrt{\sum_{j=1}^{J}\left(R_j^c\right)^2}\sqrt{\sum_{j=1}^{J}\left(R_j^m\right)^2}}, \quad (m = 1, 2, \ldots, L), \end{aligned} \tag{1}$$

where S_m^c represents the similarity between considered company A_c and candidate matching company $\#m$ ($m = 1, 2, \ldots, L$), $\vec{R^c} = (R_1^c, R_2^c, \ldots, R_J^c)$ represents the financial ratio vector of considered company A_c, and

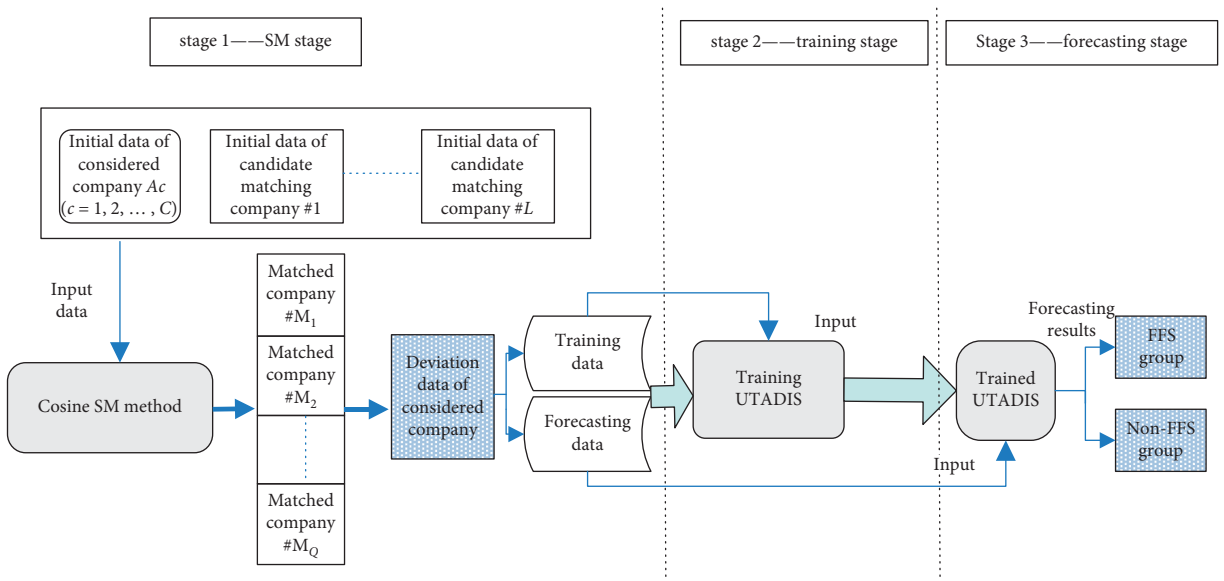

Figure 1: The procedure of FFS detection with the SM-UTADIS method.

$\overrightarrow{R^m} = (R_1^m, R_2^m, \ldots, R_J^m)$ represents the financial ratio vector of candidate matching company #m.

Step 2—Selection of matched companies: Now we rank the matched companies $\{S_m^c, m = 1, 2, \ldots, L\}$ in descending order with $\widehat{S}_1^c \geq \widehat{S}_2^c \geq \cdots \geq \widehat{S}_L^c$ and choose the first several companies as the matched companies. For example, if we choose P $(P \leq L)$ matched companies, we only choose companies with $\widehat{S}_1^c, \widehat{S}_2^c, \ldots, \widehat{S}_P^c$, and denote the P companies as the matched companies for considered company A_c.

Step 3—Computation of data deviation: Denote $\widehat{R}_j^p$ as the j-th financial ratio of the p-th $(p = 1, 2, \ldots, P)$ matched company, and using the corresponding financial data of the matched companies, we derive the deviation data of the j-th financial ratio for considered company A_c as follows:

$$\overline{R_j^c} = \frac{\left| R_j^c - (1/P) \sum_{p=1}^{P} \widehat{R}_j^p \right|}{(1/P) \sum_{p=1}^{P} \widehat{R}_j^p}. \tag{2}$$

Thus, we derive the deviation sample data for inputting into the UTADIS.

2.2. UTADIS Method. Following Zopounidis and Doumpos [27] and Spathis et al. [31], we give a brief description of the UTADIS method as follows.

Let $A = (a_1, a_2, \ldots, a_n)$ be a set of n annual financial datasets described along a set of m attributes or evaluation criteria $x_1, x_2, \ldots, x_J$; here, the attributes correspond to financial ratios. The goal is to classify the n annual financial datasets into q ordered classes $C_1, C_2, \ldots, C_q$, which are defined as $C_1 \succ C_2 \succ \cdots \succ C_q$ (C_1 is preferred to C_2, C_2 is preferred to C_3, and so on).

For each evaluation criterion x_j $(j = 1, 2, \ldots, J)$, the interval $X_j = [x_j^{\min}, x_j^{\max}]$ of its values is defined; here $x_j^{\min}$ and $x_j^{\max}$ represent the minimal and maximal values, respectively, of criterion X_j for all of the alternatives belonging to A. The internal X_j can be divided into a_{j-1} equal intervals $[x_j^i, x_j^{i+1}]$, $i = 1, 2, \ldots, a_{j-1}$, $x_j^1 = x_j^{\min}$, $x_j^{a_j} = x_j^{\max}$. a_j depends on the number of estimate points of the marginal utility u_j. Every break point x_j^i can be determined using the following formula:

$$x_j^i = x_j^{\min} + \frac{i-1}{a_j - 1}\left(x_j^{\max} - x_j^{\min}\right). \tag{3}$$

Our aim is to estimate the marginal utilities at each of these breakpoints. Suppose that the evaluation of each alternative a on the criterion x_j is $x_j(a) \in [x_j^i, x_j^{i+1}]$, and the marginal utility of each alternative $a \in A$ and $u_j[x_j(a)]$ can be roughly estimated through the linear interpolation:

$$u_j\left[x_j(a)\right] = u_j\left(x_j^i\right) + \frac{x_j(a) - x_j^i}{x_j^{i+1} - x_j^i}\left(u_j\left(x_j^{i+1}\right) - u_j\left(x_j^i\right)\right). \tag{4}$$

To achieve monotonicity of the criteria, the following conditions and the monotonicity constraints must be satisfied:

$$\left.\begin{aligned} & u_j\left(x_j^{i+1}\right) - u_j\left(x_j^i\right) \geq 0, \quad \forall j, \\ & \omega_{ji} = u_j\left(x_j^{i+1}\right) - u_j\left(x_j^i\right) \geq 0, \quad \forall i, j \\ & u_j\left(x_j^{\min}\right) = 0 \\ & u_j\left(x_j^i\right) = \sum_{k=1}^{i-1} \omega_{jk} \end{aligned}\right\}. \tag{5}$$

Using these transformations, (4) can be rewritten as

$$u_j\left[x_j(a)\right] = \sum_{k=1}^{i-1} \omega_{ji} + \frac{x_j(a) - x_j^i}{x_j^{i+1} - x_j^i}\omega_{ji}. \tag{6}$$

The total utility $U(a)$ of each alternative $a \in A$ can be expressed as

$$U(a) = \sum_{j=1}^{m} u_j\left(x_j(a)\right) \in [0,1]. \tag{7}$$

Estimations of the total utility model (marginal utilities of all breakpoints $x_j^i (i = 1, 2, \ldots, a_{j-1})$) and utility thresholds are accomplished through the solution of the following linear program:

$$\begin{aligned} \min F = & \sum_{a \in C_1} \sigma^+(a) + \cdots + \sum_{a \in C_k} \left[\sigma^+(a) + \sigma^-(a)\right] \\ & + \cdots + \sum_{a \in C_q} \sigma^-(a), \end{aligned} \tag{8}$$

subject to

$$\begin{aligned} & \sum_{j=1}^{m} u_j\left[x_j(a)\right] - t_1 + \sigma^+(a) \geq 0, \quad \forall a \in C_1, \\ & \left.\begin{aligned} & \sum_{j=1}^{m} u_j\left[x_j(a)\right] - t_{k-1} - \sigma^-(a) \leq -\delta \\ & \sum_{j=1}^{m} u_j\left[x_j(a)\right] - t_k + \sigma^+(a) \geq 0 \end{aligned}\right\}, \quad \forall a \in C_k, \\ & \sum_{j=1}^{m} u_j\left[x_j(a) - t_{q-1} - \sigma^-(a)\right] \leq -\delta, \quad \forall a \in C_q, \\ & \sum_{j=1}^{m} \sum_{i=1}^{a_{j-1}} \omega_{ji} = 1, \\ & t_{k-1} - t_k \geq \delta, \quad k = 2, 3, \ldots, q-1, \\ & \omega_{ji} \geq 0, \sigma^+(a) \geq 0, \sigma^-(a) \geq 0. \end{aligned} \tag{9}$$

Here, $\sigma^+(a)$ and $\sigma^-(a)$ are the two possible errors (misclassification errors) relative to the global utility $U(a)$; an overestimation error $\sigma^+(a)$ represents cases in which an alternative, according to its utility, is classified in a lower class than the class to which it belongs (e.g., an alternative is classified in class C_2 while belonging to class C_1), whereas an underestimation error $\sigma^-(a)$ represents cases in which an alternative, according to its utility, is classified in a higher class than the class to which it belongs. The threshold t_k is used to denote the strict preference relation between the utility thresholds that distinguish the classes; δ>0 is used to denote the strict preference relation between the utility thresholds that distinguish the classes.

By comparing each utility with the corresponding utility thresholds $t_k (t_1 > t_2 > \cdots > t_{q-1})$, we derive a decision rule for each alternative a to distinguish each class from the others:

$$\begin{aligned} & U(a) \geq t_1 \Longrightarrow a \in C_1, \\ & t_2 \leq U(a) < t_1 \Longrightarrow a \in C_2, \\ & \qquad \cdots \\ & t_k \leq U(a) < t_{k-1} \Longrightarrow a \in C_k, \\ & \qquad \cdots \\ & U(a) < t_{q-1} \Longrightarrow a \in C_q. \end{aligned} \tag{10}$$

Next, we examine the detection of FFS. In this study, only two classes of annual financial samples are considered, that is, non-FFS (group C_1) and FFS (group C_2), and the rule for the classification of a sample as FFS or non-FFS is as follows:

$$\begin{aligned} & U(a) \geq t \Longrightarrow a \in C_1, \\ & U(a) < t \Longrightarrow a \in C_2, \end{aligned} \tag{11}$$

where t is the corresponding utility threshold.

Based on the above classification rule, we classify the data into two classes: non-FFS and FFS. Here, the FFS class is the fraudulent financial data.

3. Experiment Results and Discussion

In this section, using the real financial data of the TCM sector in China, we evaluate the performance of the proposed SM-UTADIS approach. The computation results of this section are obtained using Matlab software.

3.1. Selection of Fraud Companies and Nonfraud Companies Experiment Results and Discussion. Currently, there are about 150 companies listed in the TCM sector in China, but most are involved in mixed business areas, and the main profit of some is not earned through traditional Chinese medicine. Such companies must be discarded; otherwise, they will obscure the evolving law of financial ratios as it relates to companies whose business is purely related to TCM. In addition, we must choose companies with financial ratios that include falsified data, but too much non-FFS data will dilute and hinder the identification process. Hence, only 24 TCM companies are used in our research. Among these 24 companies, three considered companies are regarded as fraud companies as they were accused of fraud in some years by the China Securities Regulatory Commission (CSRC). The other 21, as the candidate matching companies, are non-FFS, which are free of fraud. Of course, there is at least one annual financial data point in the fraudulent statements. For simplicity, we label the three considered companies as A1, A2, and A3, and the other 21 non-FFS candidate matching companies are labelled #1, #2, . . ., #21.

Next, we use the annual financial data of the three considered companies to evaluate the classification performance of the proposed SM-UTADIS method. The annual financial data cover the period from 2001 to 2016, and the data are collected from the Wind website (http://www.wind.com.cn/). If a company's financial statement in a specific year is identified as fraudulent by the CSRC, it is classified as a fraudulent observation. In contrast, financial statements that are free from falsified allegations are classified as nonfraudulent observations.

For each falsified company, we first identify the earliest year in which financial statement fraud was committed. Each period covers the years before and after the year of the event. Thus, seven consecutive annual financial statements are used in most cases except for some class in which consecutive annual financial statements are accused of fraud or the related data are not published. We get 36 firm-year

observations (i.e., annual financial statements) of the three considered companies as our research sample, out of which 24 are nonfraudulent (Class C_1) and 12 are fraudulent (Class C_2). Next, we divide these 36 annual observations into two groups: a training dataset and a testing dataset. To get better training and testing effects, the proportion of training and testing data is set to 1 : 1, respectively. Moreover, to maintain the rationality and validity of the division, we try to distribute the data of each company into the training and testing datasets as equally as possible. Therefore, the 6 falsified and 12 nonfalsified annual observations are treated as the training dataset, and the rest are treated as the testing dataset.

3.2. Choice of Financial Ratios. Based on Green and Choi [12], Mironiuc et al. [33], and Shin-Ying Huang et al. [34], 12 explanatory variables or financial ratios are selected as the sample variables; the definitions and measurements of these financial ratios (financial ratios that describe both the structure of the company assets and the level of the recorded performance care) are summarized in Table 1.

3.3. Similarity Computation and Matched Company Selection of Financial Ratios. For classification purposes, we match each falsified considered company with nonfalsified candidate matching companies in the same sector using cosine similarity analysis. In fact, we only need to compute the similarity of nonfraudulent years between the considered company and its candidate matching companies from the training dataset. If fraudulent data were included in the similarity computation, it would decrease identification efficiency because it would distort characteristics that are similar in the real world. Hence, for each considered company A_i $(i = 1, 2, 3)$, we select its nonfalsified annual financial data from the training dataset and the corresponding data of non-FFS candidate matching companies #1–#21 in the same years, and, using (1), we can compute the similarity between considered company (CM) A_c and its non-FFS candidate matching companies (CMC) #1–#21. The results of the similarity analysis are given in the following table.

Choosing the similarity threshold is the key issue for improving classification accuracy in the following training and forecasting stages. If the threshold value is too big, the number of matched companies will be small. However, if the threshold value is too small, the number of matching companies will become lagger. In fact, the threshold value will directly affect the selection of matched companies, and this will further affect the accuracy of the training and testing results. We hope to choose a suitable threshold that will allow for ideal training and testing accuracy. Through many trials, the 0.70 threshold value provides the best performance. After many trials and adjustments, we select our matched companies, and the similarity values are greater than 0.70. In Table 2, the first three maximal values are highlighted in grey for each considered company. There are two matched companies, #3 and #14, for considered company A_1, three matched companies, #1, #9, and #20, for considered company A_2, and one matched company, #18, for considered company A_3. Based on the initial financial data of each considered company and its matched companies, by (2), we can easily get the deviation data of all considered companies and further divide the data into a training dataset and a testing dataset. This concludes the data preparation for the next two stages.

3.4. Results and Discussion. Applying the proposed SM-UTADIS method to the training dataset, we get the marginal utility of each financial ratio as shown in Figure 2. The classification results and the utility threshold t are shown in Table 3.

In Figure 2, we see that the most significant ratios for discrimination in the training dataset are R_9, R_{10}, and R_{12}; their weights are 27.4971%, 23.7773%, and 12.8746%, respectively. The next is R_1 with a weight of 7.3757%. The other ratios show no significant differences in their contribution to FFS detection. Table 3 shows that the threshold t is 0.349489. Using classification rules (10), Table 3 shows that there are no misclassifications.

Furthermore, the prediction ability of the trained model developed by the UTADIS method is also tested using the testing dataset. Using the trained model, we derive classification results for the testing dataset. The results are presented in Table 4. To make it clear, the misclassifications are highlighted in grey. There are two misclassifications in the testing dataset; we summarize the type I error, type II error, and overall error in Table 5. Here, a type I error corresponds to an overestimation error $\sigma^-(a)$, meaning that an FFS observation is classified as non-FFS, whereas a type II error corresponds to an underestimation error $\sigma^+(a)$, meaning that a non-FFS observation is classified as FFS. According to the results in Table 5, the overall error rate is 11.1111%, and type I and type II errors are 16.6667% and 8.3333%, respectively.

3.5. Comparison with Single UTADIS Results and Discussion. To evaluate the performance of the proposed SM-UTADIS approach, we compare its classification results with those of single UTADIS using the same initial data of considered company A_c $(c = 1, 2, 3)$. Figure 3 illustrates the marginal utility of each financial ratio. Similar to the SM-UTADIS analysis, Figure 3 shows that the most significant ratios for discrimination in the training dataset are also R_9 and R_{10}, and their weights change to 24.3604% and 19.2855%, respectively. The other ratios show no significant differences in their contribution to FFS detection. Table 6 shows that there is no identification error in the training process. Using the trained model, we predict the testing dataset, and the classification results are shown in Table 7. Table 7 shows that there are 13 misclassification errors. The type I, type II, and overall errors are summarized in Table 8. Compared with Table 5, the classification results using the proposed SM-UTADIS are far superior to the results with single UTADIS.

3.6. Comparison of Logistic and SM-Logistic Models. Logistic regression is another popular method for FFS detection; it is widely used in many research areas, such as finance and social sciences. To test the performance of our

Table 1: Definition and measurement of financial ratios.

Notation	Definition of ratio	Measurement
R_1	Return on equity (ROE)	Net income/Average equity
R_2	Earnings before interest and taxes to return on assets (EBIT ROA)	Earnings before interest and taxes/Average total assets
R_3	Return on assets	[Net income + interest * (1 − tax rate)]/Total assets
R_4	Net profit to total operating income	Net profit/Total operating income
R_5	Operating profit to total operating income	Operating profit/Total operating income
R_6	Operating profit ratio	(Sales − Operating Costs − Operating expenses)/Sales
R_7	Current ratio	Current assets/Current liabilities
R_8	Quick ratio	(Current assets − Inventory − Prepaid expenses)/Current liabilities
R_9	Growth rate of net profit	(Net profit/Net profit in prior annual term) − 1
R_{10}	Growth rate of net assets	(Current net assets/Net assets in prior annual term) − 1
R_{11}	Total assets turnover ratio	Revenue/Average total assets
R_{12}	Ratio of liabilities to assets	Total debts/Total assets

Table 2: Similarity between considered company A_i ($i = 1, 2, 3$) and non-FFS candidate matching companies #1–#21.

CM	CMC	Similarity value
A_1	#1	−0.00493
	#2	−0.24823
	#3	0.927066
	#4	0.219229
	#5	−0.15931
	#6	0.128307
	#7	0.30172
	#8	0.352593
	#9	0.103669
	#10	−0.04005
	#11	−0.09919
	#12	−0.89803
	#13	0.325353
	#14	0.786474
	#15	0.474157
	#16	0.402147
	#17	0.022846
	#18	0.398805
	#19	0.034669
	#20	−0.023900
	#21	0.205393
A_2	#1	0.719112
	#2	0.235432
	#3	−0.099010
	#4	0.559809
	#5	0.129897
	#6	0.435574
	#7	0.632609
	#8	0.251317
	#9	0.705876
	#10	0.647896
	#11	0.537107
	#12	0.167555
	#13	0.378416
	#14	0.109826
	#15	0.052887
	#16	0.055193
	#17	0.670442
	#18	−0.195400
	#19	0.586637
	#20	0.781885
	#21	0.550248
A_3	#1	−0.06277
	#2	0.101022
	#3	0.301448
	#4	−0.07407
	#5	−0.002910
	#6	−0.733340
	#7	−0.101370
	#8	0.031707
	#9	−0.022590
	#10	−0.073700
	#11	−0.489810
	#12	−0.297000
	#13	−0.808090
	#14	0.147712
	#15	−0.043480
	#16	−0.019510
	#17	0.074003
	#18	0.891208
	#19	−0.354680
	#20	−0.036140
	#21	−0.012560

proposed SM-UTADIS method, we use logistic regression and SM-logistic regression (a combination of SM and logistic regression) to classify the same training and testing datasets and further compare the classification results with those of SM-UTADIS. The results of logistic and SM-logistic regression are presented in Tables 9 and 10, respectively. Comparing Table 9 with Table 10, we see fewer classification errors with the SM-logistic regression method than with single logistic regression; this implies that the SM technique improves the classification accuracy rate. However, the classification result of the SM-logistic regression method is not better than that of the SM-UTADIS method. Tables 5 and 10 show that the type I, type II, and overall errors with SM-logistic regression are far higher than those with SM-UTADIS (see Table 5). Therefore, by comparing the classification results of three approaches, we find that the superiority of the SM-UTADIS method over logistic regression and single regression is clear, whether classifying the training dataset or the testing dataset.

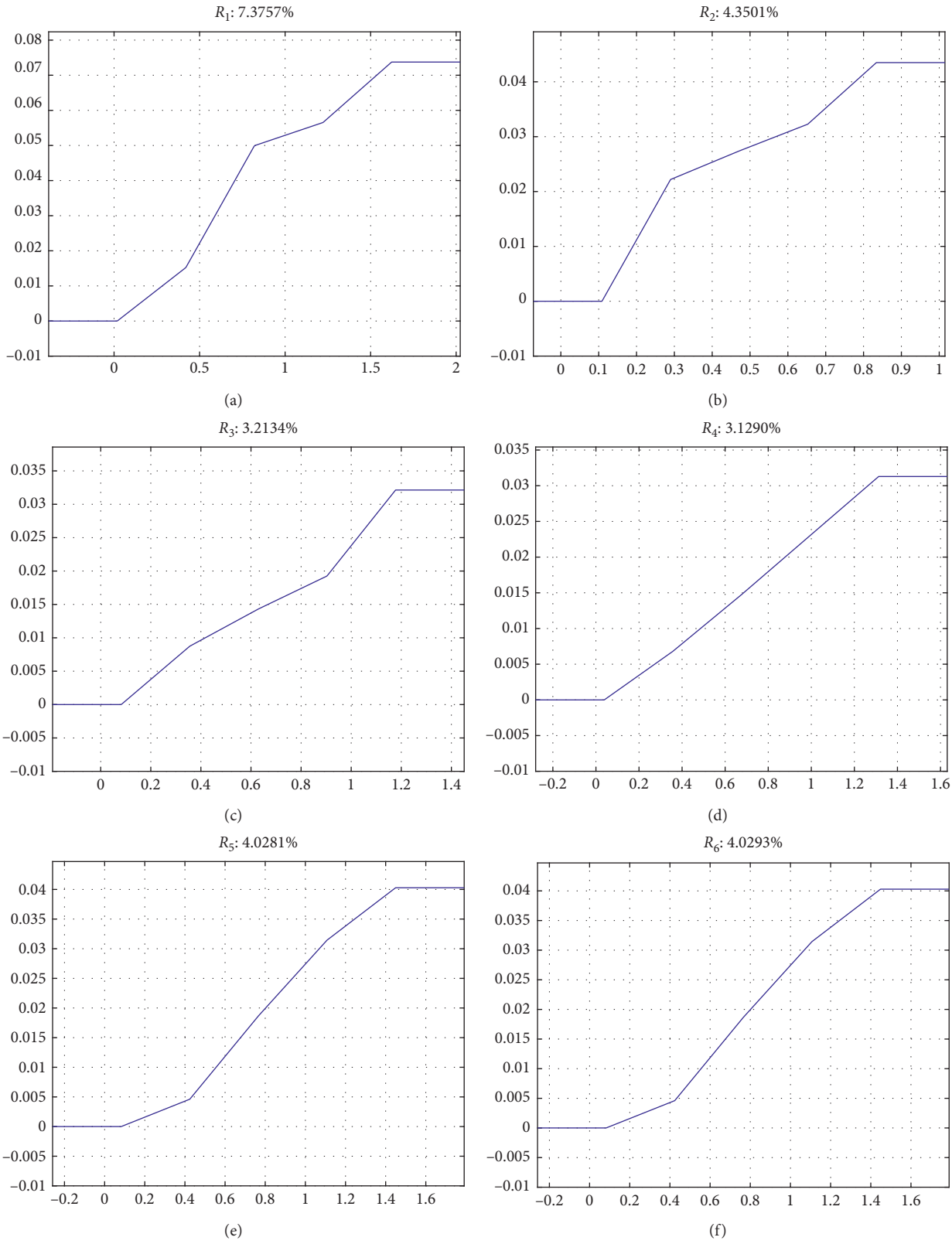

Figure 2: Continued.

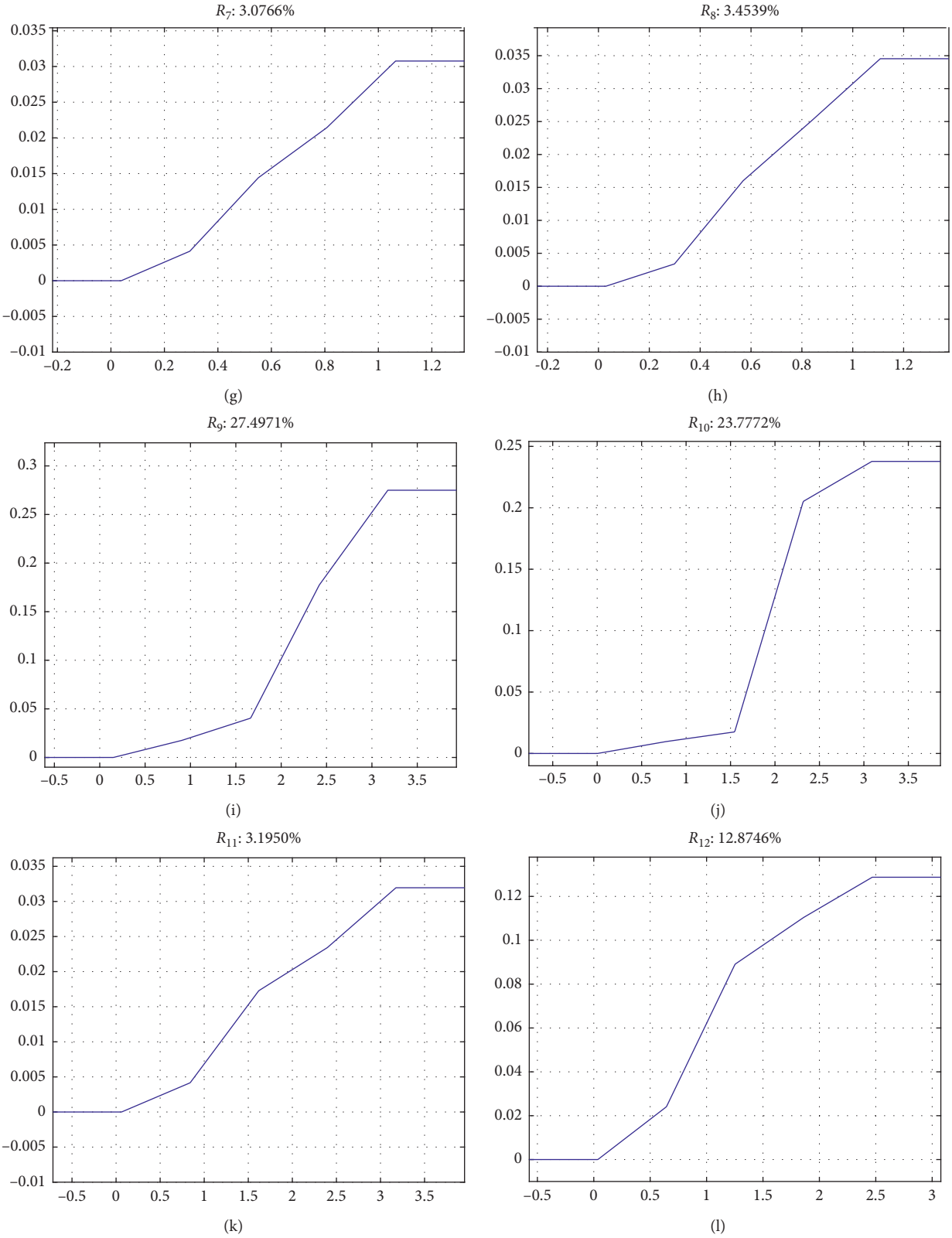

Figure 2: Marginal utility of each financial ratio with the SM-UTADIS method (training dataset).

Table 3: Classification results with the SM-UTADIS method (training dataset).

Considered company	Year	Actual class	Utility value	Estimated class
A_1	2001	C_1	0.419078	C_1
	2003	C_1	0.349516	C_1
	2009	C_1	0.358020	C_1
	2011	C_1	0.381929	C_1
A_2	2003	C_1	0.387177	C_1
	2005	C_1	0.358557	C_1
	2008	C_1	0.375054	C_1
	2013	C_1	0.413359	C_1
A_3	2003	C_1	0.350032	C_1
	2005	C_1	0.350784	C_1
	2010	C_1	0.354496	C_1
	2012	C_1	0.396284	C_1
Utility threshold t			0.349489	
A_1	2005	C_2	0.341412	C_2
	2007	C_2	0.328442	C_2
A_2	2004	C_2	0.323764	C_2
	2011	C_2	0.339731	C_2
A_3	2006	C_2	0.349420	C_2
	2008	C_2	0.346697	C_2

Table 4: Forecasting results with the trained SM-UTADIS model (testing dataset).

Considered company	Year	Actual class	Utility value	Estimated class
A_1	2002	C_1	0.349590	C_1
	2004	C_1	0.349710	C_1
	2010	C_1	0.358811	C_1
	2012	C_1	0.386387	C_1
A_2	2005	C_1	0.350307	C_1
	2006	C_1	0.380931	C_1
	2008	C_1	0.357234	C_1
	2014	C_1	0.312229	C_2
A_3	2004	C_1	0.352612	C_1
	2009	C_1	0.350430	C_1
	2011	C_1	0.351601	C_1
	2014	C_1	0.365781	C_1
Utility threshold t			0.349489	
A_1	2006	C_2	0.316715	C_2
	2008	C_2	0.353189	C_1
A_2	2010	C_2	0.325583	C_2
	2012	C_2	0.339748	C_2
A_3	2007	C_2	0.345021	C_2
	2013	C_2	0.050622	C_2

Table 5: Error summary with the SM-UTADIS method.

	Actual class	Total amounts	Number of errors identified	Type I errors	Type II errors	Overall errors
Training dataset	C_1	12	0	0	0	0
	C_2	6	0			
Testing dataset	C_1	12	1	16.6667%	8.3333%	11.1111%
	C_2	6	1			

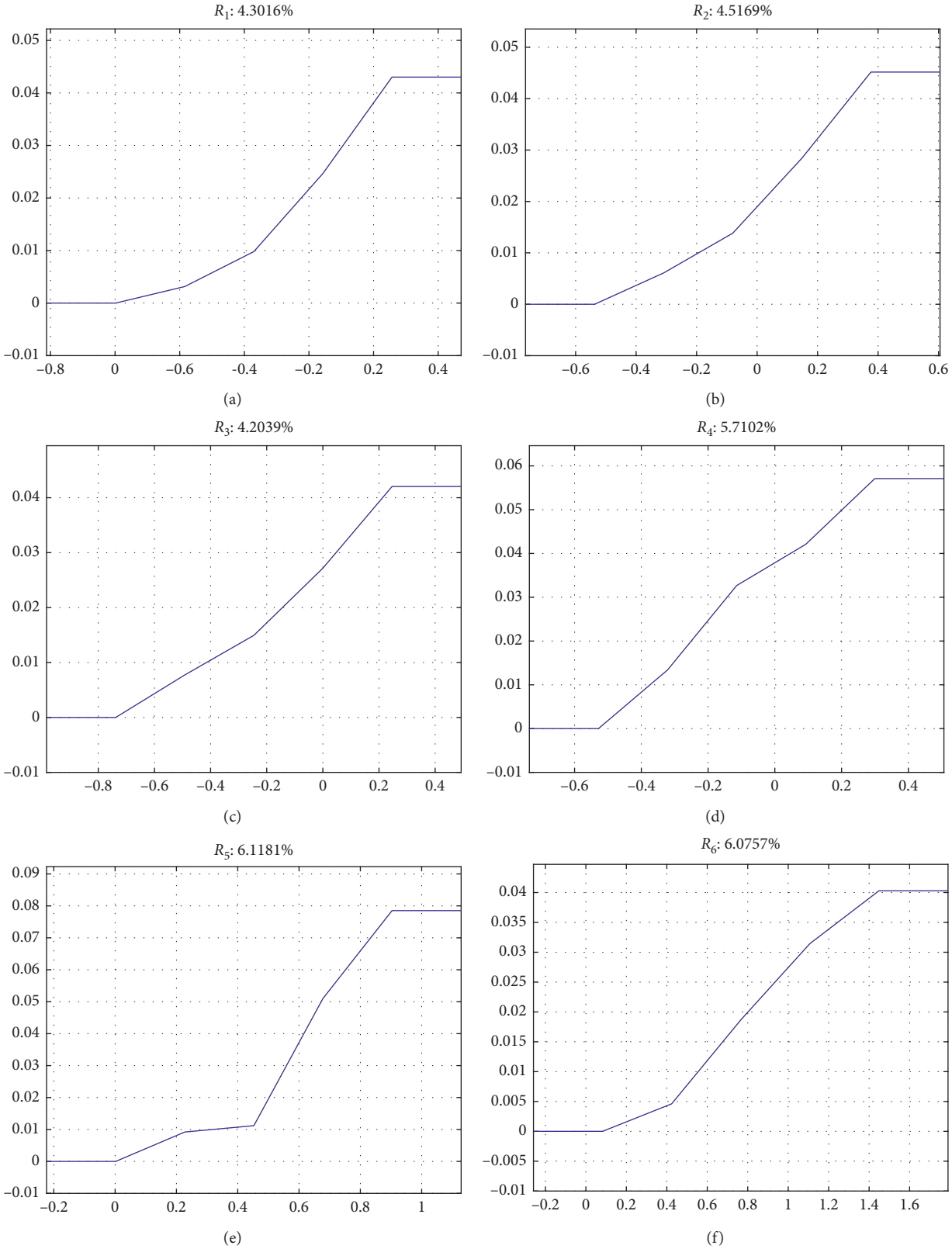

Figure 3: Continued.

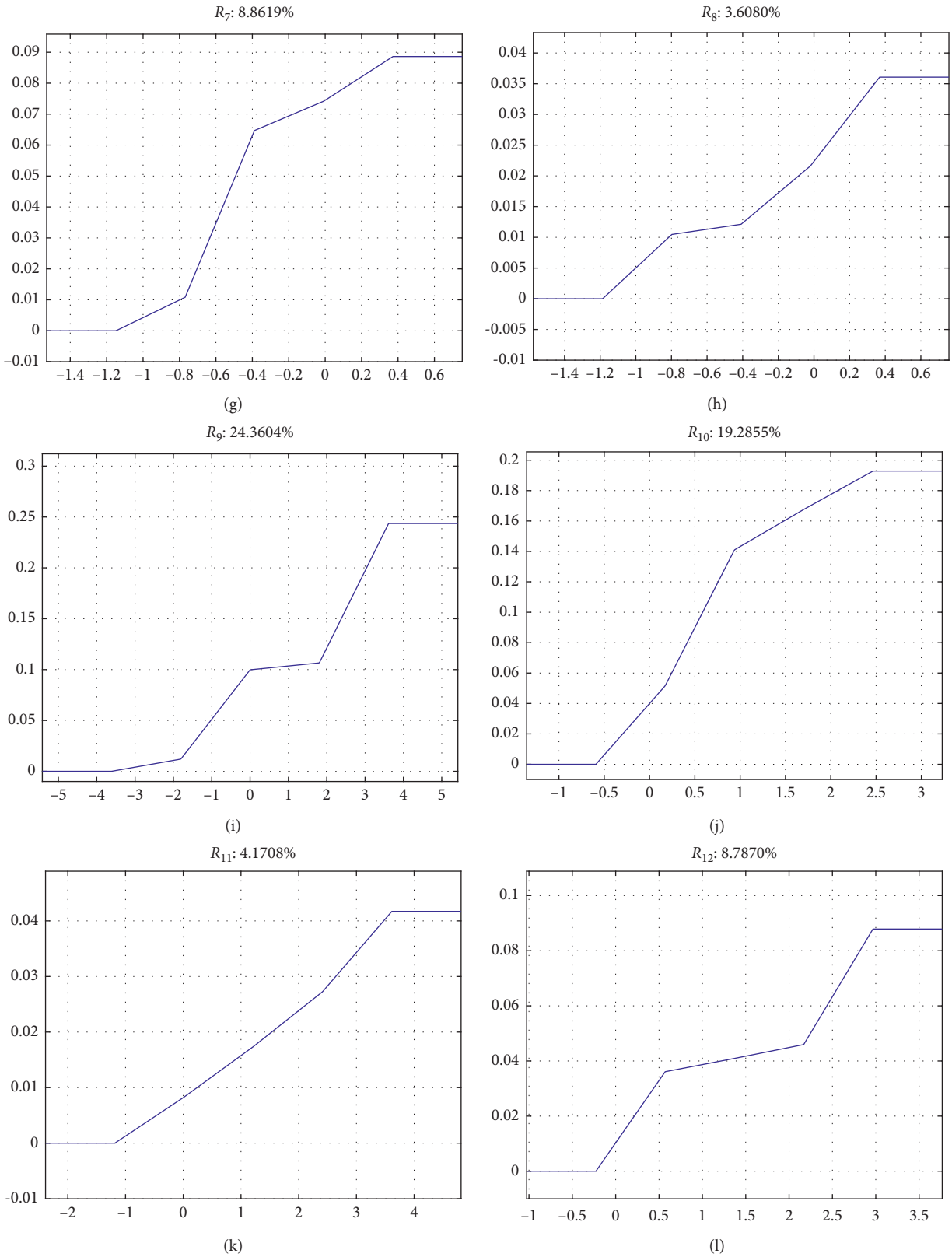

Figure 3: Marginal utility of each financial ratio using single UTADIS (training dataset).

Table 6: Classification results using the single UTADIS method (training dataset).

Considered company	Year	Actual class	Utility value	Estimated class
A_1	2001	C_1	0.337953	C_1
	2003	C_1	0.360633	C_1
	2009	C_1	0.422805	C_1
	2011	C_1	0.459080	C_1
A_2	2003	C_1	0.442766	C_1
	2005	C_1	0.360602	C_1
	2008	C_1	0.334531	C_1
	2013	C_1	0.369313	C_1
A_3	2003	C_1	0.392187	C_1
	2005	C_1	0.460778	C_1
	2010	C_1	0.319150	C_1
	2012	C_1	0.347777	C_1
Utility threshold t			0.319139	
A_1	2005	C_2	0.304669	C_2
	2007	C_2	0.283387	C_2
A_2	2004	C_2	0.290913	C_2
	2011	C_2	0.261104	C_2
A_3	2006	C_2	0.303683	C_1
	2008	C_2	0.316627	C_2

Table 7: Forecasting results using the trained single UTADIS model (training dataset).

Considered company	Year	Actual class	Utility value	Estimated class
A_1	2002	C_1	0.429631	C_1
	2004	C_1	0.474130	C_1
	2010	C_1	0.441629	C_1
	2012	C_1	0.437099	C_1
A_2	2005	C_1	0.303503	C_2
	2006	C_1	0.352361	C_1
	2008	C_1	0.382830	C_1
	2014	C_1	0.269696	C_2
A_3	2004	C_1	0.444084	C_1
	2009	C_1	0.280839	C_2
	2011	C_1	0.347777	C_1
	2014	C_1	0.469747	C_1
Utility threshold t			0.319139	
A_1	2006	C_2	0.285334	C_2
	2008	C_2	0.431728	C_1
A_2	2010	C_2	0.293083	C_2
	2012	C_2	0.363316	C_1
A_3	2007	C_2	0.363351	C_1
	2013	C_2	0.428283	C_1

Table 8: Error summary using single UTADIS method.

	Actual class	Total amounts	Number of errors identified	Type I errors	Type II errors	Overall errors
Training dataset	C_1	12	0	0	0	0
	C_2	6	0			
Testing dataset	C_1	12	3	66.67%	25%	38.8889%
	C_2	6	4			

Table 9: Error summary with single logistic regression method.

Data	Actual class	Total amount	Number of errors identified	Type I errors (%)	Type II errors (%)	Overall errors (%)
Training dataset	C_1	12	1	50	8.3333	22.2222
	C_2	6	3			
Testing dataset	C_1	12	2	83.3333	16.6667	38.8889
	C_2	6	5			

Table 10: Error summary with SM-logistic regression method.

Data	Actual class	Total amount	Number of errors identified	Type I errors	Type II errors (%)	Overall errors (%)
Training dataset	C_1	12	1	0	8.3333	5.5556
	C_2	6	0			
Testing dataset	C_1	12	3	33.3333%	25.0000	27.7778
	C_2	6	2			

4. Conclusions

Combining the SM method with UTADIS, a hybrid SM-UTADIS approach is proposed to detect falsified financial statements by classifying financial ratio data into FFS and non-FFS groups. To evaluate the performance of this hybrid method, we conduct experiments using the annual financial ratios of listed companies in the TCM sector in China. Compared with UTADIS and logistic and SM-logistic regression models, the results show that the hybrid SM method can improve the clustering accuracy, and the SM-UTADIS method has the highest prediction accuracy.

The main contributions of this paper are summarized as follows:

(1) From the candidate matching companies, the cosine similarity algorithm is introduced to select out the matched companies, similar to the considered companies. Based on this, we use the financial data of matched companies to compute the deviation of the considered company by SM method. The financial deviation data obtained by SM method can reflect the intrinsic law of a considered company more clearly and make it easier to detect FFS with UTADIS.

(2) We formulate a hybrid SM-UTADIS method by combining the cosine SM algorithm with UTADIS method for detecting FFS.

(3) We give an empirical analysis by taking the traditional Chinese medicine industry as our research sample and prove the outperformance of the proposed hybrid method.

The proposed hybrid method can also be used for FFS detection in other industries. Here, the traditional Chinese medicine industry is just chosen as an example to test our hybrid method in this paper. Note that the industry of research samples had better have the homogeneous feature in their main business. The usefulness of this study first comes from the possibility of applying current working methods in financial fraud detecting and the improvement of classification methods. The development direction of future research is to expand the sample of the analyzed companies, focus on specific activity objects, determine the characteristics of each department, and improve the proposed model according to the specific economic environment of each company, so as to provide the best possible guarantee for the existence of fraud.

The importance of this topic and its results stems from the promotion of the method to identify financial fraud, which may contribute to the successful prevention and detection of these catastrophic actions.

Data Availability

The data used to support the findings of this study can be accessed from the following online address: http://www.wind.com.cn/.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This study was partially supported by the National Natural Science Foundation of China (Grant no. 71761029), Natural Science Foundation of Inner Mongolia Autonomous Region (Grant no. 2017MS717), and the Program for Innovative Research Team in Universities of Inner Mongolia Autonomous Region (Grant no. NMGIT1405).

References

[1] G. Apparao, A. Singh, G. S. Rao, B. L. Bhavani, K. Eswar, and D. Rajani, "Financial statement fraud detection by data mining," *International Journal of Advanced Networking and Applications*, vol. 1, no. 3, pp. 159–163, 2009.

[2] M. Omidi, Q. Min, V. Moradinaftchali, and M. Piri, "The efficacy of predictive methods in financial statement fraud," *The Scientific World Journal*, vol. 2014, Article ID 968712, 9 pages, 2014.

[3] D. R. Cox, "The regression analysis of binary sequences," *Journal of the Royal Statistical Society*, vol. 21, no. 1, p. 238, 1958.

[4] M. S. Beasley, "An empirical analysis of the relation between the board of director composition and financial statements," *Accounting Review*, vol. 71, no. 4, pp. 443–465, 1996.

[5] A. Ines, A. Ben, and J. Anis, "Detection of fraud in financial statements: French companies as a case study," *International Journal of Academic Research in Business and Social Sciences*, vol. 3, no. 5, pp. 2222–6990, 2013.

[6] J. V. Hansen, J. B. Mcdonald, W. F. Messier, and T. B. Bell, "A generalized qualitative-response model and the analysis of management fraud," *Management Science*, vol. 42, no. 7, pp. 1022–1032, 1996.

[7] O. S. Persons, "Using financial statement data to identify factors associated with fraudulent financial reporting," *Journal of Applied Business Research*, vol. 11, no. 3, pp. 38–46, 1995.

[8] C. T. Spathis, "Detecting false financial statements using published data: some evidence from Greece," *Managerial Auditing Journal*, vol. 17, no. 4, pp. 179–191, 2002.

[9] S.-D. Chen, Y.-J. J. Goo, and Z.-D. Shen, "A hybrid approach of stepwise regression, logistic regression, support vector machine, and decision tree for forecasting fraudulent financial statements," *The Scientific World Journal*, vol. 2014, Article ID 968712, 9 pages, 2014.

[10] H. Ye, L. Xiang, and Y. Gan, "Detecting financial statement fraud using random forest with SMOTE," *Materials Science and Engineering*, vol. 612, no. 5, 2019.

[11] Y. Zhang, G. Chu, and D. Shen, "The role of investor attention in predicting stock prices: the Long short-term memory networks perspective," *Finance Research Letters*, Article ID 101484, 2020.

[12] B. P. Green and J. H. Choi, "Assessing the risk of management fraud through neural-network technology," *Auditing: A Journal of Practice and Theory*, vol. 16, no. 1, pp. 14–28, 1997.

[13] J. Yao, J. Zhang, and L. Wang, "A financial statement fraud detection model based on hybrid data mining methods," *International Conference on Artificial Intelligence and Big Data (ICAIBD)*, vol. 2018, pp. 57–61, Article ID 4989140, 2018.

[14] C.-L. Jan, "An effective financial statements fraud detection model for the sustainable development of financial markets: evidence from Taiwan," *Sustainability*, vol. 2018, no. 2, 14 pages, Article ID 8882253, 2018.

[15] K. M. Fanning and K. O. Cogger, "Neural network detection of management fraud using published financial data," *Intelligent Systems in Accounting Finance & Management*, vol. 7, no. 2, pp. 21–41, 1998.

[16] M. Pazarskis, G. Drogalas, and K. Baltzi, "Detecting false financial statements: evidence from Greece in the period of economic crisis," *Investment Management and Financial Innovations*, vol. 14, no. 3, pp. 102–112, 2017.

[17] G. S. Temponeras, S. N. Alexandropoulos, S. B. Kotsiantis, and M. N. Vrahatis, "Financial fraudulent statements detection through a deep dense artificial neural network," in *Proceedings of the 2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pp. 1–5, Patras, Greece, 2019.

[18] E. Kirkos, C. Spathis, and Y. Manolopoulos, "Data mining techniques for the detection of fraudulent financial statements," *Expert Systems with Applications*, vol. 32, no. 8, pp. 995–1003, 2007.

[19] R. Gupta and N. S. Gill, "Prevention and detection of financial statement fraud – an implementation of data mining framework," *International Journal of Advanced Computer Science and Applications*, vol. 3, no. 8, pp. 150–156, 2012.

[20] S. Zionts, "MCDM-if not a roman numeral, then what?" *Interfaces*, vol. 9, no. 4, pp. 94–101, 1979.

[21] B. Roy, *Multicriteria Methodology for Decision Aiding*, Kluwer Academic Publishers, Dordrecht, Netherlands, 1996.

[22] E. Jacquet-Lagreze and J. Siskos, "Assessing a set of additive utility functions for multicriteria decision-making, the UTA method," *European Journal of Operational Research*, vol. 10, no. 2, pp. 151–164, 1982.

[23] Y. Siskos and D. Yannacopoulos, "UTASTAR: An ordinal regression method for building additive value functions," *Investigação Operacional*, vol. 5, pp. 39–53, 1985.

[24] S. Corrente, M. Doumpos, S. Greco, R. Słowiński, and C. Zopounidis, "Multiple criteria hierarchy process for sorting problems based on ordinal regression with additive value functions," *Annals of Operations Research*, vol. 251, no. 1-2, pp. 117–139, 2017.

[25] V. Mousseau, Ö. Özpeynirci, and S. Özpeynirci, "Inverse multiple criteria sorting problem," *Annals of Operations Research*, vol. 253, no. 1, pp. 1–34, 2017.

[26] C. M. D. M. Mota and A. T. de Almeida, "A multicriteria decision model for assigning priority classes to activities in project management," *Annals of Operations Research*, vol. 199, no. 1, pp. 361–372, 2012.

[27] C. Zopounidis and M. Doumpos, "A multicriteria decision aid methodology for sorting decision problems: the case of financial distress," *Computational Economics*, vol. 14, no. 3, pp. 197–218, 1999.

[28] K. Kosmidou, F. Pasiouras, M. Doumpos, and C. Zopounidis, "Assessing performance factors in the UK banking sector: a multicriteria methodology," *Central European Journal of Operations Research*, vol. 14, no. 1, pp. 25–44, 2006.

[29] M. M. Reza, S. M. M. Reza, and E. M. S. M. Mohsen, "Applying the clustering and UTADIS models to form an investment portfolio," *Financial Research*, vol. 20, no. 1, pp. 53–74, 2018.

[30] M. Doumpos, C. Zopounidis, and P. Fragiadakis, "Assessing the financial performance of European banks under stress testing scenarios: a multicriteria approach," *Operational Research*, vol. 16, no. 2, pp. 197–209, 2016.

[31] C. Spathis, M. Doumpos, and C. Zopounidis, "Detecting falsified financial statements: a comparative study using multicriteria analysis and multivariate statistical techniques," *European Accounting Review*, vol. 11, no. 3, pp. 509–535, 2002.

[32] S. Grigori, G. Alexander, G.-A. Helena, and P. David, "Soft similarity and soft cosine measure: similarity of features in vector space model," *Computacióny Sistemas*, vol. 18, no. 3, pp. 491–504, 2014.

[33] M. Mironiuc, I.-B. Robu, and M.-A. Robu, "The fraud auditing: empirical study concerning the identification of the financial dimensions of fraud," *Journal of Accounting and Auditing: Research & Practice*, vol. 2012, Article ID 391631, 13 pages, 2012.

[34] S. Y. Huang, R. H. Tsaih, and W. Y. Lin, "Unsupervised neural networks approach for understanding fraudulent financial reporting," *Industrial Management & Data Systems*, vol. 112, no. 2, pp. 224–244, 2012.