FEATURE EXTRACTION OF FRAUDULENT FINANCIAL REPORTING THROUGH UNSUPERVISED NEURAL NETWORKS

Shin-Ying Huang[*], *Rua-Huan Tsaih*[†], *Wan-Ying Lin*[‡]

Abstract: The objective of this study is to apply an unsupervised neural network tool to analyze fraudulent financial reporting (FFR) by extracting distinguishing features from samples of groups of companies and converting them into useful information for FFR detection. This methodology can be used as a decision support tool to help build an FFR identification model or other financial fraud or financial distress scenarios. The three stages of the proposed quantitative analysis approach are as follows: the data-preprocessing stage; the clustering stage, which uses an unsupervised neural network tool known as a growing hierarchical self-organizing map (GHSOM) to cluster sample observations into subgroups with hierarchical relationships; and the feature-extraction stage, which uncovers common features of each subgroup via principle component analysis. This study uses the hierarchal topology mapping ability of a GHSOM to cluster financial data, and it adopts principal component analysis to determine common embedded features and fraud patterns. The results show that the proposed three-stage approach is helpful in revealing embedded features and fraud patterns, using a set of significant explanatory financial indicators and the proportion of fraud. The revealed features can be used to distinguish distinctive groups.

Key words: *Fraudulent financial reporting, growing hierarchical self-organizing map, unsupervised neural network, feature extraction*

Received: November 12, 2013 **DOI:** 10.14311/NNW.2014.24.031
Revised and accepted: October 18, 2014

[*]Shin-Ying Huang – Corresponding Author, Research Center for Information Technology Innovation, Academia Sinica, Taipei 11529, Taiwan, Tel.: 886-2-2787-2300 ext. 2341, E-mail: smichelle19@gmail.com

[†]Rua-Huan Tsaih, Department of Management Information Systems, National Chengchi University, Taipei 11605, Taiwan, Tel.: 886-2-29393091 ext. 81036, E-mail: tsaih@mis.nccu.edu.tw

[‡]Wan-Ying Lin, Department of Accounting, National Chengchi University, Taipei 11605, Taiwan, Tel.: 886-2-29393091 ext. 81130, E-mail: wanying@nccu.edu.tw

1. Introduction

For the purpose of quantitatively analyzing financial reporting fraud, this study explores the advantages of an approach based upon a growing hierarchical self-organizing map (GHSOM), which is an unsupervised neural network tool [12][13][38]. Fraudulent financial reporting (FFR), also known as financial statement fraud or management fraud, is a type of fraud that adversely affects stockholders' and stakeholders' decisions using misleading financial reports [14]. FFR involves the intentional misstatement or omission of material information in an organization's financial reports [6]. Although FFR is infrequent, it can lead to severe financial crises for capital markets and financial losses for stockholders. Given the infrequency of synthetic reporting, most auditors cannot develop sufficient experience in or knowledge about FFR detection [15]. Furthermore, internal control mechanisms usually seek to prevent employee fraud and not management fraud. Thus, top management can bypass internal controls and be involved in providing unfairly presented financial statements and deliberately defrauding auditors [33]. Beasley et al. [6] found that 83% of top managers of U.S.-listed firms — chief executive officers, chief financial officers, and occasionally both — have been associated with financial statement fraud. Zhao et al. [53] found that staggered boards lessen takeover threats and thus mitigate managers' pressure to overstate earnings. Hart et al. [19] stated, "understanding the subtle indications of lying would certainly benefit anyone wishing to detect lying and deception in others" (page 135). Thus, there is an imperative need for a quantitative approach that can effectively analyze financial statements to help detect deception.

Most previous FFR-related studies have drawn research inferences from either FFR cases or archival data (firm-year observations). As Basens et al. [5] stated, studies of the empirical approach have emphasized the predictability of models and exhibit a trend toward emphasis on the classification mechanism used as the decision support system for future risk identification. There have also been other studies focusing on understanding the nature of FFR using case study methods. To integrate the advantages of these two approaches, a method that can contribute to both the prediction and explanation of FFR is proposed. Specifically, this study proposes using an unsupervised neural networks-based method, the hierarchical self-organizing map (GHSOM) [12], which is an extension of the self-organizing map (SOM) [29], to classify financial statements. The SOM has been studied in terms of methodology and statistical features [10,21,27,48], and GHSOMs are gradually being used more and are being integrated with other methods because of their flexible and hierarchical features [34,36,42,49]. Based on the clustering results generated by the GHSOM, this study adopts principal component analysis (PCA) to discover common embedded features and fraud patterns from each group of financial data. In addition, this study illustrates the clustering results added to the FFR ratios on the topology map, which can help users to define risky areas more easily and to conduct further investigations.

The remainder of this study is organized as follows. Section two presents a literature review of the FFR-related studies. Section three explains the proposed methodology. Section four presents the experiment conducted with FFR examples from Taiwan and reports the results of the data pre-processing, the clustering,

and the extraction of features by PCA with regard to several example subgroups. Section five concludes with a summary and a discussion of the findings based on the experimental results.

2. Literature review

Studies focusing on the nature of FFR often use a case study approach to provide a descriptive analysis of the characteristics of FFR and the techniques commonly used. As shown in Tab. I, the Committee of Sponsoring Organizations (COSO) examines and summarizes certain key company and management characteristics, based on FFR samples from U.S. companies. The Association of Certified Fraud Examiners (ACFE) analyzes the nature of occupational fraud schemes of U.S. companies and provides suggestions to create adequate internal control mechanisms.

Study	Methodology	Findings
Beasley et al. (1999) [6]	• Case study • Descriptive analysis	• Nature of companies committing financial statement fraud – They were relatively small. – They were inclined to experience net losses or to be close to break-even positions in periods before the fraud. • Nature of the control environment of companies involved – Top senior executives were frequently involved. – Most audit committees only met approximately once per year, or the company had no audit committee. • Nature of the frauds – The cumulative amounts of fraud were relatively large in light of the relatively small sizes of the companies involved – Most frauds were not isolated to a single fiscal period. – Typical financial statement fraud techniques involved the overstatement of revenues and assets. • Consequences for the company and individuals involved – Severe consequences awaited companies committing fraud. – Consequences associated with financial statement fraud were severe for the individuals allegedly involved.
ACFE (2008) [2]	• Case study • Descriptive analysis	• Occupational fraud schemes tend to be extremely costly. The median loss was $175,000. More than one-quarter of the frauds involved losses of at least $1 million. • Occupational fraud schemes frequently continue for years, two years being typical, before they are detected. • There are 11 distinct categories of occupational fraud. Financial statement fraud was the most costly category, with a median loss of $2 million for the cases examined.

		• The industries most commonly victimized by fraud in the study were banking and financial services (15% of cases), government (12%) and health care (8%). • Fraud perpetrators often display behavioral traits that serve as indicators of possible illegal behavior. In financial statement fraud cases, which tend to be the most costly, excessive organizational pressure to perform was a particularly strong warning sign.
ACFE (2010) [3]	• Case study • Descriptive analysis	• The median loss caused by the occupational fraud cases studied was $160,000. Nearly one-quarter of the frauds involved losses of at least $1 million. • The frauds lasted a median of 18 months before being detected. • Financial statement fraud schemes comprised less than 5% of the frauds reported in the study but caused a median loss of more than $4 million — the most costly category. • Small organizations are disproportionately victimized by occupational fraud. These organizations are typically lacking in anti-fraud controls compared to their larger counterparts, which makes them particularly vulnerable to fraud.
ACFE (2012) [4]	• Case study • Descriptive analysis	• The median loss caused by the occupational fraud cases studied was $140,000. More than one fifth of these cases caused losses of at least $1 million. • The industries most commonly victimized were the banking and financial services, government and public administration, and manufacturing sectors. • Financial statement fraud is the most costly form of occupational fraud, causing a median loss of $1 million. • Individuals engaged in financial statement fraud were much more likely than other committers of fraud to face excessive pressure from within their organizations.

Tab. I *Research methodology and findings in selected nature-related FFR studies.*

Studies focusing on predicting FFR often use the empirical approach to analyze archival data (firm-year observations) and to identify significant variables that help to predict the occurrence of FFR. Tab. II summarizes the research methodology and findings of FFR empirical studies that are relevant to our study. Note that the matched-pairs design is typical for traditional FFR empirical studies; thus, this study adopts the matched-pair design. Furthermore, as stated in Tab. II, the neural networks algorithm has been applied in the field of FFR. For example, Fanning and Cogger [15] proposed an adaptive neural networks algorithm to help detect FFR. Kirkos et al. [28] compared a decision tree, back-propagation neural network and a Bayesian belief network in FFR detection and found that the back-propagation neural network was the more accurate method, using a training dataset. Liou [32] also applied a neural networks algorithm to help detect FFR. Note that the unsupervised neural networks have fewer applications in this literature, and there

is potential for unsupervised neural networks in helping to explain the embedded features. To help close the gap in this line of research, we adopt the unsupervised neural networks in this study.

Note that Carlos [9] applied a self-organizing map (SOM) in detecting bankruptcy. The SOM is an unsupervised neural network tool [29]. The major advantage of an SOM is its ability to represent topological relationships among high-dimensional inputs from a low-dimensional perspective. Other advantages of SOMs include their adaptive nature (i.e., the classification process can be modified if new training data are reset) and their robustness. The most widespread use of SOMs is in the identification and visualization of natural groupings of data.

To address the weaknesses of SOMs, including the predefined and fixed topology size and the inability to identify hierarchical relations among samples, Dittenbach, Merkl, and Rauber [12] developed the concept of a GHSOM, which addresses the issue of the fixed network architecture of an SOM through a multilayer hierarchical network structure. The flexible and hierarchical features of a GHSOM generate more delicate clustering results than an SOM and make a GHSOM a versatile analysis tool for tasks regarding data mining, image recognition, Web mining, and text mining problems [12,13,38,41,43,46,50]. Tsaih et al. [46] used GHSOM to cluster preliminarily non-fraud and fraud financial statements into subgroups with hierarchical relationships. Their results showed that the GHSOM can be used to classify the samples into high fraud risk groups, mixed groups, and healthy groups. However, they did not discuss the feature extraction of the FFR. Huang et al. [22] used GHSOM to cluster FFR cases and to extract FFR features from each subgroup with the assistance of domain experts. The current study extends the work of Huang et al. [22] by proposing a systematic approach (i.e., PCA) to extract FFR features. The current study also adopts visualization technique to illustrate the clustering effect regarding the fraud ratios of neighboring nodes. The visualization of clustering results can help users to focus on the risky subgroups more easily.

In summary, the void in this line of research motivates the current study to develop an approach to help extract valuable patterns from abundant observations. This study proposes a three-stage approach. The data-preprocessing stage focuses on calculating the variables (i.e., measurements) and includes the variable selection. Then, the clustering stage is designed to divide all of the observations into similar groups. Finally, the feature extraction stage uses a quantitative analysis method to depict the representative features of each subgroup. Section 3 provides more details of the proposed methodology.

3. Methodology

The proposed three-stage approach depicted in Fig. 1 is used to apply certain classification techniques to the clustering of financial ratios derived from financial statements and then to apply certain quantitative analysis techniques to reveal features — the delicate but hidden truths — of each cluster. The objective of this study is to examine the advantages of applying GHSOMs in conducting the clustering tasks of the proposed quantitative approach, which can be used to help to detect FFR or other financial distress scenarios.

Study	Methodology	Variable	Sample	Findings
Persons (1995) [37]	• Stepwise logistic model	• 9 financial ratios • Z-score	Matched – pairs design: 203 fraud samples and 203 non-fraud samples	The study identified four significant indicators: financial leverage, capital turnover, asset composition and firm size.
Fanning and Cogger (1998) [15]	• Neural networks	• 62 variables • Financial ratios • Other indicators: corporate governance, capital structure	Matched – pairs design: 102 fraud samples and 102 non-fraud samples	• The neural network is more effective. • Financial ratios such as debt to equity, ratios of accounts receivable to sales, and trend variables are significant indicators.
Bell and Carcello (2000) [7]	• Logistic regression	46 fraud risk factors	77 fraud samples and 305 non-fraud samples	Logistic regression model outperformed auditors for fraud samples but performed equally for non-fraud samples.
Kirkos et al. (2007) [28]	• Decision tree • Back-propagating neural network • Bayesian belief network	• 27 financial ratios • Z-score	Matched – pairs design: 38 fraud samples and 38 non-fraud samples	• In a training dataset, the neural network was more accurate. • In a validation dataset, the Bayesian belief network was more accurate.
Hoogs et al. (2007) [20]	• Genetic algorithm	• 38 financial ratios • 9 qualitative indicators	Matched – pairs design: 51 fraud samples and 51 non-fraud samples	An integrated pattern had wider coverage for companies suspected of fraud and a lower rate of false classification for companies that had not committed fraud.
Liou (2008) [32]	• Logistic regression • Neural networks • Decision tree	• 52 financial ratios	20 fraud samples and 515 non-fraud samples	Logistic regression outperformed the other two algorithms in detecting fraudulent financial reporting.
Gupta and Gill (2012) [18]	• Decision tree • Naive Bayesian classifier • Genetic programming	• 62 financial ratios	29 fraud samples and 85 non-fraud samples	The decision tree produced the best sensitivity and genetic programming the best specificity compared with the other two methods.

Tab. II *Research methodology and findings of selected FFR empirical studies.*

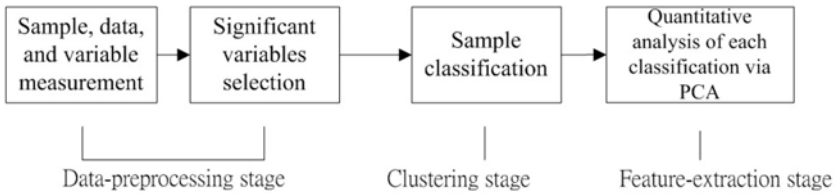

Fig. 1 *The proposed quantitative approach to conducting the data-preprocessing task, the clustering task, and the feature-extracting task in sequence.*

The data pre-processing stage in Fig. 1 includes the tasks of sampling, data and variable measurement, and selection of significant variables. Discriminant analysis is applied to the financial ratio data derived from financial statements, to identify the significant variables that help to predict the occurrence of FFR.

In the clustering stage, a classification technique is used to cluster samples into small-sized subgroups based on the significant variables identified. The clustering performance of the GHSOM is compared with that of three methods — K-means, two-step clustering, and SOM — that cluster samples based on Euclidean distance without knowing the dependent variable. The differences among these four methods are briefly described as follows. K-means clustering seeks to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. Two-step clustering uses an agglomerative hierarchical clustering approach. The SOM and GHSOM approaches perform topology-preserving mapping from a high-dimensional space to map units. The GHSOM is used to perform data clustering because of its hierarchical visualizing and topological preserving abilities. For the training process of a GHSOM, please refer to Dittenbach et al. [12] and Huang et al. [22]. This stage focuses on clustering performance comparisons among various methods. Therefore, the performance comparison is based on the mean quantization error (MQE) [12][13][40] rather than on the contribution to reducing the test error rate for fraud. The MQE value is an indicator that represents the dissimilarity of the clustered samples. Conceptually, the smaller the MQE value is, the smaller the dissimilarity of the clustered samples is. As stated in Dittenbach et al. [13], the MQE value "is commonly used as a quality measure for data representation with SOMs" (page 5).

In the feature extraction stage, PCA is used to reveal the features of each subgroup. Because each subgroup is small in size, we can assume that linear relationships exists between the independent variables and the dependent variable such that PCA is appropriate for identifying a set of variables that reveal the features of each subgroup by transforming the original input variables into a new set of principal components [8,23,24,35].

4. Experiment and results

We use the matched-pairs concept to create a sample pool of 58 fraud firms and 58 non-fraud firms, all of which are companies publicly traded in Taiwan. For each fraud firm, we pick a non-fraud counterpart in the same industry with similar total

assets. For the fraud firms, indictments and judgments issued by the Department of Justice state the detection years that were investigated by the prosecutors' offices and the fiscal years of financial statements that were fraudulent. For each fraud firm, we establish a five-year sampling period of financial statements, the center of which is the fraud year. We collect a total of 113 fraudulent financial statement samples and 467 non-fraudulent financial statement samples.

Tab. III lists the relevant FFR literature and the definitions and measurements of the variables. As stated in Huang et al. [22], we reviewed the FFR literature listed in the second column of Tab. III and summarized the 25 indicators that are significantly related to the profitability, liquidity, operating ability, financial structure and cash flow ability of a firm. In Tab. III, the dependent variable is FRAUD, which is dichotomous and is expressed as 1 or 0, depending on whether the year's financial statement was fraudulent or non-fraudulent. The explanatory variables are collected from the *Taiwan Economic Journal* (TEJ) database. Among the 25 independent variables, there are financial ratios that measure the profitability, liquidity, operating ability, financial structure and cash flow ability of a firm. There are also corporate governance variables and a Z-score, which is utilized to examine the probability of financial distress.

Variable Definition	**Study**	**Measurement**
Dependent variable:		
FRAUD	[37]	If a company's financial statements for specific years are confirmed to be fraudulent by indictments and sentences for major securities crimes issued by the Department of Justice, the firm-year data are classified as fraud observations, and the variable *FRAUD* is set to 1; otherwise, FRAUD is set to 0.
Independent variable:		
Profitability		
Gross profit margin (*GPM*)	[11]	$\frac{\text{Operating income-Operating costs}}{\text{Operating income}}$
Operating profit ratio (*OPR*)	[17]	$\frac{\text{Operating income-Operating costs-Operating expenses}}{\text{Operating income}}$
Return on assets (*ROA*)	[20][37]	$\frac{\text{Net income+Interest expenses}\times\text{(1-Tax rate)}}{\text{Average total assets}}$
Growth rate of sales (*GROS*)	[11][44][45]	$\left(\frac{\text{Sales}}{\text{Net sales in prior fiscal year}}\right) - 1$
Growth rate of net income (*GRONI)*	[7][45]	$\left(\frac{\text{Net sales}}{\text{Net income in prior fiscal year}}\right) - 1$
Liquidity		
Current ratio (*CR*)	[28]	$\frac{\text{Current assets}}{\text{Current liabilities}}$
Quick ratio (*QR*)		$\frac{\text{Current assets-Inventories-Prepaid expenses}}{\text{Current liabilities}}$
Operating ability		
Accounts receivable turnover (*ART*)	[17]	$\frac{\text{Net credit sales}}{\text{Average accounts receivable}}$

Total asset turnover (*TAT*)	[28][37]	$\frac{\text{Net sales}}{\text{Total assets}}$
Growth rate of accounts receivable (*GROAR*)	[11]	$\left(\frac{\text{Accounts receivable}}{\text{Accounts receivable in prior fiscal year}}\right) - 1$
Growth rate of inventory (*GROI*)	[11]	$\left(\frac{\text{Inventory}}{\text{Inventory in prior fiscal year}}\right) - 1$
Growth rate of accounts receivable to sales (*GRARTS*)	[45]	$\frac{\text{Accounts receivable}_{t}}{\text{Gross sales}_{t}} - \frac{\text{Accounts receivable}_{t\text{-}1}}{\text{Gross sales}_{t\text{-}1}}$
Growth rate of inventory to gross sales (GRITGS)	[45]	$\frac{\text{Inventory}_{t}}{\text{Gross sales}_{t}} - \frac{\text{Inventory}_{t\text{-}1}}{\text{Gross sales}_{t\text{-}1}}$
Accounts receivable to total assets (*ARTTA*)	[17][37][44]	$\frac{\text{Accounts receivable}}{\text{Total assets}}$
Inventory to total assets (*ITTA*)	[37][44]	$\frac{\text{Inventory}}{\text{Total assets}}$
Financial structure		
Debt ratio (*DR*)	[28][37]	$\frac{\text{Total liabilities}}{\text{Total assets}}$
Long-term funds to fixed assets (*LFTFA*)	[28]	$\frac{\text{Equity+Longterm liabilities}}{\text{Fixed assets}}$
Cash flow ability		
Cash flow ratio (*CFR*)		$\frac{\text{Cash flows from operating activities}}{\text{Current liabilities}}$
Cash flow adequacy ratio (*CFAR*)	[11]	$\frac{\text{Five year sum of cash flows from operating activities}}{\text{(Five year sum of capital expenditures, inventory additions and cash dividends)}}$
Cash flow reinvestment ratio (*CFRR*)		$\frac{\text{Cash flows from operating activities-Cash dividends}}{\text{(Gross fixed assets + Long term investments + Other assets + Working capital)}}$
Financial difficulty		
Z-score	[1][15][16] [44][45]	$1.2 \times \left(\frac{\text{Working capital}}{\text{Total assets}}\right) + 1.4 \times (\frac{\text{Retained earnings}}{\text{Total assets}}) + 3.3 \times \left(\frac{\text{Earnings before interest and taxes}}{\text{Total assets}}\right) + 0.6 \times \left(\frac{\text{Market value of equity}}{\text{Book value of total debt}}\right) + 1.0 \times \text{TAT}$
Corporate governance		
Stock pledge ratio (*SPR*)	[31]	$\frac{\text{large shareholders' shareholdings in pledge}}{\text{large shareholders' shareholdings}}$
Sum of percentage of major shareholders' shareholdings (*SMLSR*)	[6][52]	Σ (Percentage of shareholdings >10%)
Deviation between CR and CFR (*DBCRCFR*)	[30][31]	Voting rights – Cash flow rights
Deviation between CBS and CFR (*DBCBSCFR*)	[31][51]	Percentage of board seats controlled – Cash flow rights

Tab. III *Variable definitions and measurements.*

These 25 variables are then incorporated into the variable selection process, which involves a multicollinearity test and canonical discriminant analysis (CANDISC). The structure coefficients (i.e., discriminant loadings) are used to compare the discriminant power of individual variables.

The result of the multicollinearity test suggests that GRITGS should be excluded because its tolerance is much lower than that of the other independent variables. As a result, 24 independent variables are incorporated in the CANDISC shown in Eq. (1):

$$\begin{aligned}\text{FRAUD} = {} & \alpha_1 \times \text{GPM} + \alpha_2 \times \text{OPR} + \alpha_3 \times \text{ROA} + \alpha_4 \times \text{GROS} + \alpha_5 \times \text{GRONI} + \\ & + \alpha_6 \times \text{CR} + \alpha_7 \times \text{QR} + \alpha_8 \times \text{ART} + \alpha_9 \times \text{TAT} + \alpha_{10} \times \text{GROAR} + \alpha_{11} \times \\ & \times \text{GROI} + \alpha_{12} \times \text{GRARTS} + \alpha_{13} \times \text{ARTTA} + \alpha_{14} \times \text{ITTA} + \alpha_{15} \times \text{DR} + \\ & + \alpha_{16} \times \text{LFTFA} + \alpha_{17} \times \text{CFR} + \alpha_{18} \times \text{CFAR} + \alpha_{19} \times \text{CFRR} + \alpha_{20} \times \\ & \times \text{Z-Score} + \alpha_{21} \times \text{SPR} + \alpha_{22} \times \text{SMLSR} + \alpha_{23} \times \text{DBCRCFR} + \alpha_{24} \times \\ & \times \text{DBCBSCFR}.\end{aligned} \tag{1}$$

The results show that the Wilks' Λ value equals 0.766, and χ^2 equals 151.095 (both are significant at the p-value level <0.01), indicating that the discriminant model employed has adequate explanatory power. The corresponding p-values indicate that the following eight variables have statistically significant effects: ROA, CR, QR, DR, CFR, CFAR, Z-Score and SPR. These eight chosen variables are collected for each sample and are used as the training data for a clustering model. Theoretically, these eight variables serve as proxies for a company's attributes in the following respects.

1. Profitability: ROA can be used to assess a firm's ability to generate profits by the use of its own assets. Persons [37] indicated that underperforming firms can give management an incentive to overstate revenues or understate expenses.

2. Liquidity: CR and QR can be used to measure a firm's liquidity which means its short-term ability to pay a debt. QR excludes inventory and prepaid expenses, the liquidity of which is lower than that of cash or accounts receivable.

3. Financial structure: DR can be used to measure a firm's financial structure. Persons [37] found that fraud firms have higher financial leverage than non-fraud firms.

4. Cash flow ability: CFR and CFAR can be used to test a company's ability to pay debts and other disbursements, such as capital expenditures, inventory additions and cash dividends, using cash flows from operating activities.

5. Stock pledge ratio: SPR can be used to measure the degree of financial pressure on directors and supervisors to pledge their stocks to obtain funds.

6. Financial condition: The Z-score can be used to measure a company's financial situation to determine the relationship between financial distress and fraud.

Based on the examination of the FFR detection results in Tab. IV, we can say that the CANDISC prediction model, based on these eight variables and shown in Eq. (2), is not bad. The prediction power is 79.1%, with a 16.5% probability of a type I error and a 38.9% probability of a type II error. Eq. (2) identifies the dependent variables with statistically significant effects that will be used as the input variables of the clustering methods in the clustering stage:

$$\text{FRAUD} = 0.77\text{ROA} + 0.34\text{CR} + 0.28\text{QR} - 0.42\text{DR} + 0.33\text{CFR} + 0.24\text{CFAR} + 0.64\text{Z-score} - 0.47\text{SPR} \quad (2)$$

	Class		**Prediction**	
			non-fraud	fraud
Original	No. of observations	non-fraud	390	77
		fraud	44	69
	%	non-fraud	83.5	16.5
		fraud	38.9	61.1

Tab. IV *FFR detection results of the canonical discriminant prediction model.*

In the clustering stage, we use the following four unsupervised classification techniques: GHSOM, SOM, K-means and two-step clustering. We use the GHSOM toolbox and SOM toolbox in the platform of MATLAB R2007a, and we use the K-means package and two-step clustering package in SPSS software, version 13. The classification technique with the best clustering performance is chosen, and the clustering results are used in the feature extraction stage.

In applying GHSOM, we obtain several hierarchical structures for different values of the breadth parameter (τ_1) and depth parameter (τ_2). While τ_1 controls the size of individual SOMs in the GHSOM, τ_2 determines the minimum data granularity and the global termination criterion. We then select a suitable GHSOM based on the following criteria:

1. The depth of a model should be greater than two layers;
2. The breadth of an individual node should at least consist of two firms; and
3. New nodes should not be overly clustered in a minority of their parent nodes

Tab. V shows 13 candidate GHSOM configurations conducted under different τ_1 andτ_2 settings. As shown in Tab. V, when the depth value is 0.01, we find that a small breadth value results in a flat structure, and the number of mappings in each layer and the total number of leaf mappings converge when the breadth value is greater than 0.7. Then, we attempt to increase the depth value under the breadth values 0.5 and 0.7, and we find that test No. 12 with three layers and 41 leaf mappings fit the predefined selection criteria.

As shown in Fig. 2, the selected GHSOM is a tree that has three layers, 52 nodes (subgroups) and 41 leaf nodes. In each node, there is a given name based on the layer number, node order, and the path from the root node to its position.

No	Parameter		Total of Layers	Number of Mappings				Number of Leaf Mappings
	Breadth (τ_1)	Depth (τ_2)		Layer 1	Layer 2	Layer 3	Layer 4	
1	0.1	0.01	1	144				144
2	0.2	0.01	1	63				63
3	0.3	0.01	2	21	222			231
4	0.4	0.01	2	9	125			125
5	0.5	0.01	3	6	59	4		62
6	0.6	0.01	3	4	25	85		95
7	0.7	0.01	4	4	16	54	6	63
8	0.8	0.01	4	4	16	48	4	55
9	0.9	0.01	4	4	16	48	4	55
10	1.0	0.01	4	4	16	48	4	55
11	0.5	0.02	2	6	59			59
12*	0.7	0.02	3	4	16	32		41
13	0.7	0.03	3	4	16	10		24

* chosen GHSOM tree

Tab. V Thirteen GHSOM configurations.

For example, node "L1m2-L2m1" is developed from the second node of layer 1, and it is the first child node of layer 2. In each node, the three numbers within parentheses indicate the number of fraudulent financial statements, restated financial statements (statements that were restated and re-announced at the request of a government agency), and non-fraud financial statements, respectively. The number of fraud financial statements is the sum of the number of fraudulent financial statements and the number of restated financial statements. The MQE value of the selected GHSOM is 0.121.

The overall ratio of fraud samples to non-fraud samples is 113:467. This ratio information is adopted as the norm for the nodes. For instance, among the four nodes in the first layer of Fig. 2, L1m2 is high risk because it has a fraud ratio much greater than 113:467, while L1m3 is healthy, with a much lower fraud ratio. The fraud ratio is the ratio of fraud samples to non-fraud samples for each node.

In applying SOM, we obtain the MQEs of SOMs of six different sizes shown in Tab. VI. In summary, as Tab. VI shows, the corresponding MQE value becomes smaller as the size of the SOM becomes larger. For purposes of comparison with the selected GHSOM, the number of subgroups of which is 41, the reasonable counterpart of SOM, shown in Tab. VI is that with a size of 8×8. Because the counterpart SOM has a larger MQE than the GHSOM, the clustering performance of the GHSOM is better than that of the SOM.

In applying the K-means and two-step clustering methods, the K value of K-means is set to 41, and the corresponding MQE value is 0.225, while the number of groups generated by the two-step clustering method is 5, and the corresponding MQE value is 0.3. Compared with the selected GHSOM, the clustering qualities of the K-means and two-step clustering methods are poorer because their corresponding MQE values are much larger.

Briefly, the MQE results show that the clustering performance of the GHSOM is better than that of the other three clustering methods. The clustering method

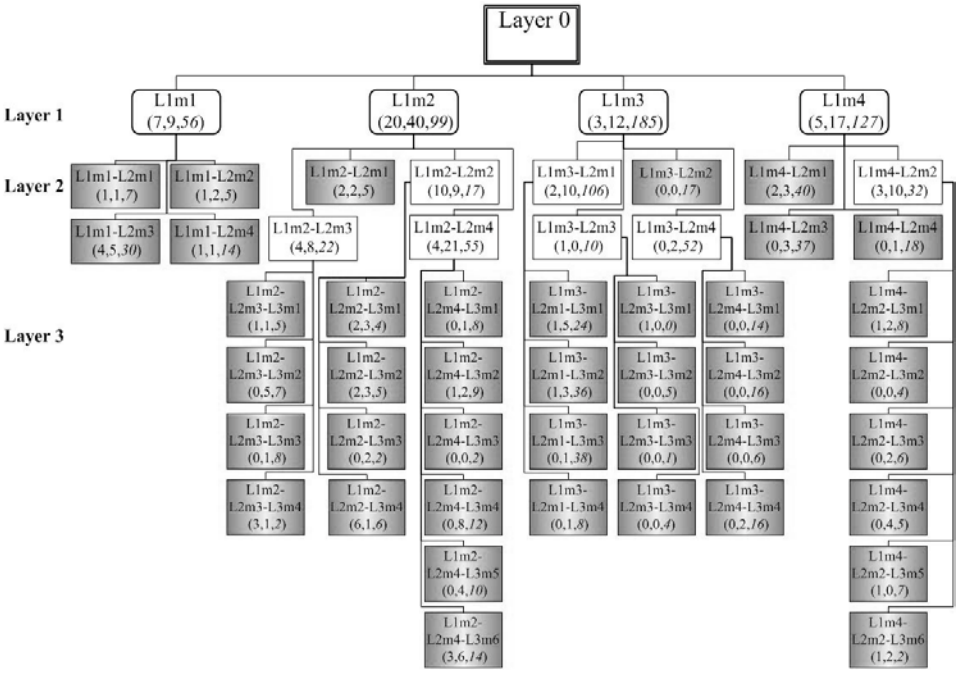

Fig. 2 *The selected GHSOM; leaf nodes are shaded.*

Size	MQE
15 × 8	0.234
8 × 8	0.272
10 × 10	0.244
16 × 16	0.186
20 × 20	0.158
24 × 24	0.137

Tab. VI *The MQEs of SOMs of different sizes.*

with the best clustering performance is selected, and the clustering results are used in the feature extraction stage. It is worth noting that the clustering method selected is better in terms of its clustering ability than in its ability to reduce the test error rate for FRAUD. We believe that the chosen clustering results are helpful in revealing distinctive FFR features.

Based on the GHSOM results, several natural groupings of the GHSOM are found in terms of FFR-related characteristics. For example, Tab. VII shows the fraud ratios for the nodes L1m1, L1m2, L1m3, and L1m4 in layer 1. The fraud ratio of a node is defined as the ratio of its fraud samples to its non-fraud samples. In layer 1, the node L1m2 has the highest fraud ratio, and it clusters more than half of the fraud samples, while node L1m3 has the lowest fraud ratio, with very few fraud samples.

Node	Number of samples		Fraud ratio (%)
	Fraud	Non-fraud	
L1m1	16	56	28.57
L1m2	60	99	60.61
L1m3	15	185	8.11
L1m4	22	127	17.32

Tab. VII *Fraud ratio of each node in layer 1.*

Fig. 3 shows the fraud ratios for the leaf nodes of the GHSOM, with nodes at deeper levels marked in deeper shades of color. It is obvious that if a parent node has a high fraud ratio (e.g., L1m2), its child nodes also tend to have higher fraud ratios (see the top right corner of Fig. 3), and vice versa.

Fig. 4 shows the fraud ratios for all of the nodes of the SOM. The processes of determining the map size of the SOM are described as follows. The map size depends on training data and the number of map units. The number of map units depends on number of training samples (dlen). First, the number of map units is determined (unless it is given). A heuristic formula of map size = 5*sqrt(dlen)' is used to calculate it. Subsequently, the map size is determined. Basically, the two largest eigenvalues of the training data are calculated, and the ratio between

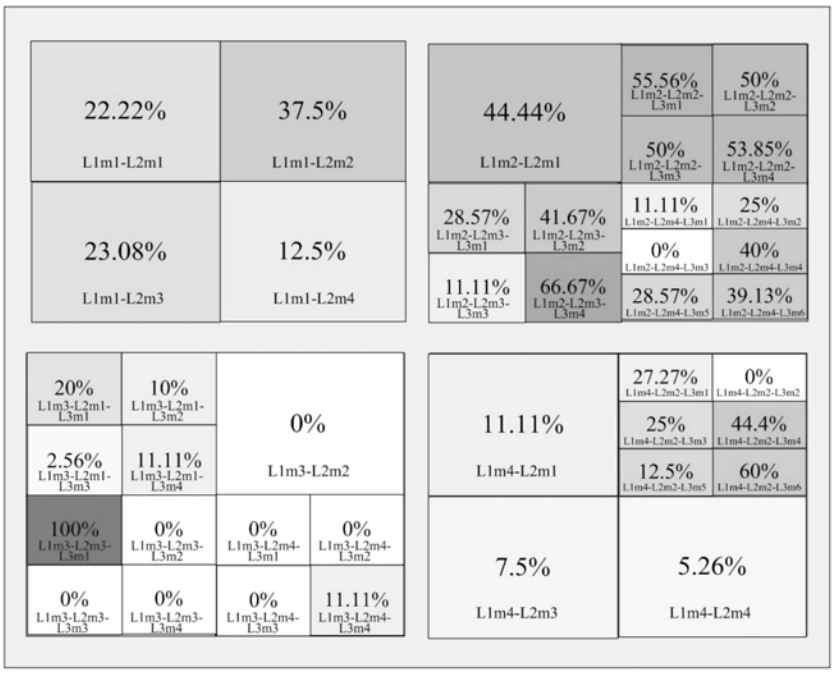

Fig. 3 *Fraud ratios for each node of the GHSOM.*

side lengths of the map grid is set to the square root of this ratio. The actual side lengths are then determined so that their product is as close to the desired number of map units as possible. As a result, the adaptive map size of the SOM is 15×8.

Comparing the results of Figs. 3 and 4, we can see that the GHSOM has denser subgroups with samples, and the risky subgroups are located in the upper portion. The risk level of the risky subgroups decreases gradually despite there being a subgroup in the lower left side with a 100% fraud ratio but with only one sample. In contrast, some subgroups of the SOM have no samples in them (the nil subgroups), and the fraud ratios are not gradually changed due to some healthy nodes being located beyond the risky ones, indicating that the clustering results of the SOM are less efficient for further analysis, compared with the results of the GHSOM.

Without loss of generality, the GHSOM clustering result is used to illustrate the feature extraction stage, and we demonstrate only the features of the following four

40%	50%	80%	75%	50%	16.67%	66.67%	0%
22.22%	0%	60%	12.5%	42.86%	0%	14.29%	40%
60%	25%	25%	100%	0%	0%	Nil	0%
18.18%	0%	20%	0%	0%	10%	100%	50%
42.86%	33.33%	100%	0%	0%	50%	0%	0%
0%	0%	10%	0%	18.18%	18.18%	18.18%	50%
40%	50%	25%	66.67%	33.33%	0%	0%	0%
0%	0%	0%	21.43%	45.45%	25%	0%	0%
25%	0%	33.33%	0%	0%	0%	0%	0%
0%	33.33%	12.5%	33.33%	66.67%	25%	33.33%	9.09%
25%	0%	16.67%	14.29%	0%	0%	Nil	0%
0%	15.38%	55.56%	0%	16.67%	33.33%	16.67%	0%
0%	0%	0%	0%	0%	0%	0%	0%
0%	16.67%	75%	40%	20%	25%	9.09%	20%
0%	Nil	0%	0%	Nil	Nil	9.09%	0%

Fig. 4 *Fraud ratios for each node of the SOM.*

leaf nodes located in different branches: L1m2-L2m3-L3m4, L1m1-L2m2, L1m3-L2m1-L3m1, and L1m4-L2m2-L3m6. For each leaf node of the GHSOM, the values of the eight significant variables for all of the clustered samples are the inputs into the PCA. The PCA is appropriate for identifying a set of variables that reveal the features of each subgroup by transforming the original input variables into a new set of principal components. That is, PCA can help to summarize the representative features from the input variables [8,23,24,35]. In the current study, we are interested in exploring the heterogeneity of each subgroup; thus, PCA helps to perform *systematic* quantitative analysis. The central idea of PCA is to reduce the dimensionality of a dataset consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the dataset. This goal is achieved by transforming to a new set of uncorrelated principal components (PCs), which are ordered so that the first few retain most of the variation present in all of the original variables [25]. PCA has been used in the financial field as a data-preprocessing and explanation tool. For example, Canbas et al. [8] used PCA to construct an integrated early warning system (IEWS) for bank examination, in which PCA is used to explore the underlying features of the financial ratios. In this study, the obtained variable sets represent the statistical patterns of a subgroup and can be used to *label* the difference between each subgroup.

According to Kaiser [26], only those factors with variances greater than 1 (i.e., the corresponding eigenvalues of which are greater than 1) are retained as principal components. Tab. VIII presents the estimated eigenvalues of the eight variables for the leaf nodes L1m2-L2m3-L3m4, L1m1-L2m2, L1m3-L2m1-L3m1, and L1m4-L2m2-L3m6. According to the factor selection criterion, L1m2-L2m3-L3m4, for instance, retains the first three factors as principal components, with Factor 1 explaining 42.155% of the total variance of the input variables, Factor 2 explaining 32.621% and Factor 3 explaining 15.823%.

To enhance the interpretability of the principal components obtained, the varimax factor rotation method [8] is used. This method minimizes the number of variables that have high loadings of a factor. To differentiate the features of each principal component, the variables with absolute values of the corresponding factor loadings less than 0.6 are omitted. Tab. IX shows the results of the varimax factor rotation method for the leaf nodes L1m2-L2m3-L3m4, L1m1-L2m2, L1m3-L2m1-L3m1, and L1m4-L2m2-L3m6. As Canbas et al. [8] demonstrated, an early-warning model for the observations can be estimated based on these major factor loadings.

As shown in Tab. IX, the principal components extracted for the different leaf-nodes are heterogeneous in the composition of their variables. For instance, the first principal component of leaf node L1m2-L2m3-L3m4 consists of four liquidity-related ratios (CR, QR, CFR, and CFAR), its second principal component consists of three earnings- and debt-related ratios (ROA, DR, and the Z-score), and its third principal component consists of one corporate governance ratio (SPR). Hence, the first principal component represents the liquidity of a firm, so it can be named as the liquidity factor. The second principal component represents the earnings and debt of a firm, so it can be named as debt paying ability factor. The third principal component represents the corporate governance health of a firm, so it can be named as corporate governance factor. For leaf node L1m1-L2m2, the first principal component (consisting of CR, DR, and SPR) represents the debt-paying

Node	Factor	Eigenvalue	% of Variance
L1m2-L2m3-L3m4	1	3.372417996	42.2
	2	2.609656893	32.6
	3	1.265835653	15.8
	4	0.467570397	5.8
	5	0.284519061	3.6
	6	1.34839E-16	0.0
	7	-7.15835E-17	0.0
	8	-1.95125E-16	0.0
L1m1-L2m2	1	3.300332163	41.3
	2	2.710120887	33.9
	3	1.004873386	12.6
	4	0.720253691	9.0
	5	0.150821476	1.9
	6	0.089632737	1.1
	7	0.023965661	0.3
	8	1.76194E-16	0.0
L1m3-L2m1-L3m1	1	2.615527285	32.7
	2	1.592850616	19.9
	3	1.183232213	14.8
	4	0.924914079	11.6
	5	0.745150561	9.3
	6	0.48356399	6.0
	7	0.271147814	3.4
	8	0.183613441	2.3
L1m4-L2m2-L3m6	1	5.019731655	62.7
	2	2.606681384	32.6
	3	0.373586961	4.7
	4	5.21943E-16	0.0
	5	1.59383E-16	0.0
	6	5.39578E-17	0.0
	7	-1.109E-16	0.0
	8	-2.53053E-16	0.0

Tab. VIII *The estimated eigenvalues of eight variables for leaf nodes L1m2-L2m3-L3m4, L1m1-L2m2, L1m3-L2m1-L3m1, and L1m4-L2m2-L3m6.*

ability and financial pressure of a firm, so it can be named as debt pressure factor. The second principal component (consisting of ROA, QR, and CFR) represents the earnings and liquidity of a firm, so it can be named as the earning power and liquidity factor. The third principal component (consisting of CFAR) represents the cash flow of a firm, so it can be named as the cash flow factor. For leaf node L1m3-L2m1-L3m1, the first principal component (consisting of CR, CFR, and CFAR) represents the liquidity of a firm, so it can be named as the liquidity factor. The second principal component (consisting of ROA and the Z-score) represents the earning ability and financial health of a firm, so it can be named as the earning

Node	Comp.	ROA	CR	QR	DR	CFR	CFAR	SPR	Z-score
L1m2-	1		-0.754	-0.742		0.875	0.919		
L2m3-	2	0.854			-0.971				0.93
L3m4	3							0.955	
L1m1-	1		0.913		-0.974			0.770	
L2m2	2	0.867		0.809		0.677			
	3						0.957		
L1m3-	1		-0.683			0.779	0.900		
L2m1-	2	0.921							0.892
L3m1	3			0.872					
L1m4-	1	0.965		0.883		0.984	0.984		0.917
L2m2-	2		0.945		0.871			0.985	
L3m6									

Tab. IX *The factor loadings of the principal components for the leafnodes L1m2-L2m3-L3m4, L1m1-L2m2, L1m3-L2m1-L3m1, and L1m4-L2m2-L3m6. Corresponding factor loadings with absolute values less than 0.6 are omitted.*

power and bankruptcy factor. The third principal component (consisting of QR) represents the liquidity of a firm, so it can be named as the short-term financial liabilities factor. For leaf node L1m4-L2m2-L3m6, the first principal component (consisting of ROA, QR, CFR, CFAR, and the Z-score) represents the composite financial health of a firm, so it can be named as the financial distress factor. The second principal component (consisting of CR, DR and SPR) represents the debt-paying ability of and financial pressure on a firm, so it can be named as debt pressure factor. The physical meanings of these results are discussed below.

The samples in leaf node L1m2-L2m3-L3m4 have a 66.67% probability of committing FFR, and the risk factors are CR, QR, CFR, and CFAR, which represent the liquidity and the debt-paying ability of a company. The parent node of L1m2-L2m3-L3m4 is L1m2, which has the highest fraud ratio (60.61%) among the nodes in layer 1. Therefore, any sample belonging to L1m2-L2m3-L3m4 should be on high alert.

The samples in leaf node L1m1-L2m2 have a 37.5% probability of committing FFR, and the risk factors are CR, DR, and SPR, which represent the debt-paying ability of a company. The parent node of L1m1-L2m2 is L1m1, which has the second largest fraud ratio (28.57%) among the nodes in layer 1. Therefore, any sample belonging to L1m1-L2m2 should be highly concerned as well.

The samples in leaf node L1m3-L2m1-L3m1 have a 20% potentiality of the fraud ratio to commit FFR, and the risk factors are CR, CFR, and CFAR, which represent the liquidity of a company. Although the parent node of L1m3-L2m1-L3m1 is L1m3, which has the lowest fraud ratio (8.11%) among the nodes in layer 1, L1m3-L2m1-L3m1 is the riskiest node among the child nodes of L1m3 that have more than one sample. Therefore, any sample belonging to L1m3-L2m1-L3m1 should be monitored.

The samples in leaf node L1m4-L2m2-L3m6 have a 60% probability of committing FFR, and the risk factors are ROA, QR, CFR, CFAR, and Z-score, which represent the overall financial health of a company. Although the parent node of L1m4-L2m2-L3m6 is L1m4, which has the second lowest fraud ratio (17.32%)

among the nodes in layer 1, L1m4-L2m2-L3m6 is the riskiest node among the child nodes of L1m4, and the fraud ratio of L1m4-L2m2-L3m6 is obviously high. Therefore, any sample belonging to L1m4-L2m2-L3m6 should be further investigated as well.

5. Conclusions

In summary, this study proposes a quantitative approach with three stages — data-preprocessing, clustering, and feature extraction — in each of which quantitative tools are used intensively. This study demonstrates the advantages of an approach that adopts the GHSOM in the clustering stage and extracts the features of subgroups via PCA. The experimental results show that GHSOM leads to better-quality clustering than SOM, K-means, and the two-step clustering method. Furthermore, from the different leaf nodes of the GHSOM, PCA extracts distinctive principal components with associated financial meanings. These results confirm the following theoretical benefits of GHSOM. First because of its unsupervised learning nature, GHSOM involves no predefined categories into which samples must be classified; rather, GHSOM develops its own feature representation of a sample via a competitive learning algorithm. Second, the GHSOM classifies the sample into many small-sized leaf nodes with hierarchical relationships, making more delicate analyses feasible. Third, due to its competitive learning nature, GHSOM works as a regularity detector that discovers the statistically salient features of the sample population [39]. That is, the extracted features of different leaf nodes are distinctive.

The experimental results also show that the proposed quantitative approach using GHSOM and PCA is helpful in identifying useful features and can be used to help detect deception regarding FFR or other financial distress scenarios. For instance, with samples clustered into many small-sized leaf nodes, a linear feature-extracting tool (e.g., PCA) can be used to extract the features of each leaf node. The principal components extracted from different leaf nodes have distinctive features and thus can contribute to further pattern analysis. For instance, the first principal component of leaf node L1m2-L2m3-L3m4, which has a high fraud ratio, represents the liquidity of a firm and thus suggests a hypothesis that firms in this subgroup have liquidity pressure and are weak in terms of short-term debt-paying ability such that they tend to record fictitious revenues, which is a type of FFR, to manipulate earnings-related financial indicators. For another high-risk leaf node, L4-L2m2-L3m6, the first principal component represents the composite financial health status of a firm and thus suggests a hypothesis that firms in this subgroup have poor cash flow conditions and thus overall financial pressure such that they tend to overstate existing assets, also a type of FFR, to manipulate cash flow-related financial indicators. Anyone interested in a decision support system for identifying FFR should further verify these hypotheses and establish warning signals based upon the confirmed hypotheses. Designing such a decision support system based upon the proposed quantitative approach is among our future research objectives. A limitation of this study is the subjective parameter setting of the GHSOM, which is common to this line of research.

The following additional directions for future work are suggested: (1) further discussion of the relationships among subgroups of GHSOM; (2) testing of the predictive capability of an early warning model based on the major factor loadings obtained; and (3) applying the proposed approach to other financial scenarios, such as bankruptcy prediction and credit rating. For instance, to predict bankruptcy or to rate credit worthiness, we propose the following three-stage approach: a data-preprocessing stage, a clustering stage and a risk-scoring stage. In the clustering stage, a classification technique is used to cluster the samples into subgroups. Prediction tools, such as the support vector machine (SVM) [47] or linear discriminant analysis (LDA) [9], can be applied to each subgroup to develop a score as a function of credit risk level. In the risk-scoring stage, the score function can be used to predict the credit risk of any investigated sample, based upon the subgroup to which it belongs.

References

[1] ALTMAN E.I. Financial ratios discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*. 1968, 23(4), pp. 589-609, doi: 10.1111/j.1540-6261.1968.tb00843.x.

[2] ASSOCIATION OF CERTIFIED FRAUD EXAMINERS. 2008 Report to the Nation on Occupational Fraud & Abuse [online]. USA: Association of Certified Fraud Examiners, 2008 [viewed 2014-10-15]. Available from: http://www.acfe.com/uploadedFiles/ACFE_Website/Content/documents/2008-rttn.pdf.

[3] ASSOCIATION OF CERTIFIED FRAUD EXAMINERS. 2010 Report to the Nation on Occupational Fraud & Abuse [online]. USA: Association of Certified Fraud Examiners, 2010 [viewed 2014-10-15]. Available from: http://www.acfe.com/uploadedFiles/ACFE_Website/Content/documents/rttn-2010.pdf.

[4] ASSOCIATION OF CERTIFIED FRAUD EXAMINERS. 2012 Report to the Nation on Occupational Fraud & Abuse [online]. USA: Association of Certified Fraud Examiners, 2012 [viewed 2014-10-15]. Available from: https://www.acfe.com/uploadedFiles/ACFE_Website/Content/rttn/2012-report-to-nations.pdf.

[5] BASENS B. et al. Using neural network rule extraction and decision tables for credit-risk evaluation. *Management Science*. 2003, 49(3), pp. 312-319, doi: 10.1287/mnsc.49.3.312.12739.

[6] BEASLEY M.S., CARCELLO, J.V., HERMANSON, D.R. *Fraudulent financial reporting: 1987-1997 An analysis of U.S. public companies.* New York: Committee of Sponsoring Organizations of the Treadway Commission (COSO), 1999. Research report.

[7] BELL T.B., CARCELLO J.V. A Decision Aid for Assessing the Likelihood of Fraudulent Financial Reporting. *Auditing*. 2000, 19(1), pp. 169-184, doi: 10.2308/aud.2000.19.1.169.

[8] CANBAS S. et al. Prediction of commercial bank failure via multivariate statistical analysis of financial structures: The Turkish case. *European Journal of Operational Research*. 2005, 166(2), pp. 528-546, doi: 10.1016/j.ejor.2004.03.023.

[9] CARLOS S.C. Self-organizing neural networks for financial diagnosis. *Decision Support System*. 1996, 17(3), pp. 227-238, doi: 10.1016/0167-9236(95)00033-X.

[10] DE BODT E. et al. Statistical tools to assess the reliability of self-organizing maps. *Neural Networks*. 2002, 15(8-9), pp. 967-978, doi: 10.1016/S0893-6080(02)00071-0.

[11] DECHOW P.M. et al. Predicting material accounting manipulations. *Contemporary Accounting Research*. 2011, 28(1), pp. 17-82.

[12] DITTENBACH M. et al. The Growing Hierarchical Self-Organizing Map. In: *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks- IJCNN*, Como, Italy. IEEE Computer Society, 2000, pp. 15-19.

[13] DITTENBACH M. et al. Uncovering hierarchical structure in data using the growing hierarchical self-organizing map. *Neurocomputing*. 2002, 48(1-4), pp. 199-216, doi: 10.1016/S0925-2312(01)00655-5.

[14] ELLIOT R., WILLINGHAM J. *Management fraud: detection and deterrence*. Petrocelli, New York: Petrocelli Books, Inc., 1980.

[15] FANNING K.M., COGGER K.O. Neural network detection of management fraud using published financial data. *International Journal of Intelligent Systems in Accounting, Finance & Management*. 1998, 7(1), pp. 21-41, doi: 10.1002/(SICI)1099-1174(199803)7:13.3.CO;2-B.

[16] FICH E.M., SLEZAK S.L. Can corporate governance save distressed firms from bankruptcy? An empirical analysis. *Review of Quantitative Finance and Accounting*. 2008, 30(2), pp. 225-251, doi: 10.1007/s11156-007-0048-5.

[17] GREEN P., CHOI J.H. Assessing the risk of management fraud through neural network technology. *Auditing: A Journal of Practice & Theory*. 1997, 16(1), pp. 14-28.

[18] GUPTA R., GILL N.S. Prevention and Detection of Financial Statement Fraud – An Implementation of Data Mining Framework. *International Journal of Advanced Computer Science and Applications*. 2012, 3(8), pp. 150-156.

[19] HART C.L. et al. Indirect Detection of Deception: Looking for Change. *Current Research in Social Psychology*. 2009, 14(9), pp. 134-142, doi: 10.1.1.364.1259.

[20] HOOGS B. et al. A genetic algorithm approach to detecting temporal patterns indicative of financial statement fraud. *Intelligent Systems in Accounting Finance and Management*. 2007, 15(1/2), pp. 41-56, doi: 10.1002/isaf.284.

[21] HSU A.L., HALGAMUGE S.K. Enhancement of topology preservation and hierarchical dynamic self-organising maps for data visualisation. *International Journal of Approximate Reasoning*. 2003, 32(2-3), pp. 259-279, doi: 10.1016/S0888-613X(02)00086-5.

[22] HUANG S.Y. et al. Unsupervised Neural Networks Approach for Understanding Fraudulent Financial Reporting. *Industrial Management & Data Systems*. 2012, 112(2), pp. 224-244.

[23] HUMPHERYS S.L. et al. Identification of fraudulent financial statements using linguistic credibility analysis. *Decision Support Systems*. 2011, 50(3), pp.585-594, doi: 10.1016/j.dss.2010.08.009.

[24] JOLLIFFE I.T. *Principal Component Analysis*. New York: Springer, 1986.

[25] JOLLIFFE I.T. *Principal Component Analysis*. New York: Springer, 2002.

[26] KAISER H.F. The application of electronic computers to factor analysis. *Educational and Psychological Measurement*. 1960, 20(1), pp. 141-151, doi: 10.1177/001316446002000116.

[27] KIANG M.Y. Extending the Kohonen self-organizing map networks for clustering analysis. *Computational Statistics & Data Analysis*. 2001, 38(2), pp. 161-180, doi: 10.1016/S0167-9473(01)00040-8.

[28] KIRKOS E. et al. Data mining techniques for the detection of fraudulent financial statements. *Expert Systems with Applications*. 2007, 32(4), pp. 995-1003, doi: 10.1016/j.eswa.2006.02.016.

[29] KOHONEN T. Self-organized formation of topologically correct features maps. *Biol. Cyber*. 1982, 43(1), pp. 59-69, doi: 10.1007/BF00337288.

[30] LA PORTA R. et al. Corporate ownership around the world. *Journal of Finance*. 1999, 54(2), pp. 471-517.

[31] LEE T.S., YEH Y.H. Corporate governance and financial distress: evidence from Taiwan. *Corporate Governance: An International Review*. 2004, 12(3), pp. 378-388, doi: 10.1111/j.1467-8683.2004.00379.x.

[32] LIOU F.M. Fraudulent financial reporting detection and business failure prediction models: a comparison. *Managerial Auditing Journal*. 2008, 23(7), pp. 650-662, doi: 10.1108/02686900810890625.

[33] LOEBBECKE J.K. et al. Auditors' experience with material irregularities: frequency, nature, and detectability. *Auditing*. 1989, 9(1), pp. 1-28.

[34] LU C.J., WANG Y.W. Combining Independent Component Analysis and Growing Hierarchical Self-Organizing Maps with Support Vector Regression in Product Demand Forecasting. *International Journal of Production Economics*. 2010, 128(2), pp. 603-613.

[35] MIN J.H., LEE Y.C. Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert Systems with Applications*. 2005, 28(4), pp. 603-614, doi: 10.1016/j.eswa.2004.12.008.

[36] PAMPALK E. et al. A new approach to hierarchical clustering and structuring of data with Self-Organizing Maps. *Journal of Intelligent Data Analysis*. 2004, 8(2), pp. 131-149.

[37] PERSONS O.S. Using financial statement data to identify factors associated with fraudulent financial reporting. *Journal of Applied Business Research*. 1995, 11(3), pp. 38-46.

[38] RAUBER A. et al. The Growing Hierarchical Self-Organizing Map: Exploratory Analysis of High-Dimensional Data. *IEEE Transactions on Neural Networks*. 2002, 13(6), pp. 1331-1341, doi: 10.1109/TNN.2002.804221.

[39] RUMELHART D.E., ZIPSER D. Feature discovery by competitive learning. *Cognitive Science*, 1985, 9(1), pp. 75-112, doi: 10.1207/s15516709cog0901_5.

[40] RUSINKIEWICZ S., LEVOY M. QSplat: A Multiresolution Point Rendering System for Large Meshes. In: *Proceedings of the 27th annual conference on computer graphics and interactive techniques*, New York, USA. New York: ACM, 2000, pp. 343-352.

[41] SCHWEIGHOFER E. et al. Automatic text representation classification and labeling in European law. In: *Proceedings of the International Conference on Artificial Intelligence and Law (ICAIL)*, Amsterdam, Netherlands. New York: ACM, 2001, pp. 78-87, doi: 10.1145/383535.383544.

[42] SHI H. et al. A method for classifying packets into network flows based on GHSOM. *Mobile Networks and Applications*. 2012, 17(6), pp. 730-739, doi: 10.1007/s11036-012-0383-1.

[43] SHIH J.Y., et al. Using GHSOM to construct legal maps for Taiwan's securities and futures markets. *Expert Systems with Applications*. 2008, 34(2), pp. 850-858, doi: 10.1016/j.eswa.2006.10.031.

[44] STICE J.D. Using financial and market information to identify pre-engagement factors associated with lawsuits against auditors. *Accounting Review*. 1991, 66(3), pp.516-533.

[45] SUMMERS S.L., SWEENEY J.T. Fraudulently misstated financial statements and insider trading: An empirical analysis. *Accounting Review*, 1998. 73(1), pp. 131-146.

[46] TSAIH R.H. et al. Exploring fraudulent financial reporting with GHSOM. In H. Chen, ed. *Intelligence and Security Informatics*. Berlin Heidelberg: Springer, 2009, pp. 31-41, doi: 10.1007/978-3-642-01393-5_5.

[47] VAPNIK V.N. *The nature of statistical learning theory*. New York: Springer, 1995.

[48] WICKRAMASINGHE L.K. et al. A hybrid intelligent multiagent system for e-business. *Computational Intelligence*. 2004, 20(4), pp. 603-623, doi:10.1111/j.0824-7935.2004.00256.x.

[49] YANG H.C. et al. A method for multilingual text mining and retrieval using growing hierarchical self-organizing maps. *Journal of Information Science*. 2009, 35(1), pp. 3-23, doi: 10.1177/0165551508088968.

[50] YANG Y.X., TSAIH R.H. An Investigation of Research on Evolution of Altruism using Informatics Methods and the Growing Hierarchical Self-Organizing Map. *Malaysian Journal of Library & Information Science*. 2010, 15(3), pp. 1-17.

[51] YEH Y. et al. Family control and corporate governance: Evidence from Taiwan. *International Review of finance*. 2001, 2(1-2), pp. 21-48.

[52] YOUNG C.S. et al. The effect of controlling shareholders' excess board seats control on financial restatements: evidence from Taiwan. *Review of Quantitative Finance and Accounting*. 2008, 30(3), pp. 297-314, doi: 10.1007/s11156-007-0054-7.

[53] ZHAO Y. et al. Effects of takeover protection on earnings overstatements: evidence from restating firms. *Review of Quantitative Finance and Accounting*. 2009, 33(4), pp. 347-369, doi: 10.1007/s11156-007-0054-7.