International Journal of Accounting Information Systems 32 (2019) 1–23

Contents lists available at ScienceDirect

International Journal of Accounting Information Systems

journal homepage: www.elsevier.com/locate/accinf

Fraud detection for financial statements of business groups

Yuh-Jen Chen[a,*], Wan-Ching Liou[a], Yuh-Min Chen[b], Jyun-Han Wu[b]

[a] *Department of Accounting and Information Systems, National Kaohsiung University of Science and Technology, Kaohsiung, Taiwan, ROC*
[b] *Institute of Manufacturing Information and Systems, National Cheng Kung University, Tainan, Taiwan, ROC*

ARTICLE INFO

Keywords:
Fraud detection
Business group
Financial statement
Texting mining

ABSTRACT

Investors rely on companies' financial statements and economic data to inform their investment decisions. However, many businesses manipulate financial statements to raise more capital from investors and financial institutions, which reduces the practicality of financial statements. The modern business environment is highly information-oriented, and firms' information systems and activities are complex and dynamic. Technology for avoiding fraud detection is continually updated. Recent studies have focused on detecting financial statement fraud within a single business, but not within a business group. Development of methods for using diverse data to detect financial statement fraud in business groups is thus a high priority in the advancement of fraud detection.

This study develops an approach for detecting fraud in the financial statements of business groups. The proposed approach is applied to reduce investment losses and risks and enhance investment benefits for investors and creditors. The study objectives are achieved through the following steps: (i) design of a process for detecting fraud in the financial statements of business groups, (ii) development of fraud detection techniques for use with such statements, and (iii) demonstration and evaluation of the proposed approach.

1. Introduction

Since the Procomp case and Enron event, investors, governments, and regulatory authorities have begun to focus on financial statement fraud committed by business groups. Falsified financial statements may result in large losses for investors and creditors in capital markets. The modern business environment is highly information-oriented, and firms' systems and activities are complex and dynamic. Technology used to avoid fraud detection is constantly updated. Development of methods for using diverse data to detect financial statement fraud in business groups is thus a high priority in the advancement of fraud detection.

Various approaches have been developed to detect fraud in corporate financial statements. Kirkos et al. (2007) explored the effectiveness of data mining (DM) classification techniques for detecting firms' fraudulent financial statements (FFS) and identified factors associated with FFS. Auditors assisted in detecting fraud using DM techniques. The study also investigated the usefulness of decision trees, neural networks, and Bayesian belief networks in identifying FFS. Ravisankar et al. (2011) also used DM techniques such as multilayer feed forward neural network (MLFF), support vector machine (SVM), genetic programming (GP), group method of data handling (GMDH), logistic regression (LR), and probabilistic neural network (PNN) to identify companies that had committed financial statement fraud. Each of these techniques was tested on a dataset covering 202 Chinese companies, and the results of tests with and without feature selection were compared. Among the techniques, PNN was the most accurate without feature selection, and GP and PNN were the most accurate with feature selection (with marginally equal accuracies). Glancy and Yadav (2011) proposed a

[*] Corresponding author.
E-mail address: yjchen@nkust.edu.tw (Y.-J. Chen).

https://doi.org/10.1016/j.accinf.2018.11.004
Received 14 June 2017; Received in revised form 1 November 2018; Accepted 29 November 2018
Available online 06 December 2018

quantitative model for detecting attempts to conceal information or present incorrect information in annual filings within the US Securities and Exchange Commission. The model essentially used all the information contained in a text document for fraud detection and was validated as a consistently accurate screening tool for early detection of fraud. Zhou and Kapoor (2011) considered DM-based financial fraud detection techniques (such as regression, decision trees, neural networks, and Bayesian networks) that support fraud identifications. In their study, the effectiveness of these DM methods and their limitations were examined, particularly regarding the adaptability of the methods for new fraud detection schemes. A self-adaptive framework (based on a response surface model) with domain knowledge was then employed to detect financial statement fraud. Gray and Debreceny (2014) explored the application of DM techniques in fraud detection within financial statement audits and proposed a taxonomy for guiding future research. To develop a structure for research on DM, a taxonomy was created that combined research into observed fraud scheme patterns with an appreciation of areas that benefit from the productive application of DM. In addition, traditional views of DM were encapsulated to ensure that the mining primarily operated on quantitative data, such as financial statements. Dutta et al. (2017) employed all widely used data mining techniques to detect fraudulent financial restatements, including decision tree (DT), artificial neural network (ANN), naïve Bayes (NB), support vector machine (SVM), and Bayesian belief network (BBN) Classifier. The prior studies have focused primarily on using DM techniques to detect financial statement fraud within a single enterprise; few studies have focused on detecting such fraud within an entire business group by using diverse data. Thus, business groups are increasingly creating FFS when searching for buyers in corporate mergers, acquisitions, or stock undertaking, and increasing the risk for investors in capital markets.

Therefore, this study develops an approach for detecting the FFS of business groups that has high accuracy and thereby the potential to reduce investment losses and risks and benefit investors and creditors. This is achieved through the following steps: (i) design of a process for detecting fraud in the financial statements of business groups, (ii) development of fraud detection techniques for use with such statements, and (iii) demonstration and evaluation of the proposed approach.

The remainder of this paper is organized as follows. In Section 2, the design of a fraud detection process for application to the financial statements of business groups is presented. Section 3 then develops the techniques involved in detecting the FFS of business groups. Subsequently, Section 4 demonstrates the effectiveness of the proposed approach. Conclusions are drawn in Section 5, and recommendations for future research are also proposed.

2. Design of a fraud detection process for financial statements of business groups

In this section, the operational models used in FFS of business groups are first identified. Based on these models and the concept of diverse data, fraud detection processes for application to the financial statements of business groups are then designed.

2.1. Operational models of financial statement fraud for business groups

Based on the survey of infamous financial statement fraud cases related business groups (Chary, 2004; Kaplan and Kiron, 2004; Suraj and Sesia, 2011; Swartz and Watkins, 2003) (including Enron (2001), WorldCom (2002), Tyco (2004), Infodisc (2003), Procomp (2004) and China Rebar (2006)) as well as financial statement fraud definitions and types (*Ten Things About Financial Statement*

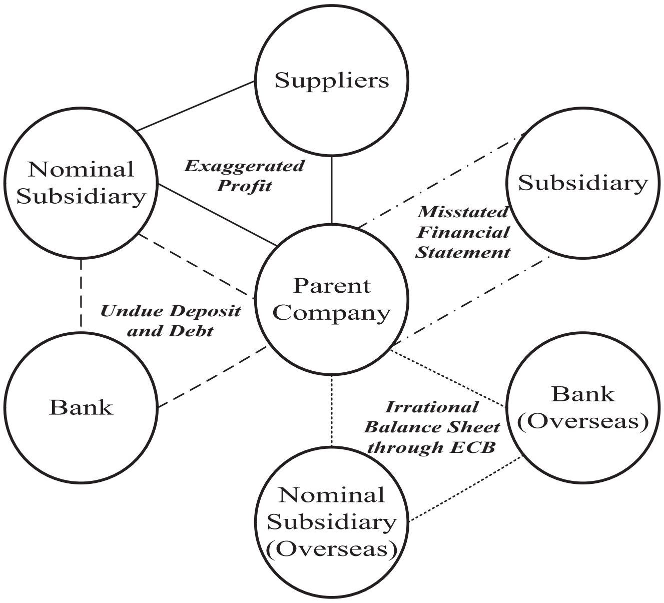

Fig. 1. Four fraud types of business group financial statements.

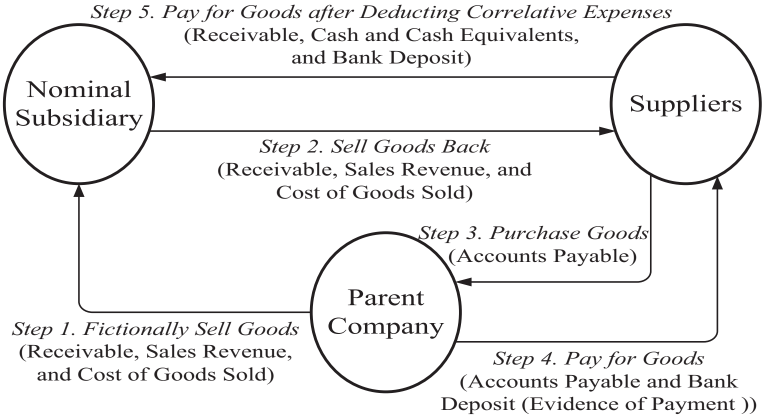
Step 5. Pay for Goods after Deducting Correlative Expenses
(Receivable, Cash and Cash Equivalents, and Bank Deposit)
Nominal Subsidiary
Suppliers
Step 2. Sell Goods Back
(Receivable, Sales Revenue, and Cost of Goods Sold)
Step 3. Purchase Goods
(Accounts Payable)
Parent Company
Step 1. Fictionally Sell Goods
(Receivable, Sales Revenue, and Cost of Goods Sold)
Step 4. Pay for Goods
(Accounts Payable and Bank Deposit (Evidence of Payment))

Fig. 2. Exaggerated profit process.

Fraud - Third Edition, a Review of SEC Enforcement Releases, 2000–2008, 2008; Nguyen, 2010), four financial statement fraud types for business groups are modeled (Fig. 1) based on a commonly used business analysis technique (business process model) (Vernadat, 1996) that can identify how a fraudulent business process works by providing a graphical notation for presenting business fraudulent activities. Meanwhile, the fraud practices of Procomp involved in exaggerated profit, undue deposit and debt, misstated financial statement, and irrational balance sheet through ECB (Euro-Convertible Bond) in 2004, while the fraud practices of Enron involved in misstated financial statement and exaggerated profit in 2001. WorldCom occurred exaggerated profit in 2002 and Tyco happened misstated financial statement in 2004. Infodisc involved in the frauds of exaggerated profit and irrational balance sheet through ECB in 2003. In 2006, China Rebar had undue deposit and debt.

In cases of exaggerated profit, a nominal subsidiary company that purchases businesses or falsely trades with the parent company is established to increase both the sales and operating revenue of the parent company. The nominal subsidiary company then results the goods purchased from the parent company to the suppliers of the parent company to form an input and output cycle of the same goods. Thus, the operating revenue of the parent company is manipulated, as represented in Fig. 2. For undue deposit and debt (Fig. 3), endorsements and guarantees for a bank loan are first proposed by the parent company. The bank then fulfills a loan to the nominal subsidiary company, which subsequently uses the loan to pay the parent company and thereby manipulate the parent company's financial structure. As illustrated in Fig. 4, in instances of misstated financial statements, the subsidiary company of a business group offers the parent company financial statement information that is inconsistent with the real data, and the parent company then uses this false information compile consolidated financial statements. These financial statements are thus questionable. Finally, irrational balance sheet through ECB, a nominal overseas subsidiary company receives a loan from an overseas bank to purchase ECBs issued by the parent company. After issuing the ECBs, the parent company usually illegally lifts the stock value, thereby changing the ECBs held by the nominal overseas subsidiary company into common stocks. Consequently, the nominal overseas subsidiary company can sell the stocks to earn the difference between the prices, and the parent company can reduce their debt ratio by changing the ECBs into capital stocks to manipulate the parent company's financial structure. This process is depicted in Fig. 5.

2.2. *Fraud detection process for financial statements of business groups*

Chen et al. (2015) stated that the use of diverse data for fraud risk management is a new trend in the world of transactions. A new

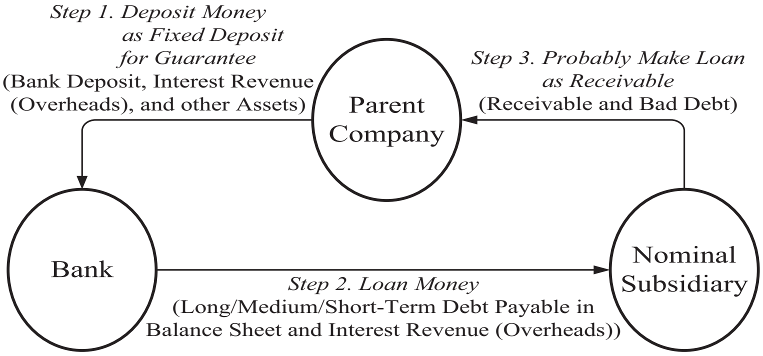
Step 1. Deposit Money as Fixed Deposit for Guarantee
(Bank Deposit, Interest Revenue (Overheads), and other Assets)
Parent Company
Step 3. Probably Make Loan as Receivable
(Receivable and Bad Debt)
Bank
Nominal Subsidiary
Step 2. Loan Money
(Long/Medium/Short-Term Debt Payable in Balance Sheet and Interest Revenue (Overheads))

Fig. 3. Undue deposit and debt process.

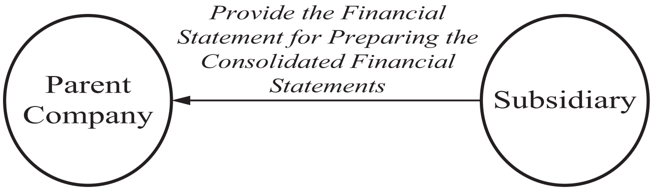
Parent Company
Provide the Financial Statement for Preparing the Consolidated Financial Statements
Subsidiary

Fig. 4. Misstated financial statement process.

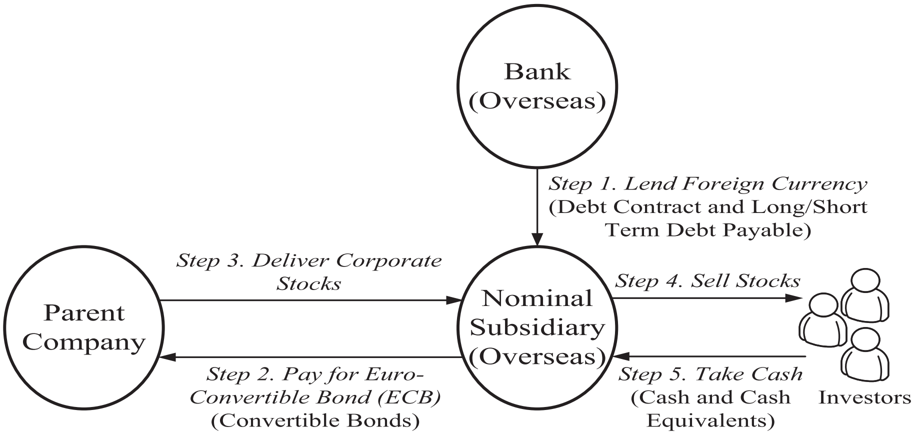
Bank (Overseas)
Step 1. Lend Foreign Currency (Debt Contract and Long/Short Term Debt Payable)
Parent Company
Step 3. Deliver Corporate Stocks
Nominal Subsidiary (Overseas)
Step 4. Sell Stocks
Step 2. Pay for Euro-Convertible Bond (ECB) (Convertible Bonds)
Step 5. Take Cash (Cash and Cash Equivalents)
Investors

Fig. 5. Irrational balance sheet through ECB process.

generation of fraud-monitoring techniques that employ processing of diverse data, computing technology, real-time fraud prevention systems, and risk models has emerged. Tian et al. (2015) indicated that data analytics offers numerous benefits to banking and financial market firms in tasks such as accurate customer analytics, risk analysis, and fraud detection. These approaches can facilitate intelligent trading, which can assist organizations in avoiding latent risks and providing more personalized services, thereby increasing firms' competitive advantage. Jina et al. (2015) compared traditional data and diverse data from new data resources. Moreover, Gepp et al. (2018) discovered that diverse data was used less in auditing than in other related fields. Literature on the use of diverse data in accounting and finance encompasses three genealogies: financial distress modeling, financial fraud modeling, and stock market prediction and quantitative modeling. Auditing is lagging behind the other fields in research on the use of diverse data. The analysis of diverse data is not only important for the operational decision of a single company, but also more necessary for business groups. In this section, the concept of diverse data is thus applied to identify the useful internal and external financial data (different quantitative and qualitative financial data) of business groups for designing processes for detecting fraud in financial statements. The proposed fraud detection processes comprise techniques for identifying exaggerated profit, undue deposit and debt, misstated financial statement, and irrational balance sheet through ECB. Each fraud detection process involves different data sources and processing methods. The relevant data sources used to detect the exaggerated profit are identified, including auditor's review reports (textual/internal data), financial news (textual/external data), stock trading volume (numerical/external data), financial statements (numerical/internal data), and security companies' predictive data (numerical/external data). The major data sources used to detect the undue deposit and debt include auditors' review reports (textual/internal data), letters to shareholders (textual/internal data) and financial news (textual/external data), debt credit rating (numerical/external data), and financial statements (numerical/internal data). The misstated financial statement detection uses auditor's review reports (textual/internal data), financial news (textual/external data), stock trading volume (numerical/external data), debt structure indicators (numerical/external data), and debt credit rating (numerical/external data). For detecting the irrational balance sheet through ECB, the useful data sources are letters to shareholders (textual/internal data), stock trading volume (numerical/external data), debt structure indicators (numerical/external data), corporate governance indicators (numerical/external data), and financial statements (numerical/internal data). The processing methods for the above-mentioned textual and numerical financial data involved in different fraud detection processes are employed. Indicator normalization is adopted to process the numerical data of stock trading volume, debt credit rating, and corporate governance indicators, while predictive data retrieval is used to capture the predictive data from security companies. The fraud detection processes with using different data sources and processing methods are described as follows:

Detection process for exaggerated profit: The Chinese Knowledge Information Processing Group (CKIP) Client (*Chinese Knowledge and Information Processing*, n.d.) is first used to pre-process *auditor review reports* and *financial news* relating to fraudulent and non-fraudulent financial statements. Preprocessing involves segmenting sentences into meaningful terms and tagging the part-of-speech (POS) characteristics of terms. According to the results of data preprocessing, term-pairs are combined, unimportant terms are deleted, and fraudulent feature terms are filtered to establish a feature term library for exaggerated profit (such as “訴訟”

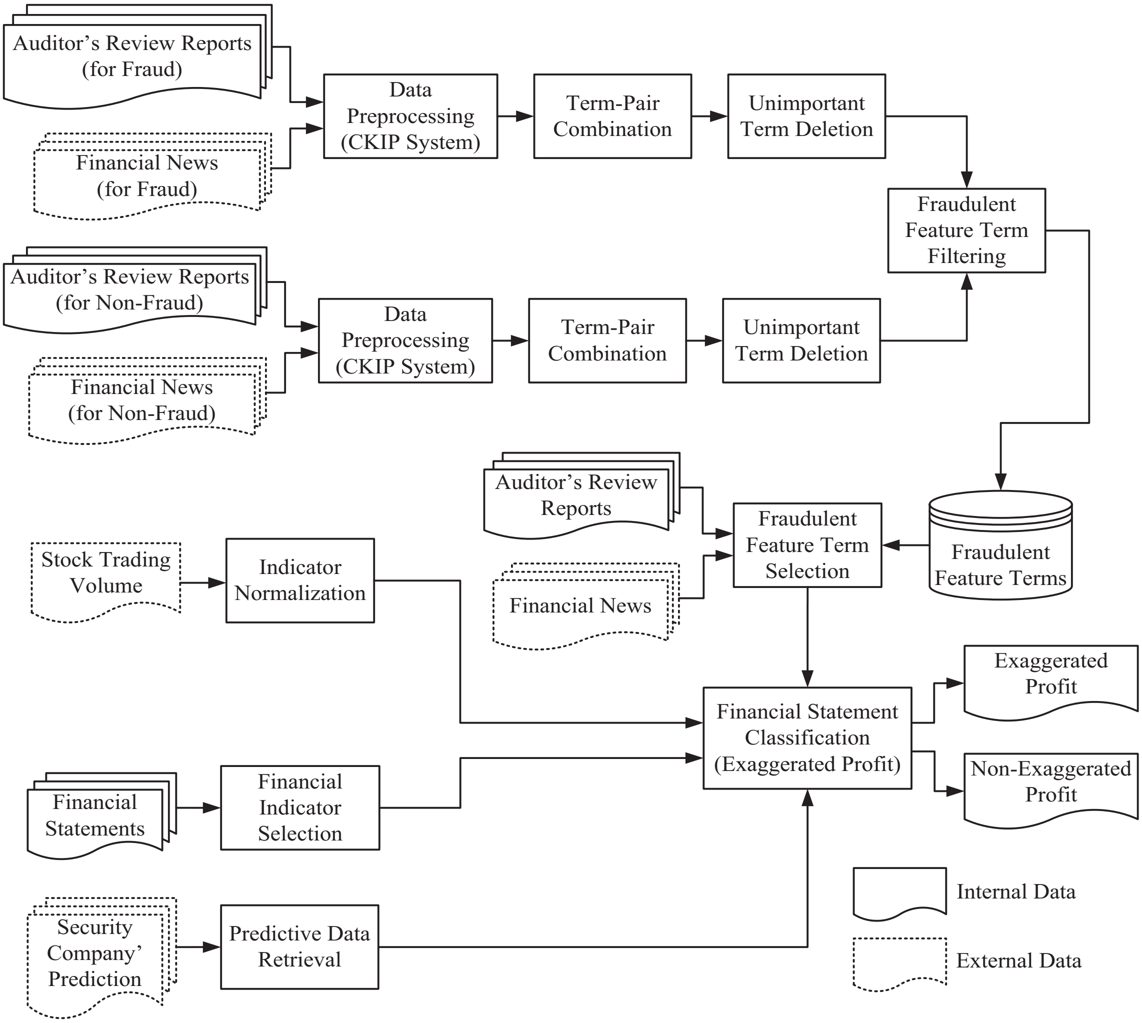
Auditor's Review Reports (for Fraud)
Financial News (for Fraud)
Data Preprocessing (CKIP System)
Term-Pair Combination
Unimportant Term Deletion
Fraudulent Feature Term Filtering
Auditor's Review Reports (for Non-Fraud)
Financial News (for Non-Fraud)
Data Preprocessing (CKIP System)
Term-Pair Combination
Unimportant Term Deletion
Auditor's Review Reports
Financial News
Fraudulent Feature Term Selection
Fraudulent Feature Terms
Stock Trading Volume
Indicator Normalization
Financial Statement Classification (Exaggerated Profit)
Exaggerated Profit
Non-Exaggerated Profit
Financial Statements
Financial Indicator Selection
Security Company' Prediction
Predictive Data Retrieval
Internal Data
External Data

Fig. 6. Detection process for exaggerated profit.

(litigation), "挪用" (misappropriate), "賠償" (indemnify) and "訴請" (appeal)). This library is then used for future identification of fraudulent feature terms in *auditors' review reports* and *financial news* relating to financial statements. In addition, an indicator of the *company's stock trading volume* is normalized, and the *relevant indicators of operating revenue* in financial statements are selected. *Predictive data* relating to operating revenue in financial statements, as predicted by security companies, are then retrieved. Finally, the results acquired from the fraudulent feature term selection, indicator normalization, financial indicator selection, and predictive data retrieval are input into the *queen genetic algorithm-support vector machine* (*QGA-SVM*) *classifier* for detecting exaggerated profits within the financial statements of business groups, as shown in Fig. 6.

Detection process for undue deposit and debt: The CKIP Client (Chinese Knowledge and Information Processing, n.d.) is first used to pre-process the *auditor's review reports*, *letters to shareholders*, and *financial news* relating to fraudulent and non-fraudulent financial statements. Preprocessing involves segmenting sentences into meaningful terms and tagging the POS characteristics of terms. According to the results of sentence segmentation and POS tagging, term-pair combination, unimportant term deletion, and fraudulent feature term filtering are sequentially performed to establish a feature term library for undue deposit and debt (including terms such as "重生" (rebirth), "導致" (resulting in), "進攻" (attack) and "謀求" (seek)). This library then serves as a base for future identification of fraudulent feature terms in *auditors' review reports*, *letters to shareholders*, and *financial news* related to financial statements. Furthermore, *deposit and debt indicators* in financial statements are selected and *indicators of debt credit rating* for business groups are normalized. Finally, undue deposit and debt in the financial statements of business groups are detected using the *QGA-SVM classifier* based on the results of fraudulent feature term selection, financial indicator selection, and indicator normalization, as illustrated in Fig. 7.

Detection process for misstated financial statement: As depicted in Fig. 8, *auditor's review reports* and *financial news* for fraudulent

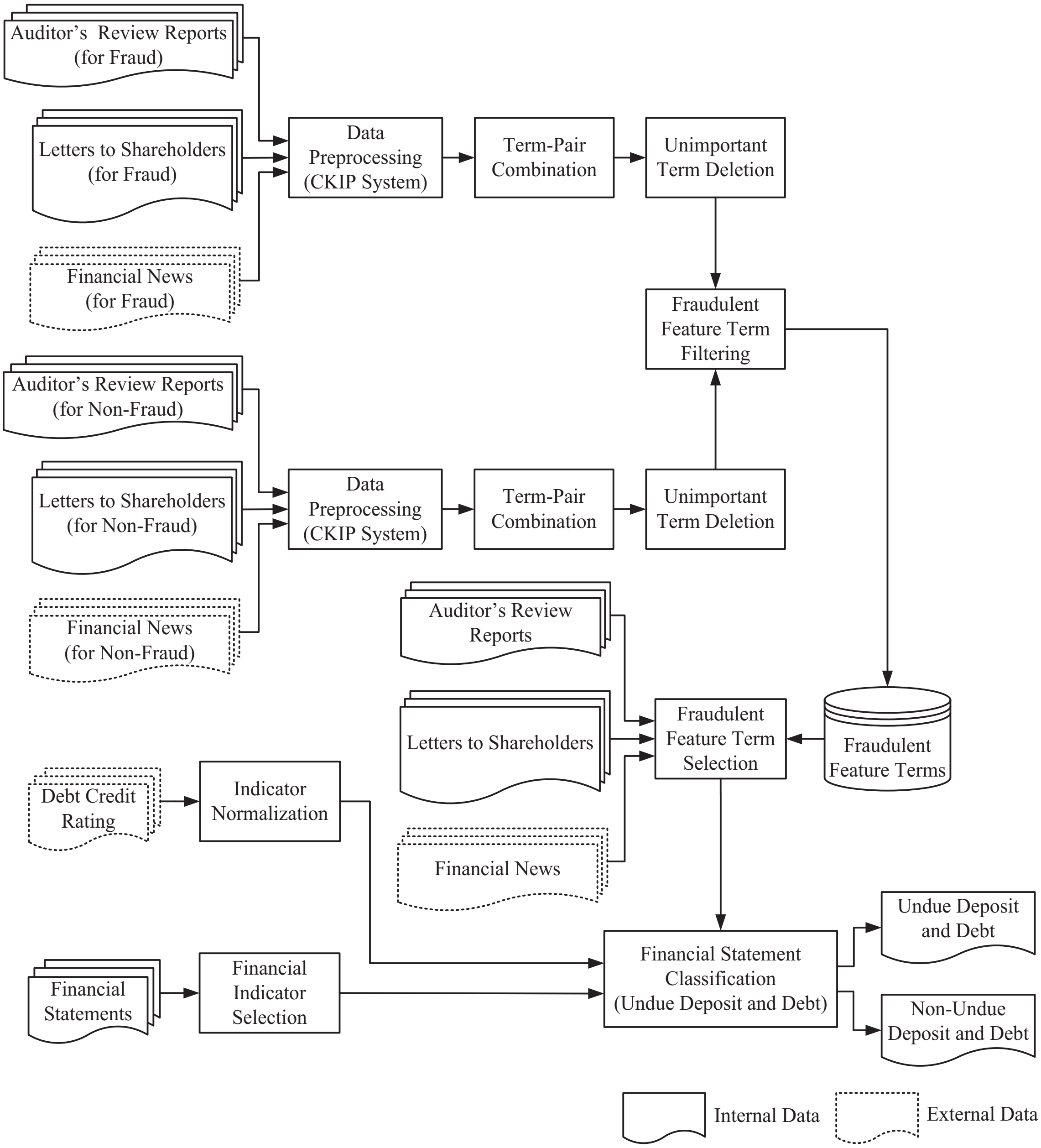
Auditor's Review Reports (for Fraud)
Letters to Shareholders (for Fraud)
Financial News (for Fraud)
Data Preprocessing (CKIP System)
Term-Pair Combination
Unimportant Term Deletion
Fraudulent Feature Term Filtering
Auditor's Review Reports (for Non-Fraud)
Letters to Shareholders (for Non-Fraud)
Financial News (for Non-Fraud)
Data Preprocessing (CKIP System)
Term-Pair Combination
Unimportant Term Deletion
Auditor's Review Reports
Letters to Shareholders
Financial News
Fraudulent Feature Term Selection
Fraudulent Feature Terms
Debt Credit Rating
Indicator Normalization
Financial Statements
Financial Indicator Selection
Financial Statement Classification (Undue Deposit and Debt)
Undue Deposit and Debt
Non-Undue Deposit and Debt
Internal Data
External Data

Fig. 7. Detection process for undue deposit and debt.

and non-fraudulent financial statements are first preprocessed through the CKIP Client (Chinese Knowledge and Information Processing, n.d.), and preprocessing involves segmenting sentences into meaningful terms and tagging the POS characteristics of terms. Term-pair combination, unimportant term deletion, and fraudulent feature term filtering are then executed based on the auditor's review reports and financial news preprocessing, with the aim of establishing a feature term library for misstated financial statement terms (such as "慘重" (heavy), "重編" (re-edit) and "和解" (reconcile)). This library is then used for future identification of fraudulent feature terms in *auditors' review reports* and *financial news* related to financial statements. In addition, indicators of the *company's stock trading volume*, *debt structure*, and *debt credit rating* are normalized, and *relevant indicators of operating revenue* in financial statements are selected. *Predictive data relating to operating revenue* in financial statements, as predicted by security companies, are then retrieved. On the basis of the results of fraudulent feature term selection, indicator normalization, financial indicator selection, and predictive data retrieval, misstated financial statements related to business groups are then detected using the *QGA-SVM classifier*.

Detection process for irrational balance sheet through ECB: As presented in Fig. 9, *letters to shareholders* for fraudulent and non-

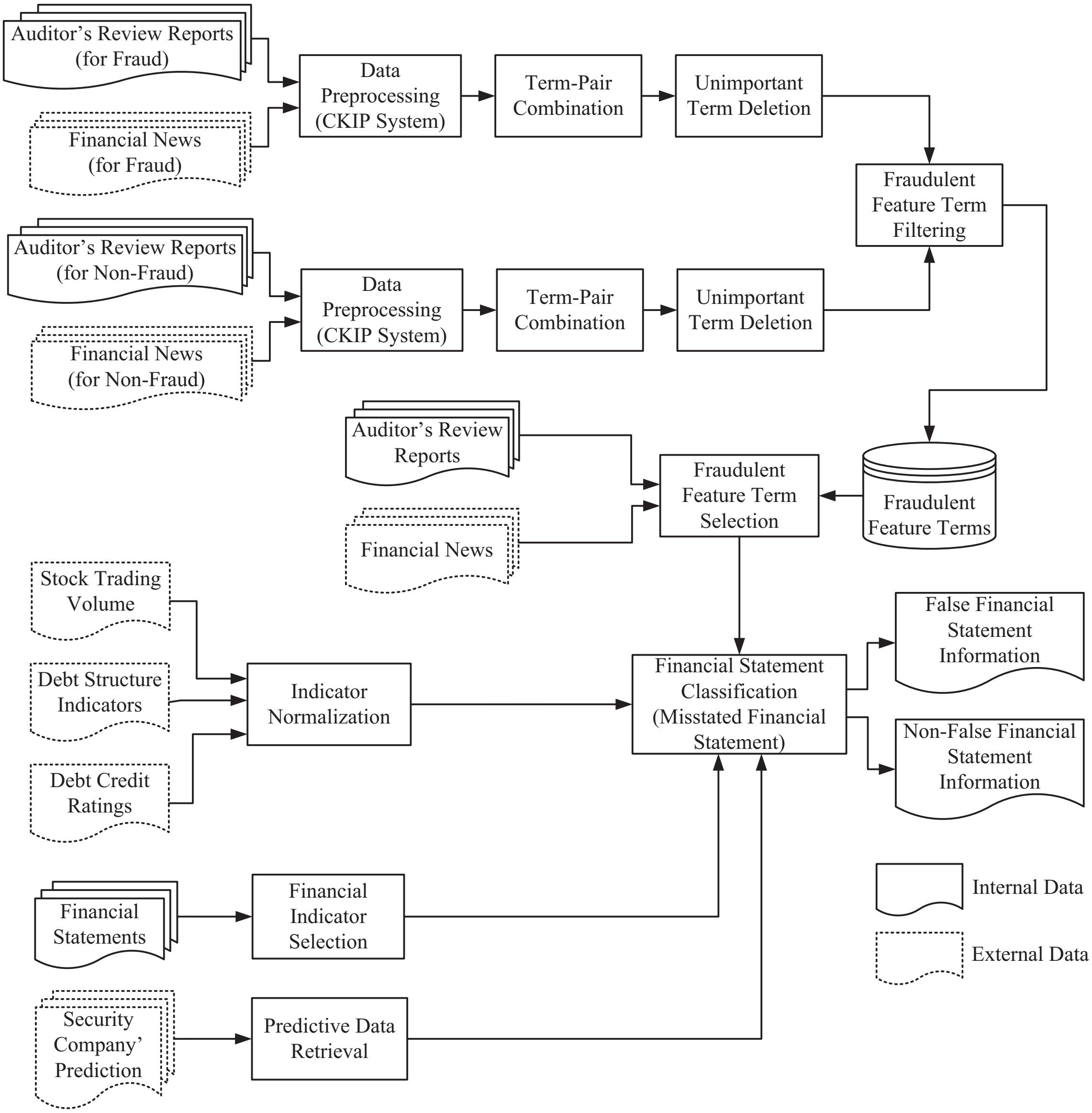
Auditor's Review Reports (for Fraud)
Financial News (for Fraud)
Data Preprocessing (CKIP System)
Term-Pair Combination
Unimportant Term Deletion
Fraudulent Feature Term Filtering
Auditor's Review Reports (for Non-Fraud)
Financial News (for Non-Fraud)
Data Preprocessing (CKIP System)
Term-Pair Combination
Unimportant Term Deletion
Auditor's Review Reports
Financial News
Fraudulent Feature Term Selection
Fraudulent Feature Terms
Stock Trading Volume
Debt Structure Indicators
Debt Credit Ratings
Indicator Normalization
Financial Statement Classification (Misstated Financial Statement)
False Financial Statement Information
Non-False Financial Statement Information
Financial Statements
Financial Indicator Selection
Security Company' Prediction
Predictive Data Retrieval
Internal Data
External Data

Fig. 8. Detection process for misstated financial statements.

fraudulent financial statements are first preprocessed using the CKIP Client (*Chinese Knowledge and Information Processing*, n.d.). Sequentially, term-pair combination, unimportant term deletion, and fraudulent feature term filtering are conducted according to the results of the letters to shareholders preprocessing, with the aim of establishing a feature term library for irrational balance sheet through ECB (including terms such as "寒冬" (slump), "擾亂" (disturb) and "醞釀" (brew)). This library may then be used in future identification of fraudulent feature terms in *letters to shareholders*. In addition, *indicators of the company's stock trading volume*, *debt structure*, and *corporate governance* are normalized, and *indicators relating to ECB in financial statements* are selected. On the basis of fraudulent feature term selection, indicator normalization, and financial statement indicator selection, irrational balance sheets through ECBs in the financial statements of business groups are identified using the *QGA-SVM classifier*.

3. Development of fraud detection techniques for financial statements of business groups

Based on the fraud detection process designed in Section 2, this section develops techniques involved in the process, comprising data pre-processing, term-pair combination, unimportant term deletion, fraudulent feature term filtering, indicator normalization, financial indicator selection, predictive data retrieval, and financial statement classification. Each element is described in the following subsections.

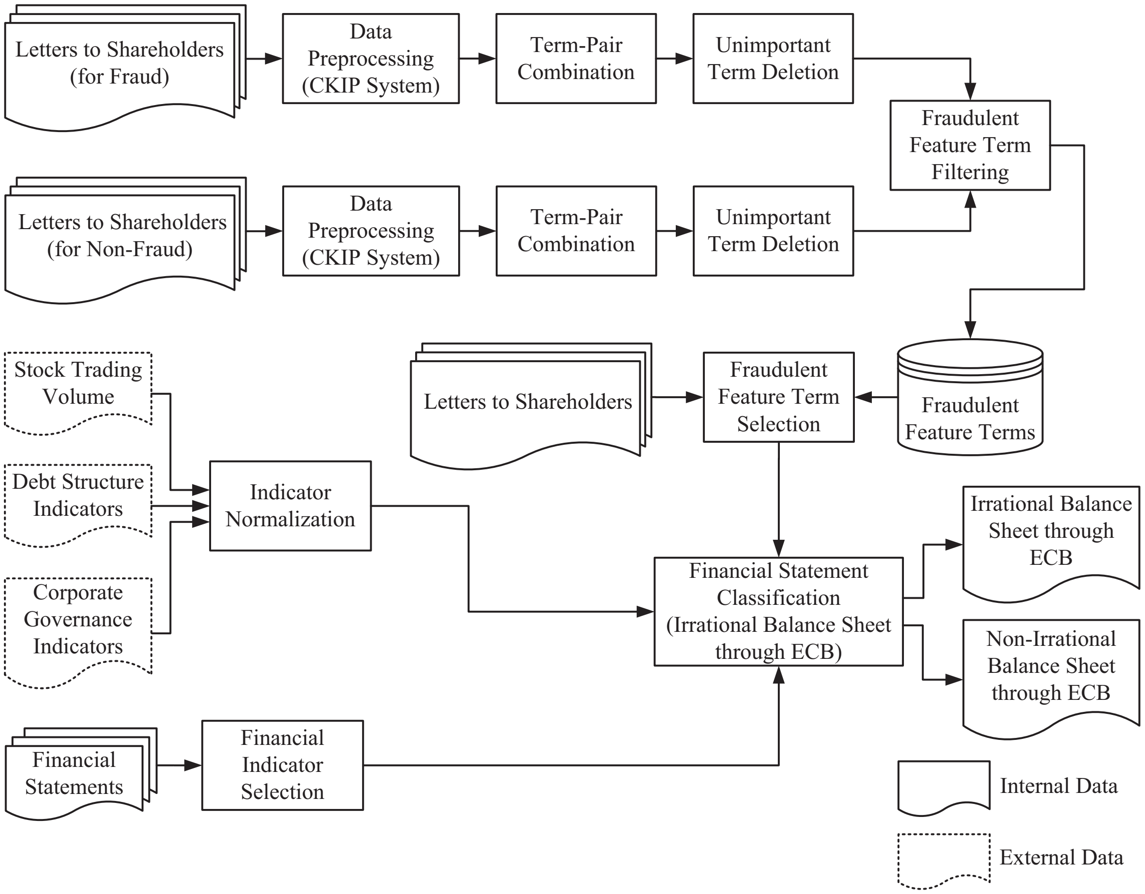
Letters to Shareholders (for Fraud)
Data Preprocessing (CKIP System)
Term-Pair Combination
Unimportant Term Deletion
Fraudulent Feature Term Filtering
Letters to Shareholders (for Non-Fraud)
Data Preprocessing (CKIP System)
Term-Pair Combination
Unimportant Term Deletion
Stock Trading Volume
Letters to Shareholders
Fraudulent Feature Term Selection
Fraudulent Feature Terms
Debt Structure Indicators
Indicator Normalization
Irrational Balance Sheet through ECB
Financial Statement Classification (Irrational Balance Sheet through ECB)
Corporate Governance Indicators
Non-Irrational Balance Sheet through ECB
Financial Statements
Financial Indicator Selection
Internal Data
External Data

Fig. 9. Detection process for irrational balance sheet through ECB.

3.1. Data pre-processing

The CKIP client is a mature and popular Chinese word segmentation system developed by the CKIP of Academia Sinica, Taiwan. In the present study, the CKIP system (*Chinese Knowledge and Information Processing*, n.d.) is utilized in the pre-processing of auditors' review reports, letters to shareholders, and financial news of fraudulent and non-fraudulent financial statements of business groups. Preprocessing involves segmenting sentences into meaningful terms, tagging the POS characteristics of terms, filtering stop-terms (e.g., particles and prepositions), and removing punctuation, respectively. This preprocessing system does not include stemming and lemmatization because it is not constructed for English-text processing. The algorithm employed for data pre-processing is presented in Fig. 10.

3.2. Term-pair combination

In the data pre-processing referred to in Section 3.1, professional terms used in the finance and accounting domain may be accidentally broken up during term segmentation of auditors' review reports, letters to shareholders, and financial news under the CKIP system. This break-up of terms can lead to the use of incorrect terms and semantics. Therefore, a term combination algorithm is designed in this study to recombine broken-up professional terms with the aim of filtering fraudulent feature terms in financial statements. The algorithm is presented in Fig. 11.

3.3. Unimportant term deletion

According to the results of data pre-processing (Section 3.1) and term combination (Section 3.2), the unimportant POS defined in Table 1 and punctuation marks listed in Table 2 are removed to retain nouns, verbs or objectives.

3.4. Fraudulent feature term filtering

Fraudulent feature terms are acquired from the auditors' review reports, letters to shareholders, and financial news of fraudulent

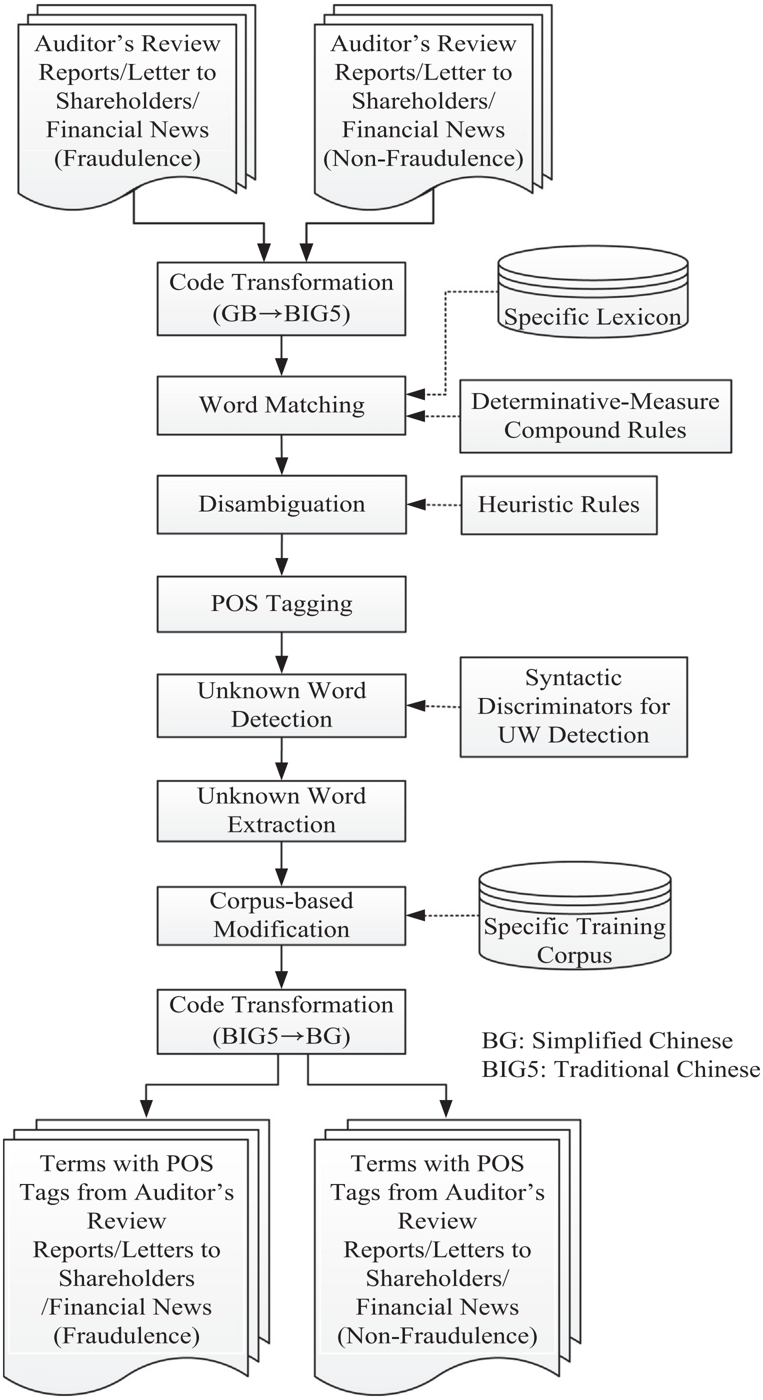
Auditor's Review Reports/Letter to Shareholders/ Financial News (Fraudulence)
Auditor's Review Reports/Letter to Shareholders/ Financial News (Non-Fraudulence)
Code Transformation (GB→BIG5)
Specific Lexicon
Word Matching
Determinative-Measure Compound Rules
Disambiguation
Heuristic Rules
POS Tagging
Unknown Word Detection
Syntactic Discriminators for UW Detection
Unknown Word Extraction
Corpus-based Modification
Specific Training Corpus
Code Transformation (BIG5→BG)
BG: Simplified Chinese
BIG5: Traditional Chinese
Terms with POS Tags from Auditor's Review Reports/Letters to Shareholders /Financial News (Fraudulence)
Terms with POS Tags from Auditor's Review Reports/Letters to Shareholders/ Financial News (Non-Fraudulence)

Fig. 10. Algorithm for data preprocessing.

and non-fraudulent financial statements using term frequency-inverse document frequency (TF-IDF) (Zhang et al., 2011). The significance level of each fraudulent or non- fraudulent term is determined. In addition, fraudulent and non-fraudulent terms are compared to identify fraudulent feature terms with high correlations in FFS and thereby establish a library of fraudulent feature terms. Fig. 12 displays the algorithm used in filtering fraudulent feature terms, and the equation representing the TF-IDF is presented as Eq. (1).

$$TFIDF_{i,j} = TF_{i,j} \times IDF_i,\ TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}},\ \mathrm{IDF_i} = \log\left(\frac{\mathrm{n}}{\mathrm{df_i}}\right) \tag{1}$$

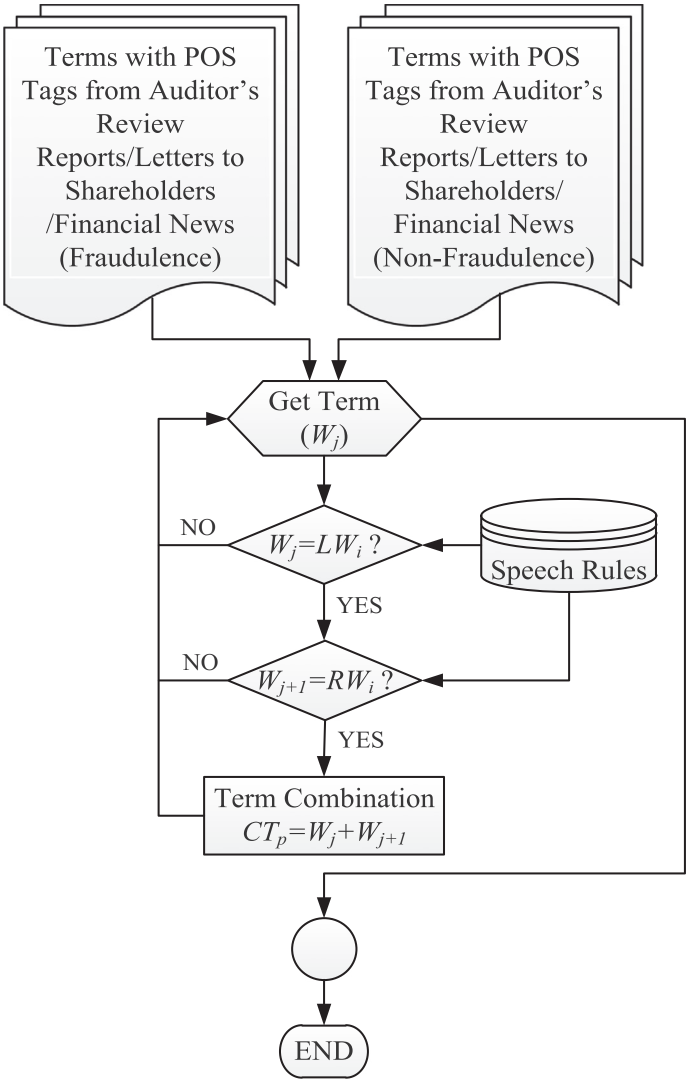
Terms with POS Tags from Auditor's Review Reports/Letters to Shareholders /Financial News (Fraudulence)
Terms with POS Tags from Auditor's Review Reports/Letters to Shareholders/ Financial News (Non-Fraudulence)
Get Term (W_j)
NO
W_j=LW_i ?
Speech Rules
YES
NO
W_{j+1}=RW_i ?
YES
Term Combination CT_p=W_j+W_{j+1}
END

Fig. 11. Algorithm for term combination.

Table 1
POS types.

POS	Type	POS	Type
Conjunction	Caa, Cab, Cba, Cbb	Pronoun	Nh
Adverb	Da, Dfa, Dfb, Di, Dk, D	Interjection	I
Infinitive	Neu, Nes, Nep, Neqa, Neqb	Preposition	P
Quantifier	NF	Expletive	T
Postposition	Ng	Stop word	V_2, DE, SHI

where

$TF_{i,\,j}$ is the frequency of a term *i* appearing in an auditor's review report/a letter to shareholders/a financial news *j* of a fraudulent/non-fraudulent financial statement;

IDF_i is the frequency of a term *i* appearing in auditor's review reports/letters to shareholders/financial news of fraudulent/non-fraudulent financial statements;

$n_{i,\,j}$ is the number of a term *i* appearing in an auditor's review report/a letter to the shareholders/a financial news *j* of a fraudulent/non- fraudulent financial statement;

$\sum_k n_{k,j}$ is the total number of all terms appearing in an auditor's review report/a letter to the shareholders/a financial news *j* of a

Table 2
Punctuation types.

Punctuation	Symbol	Punctuation	Symbol
Period	。	Exclamation point	!
Comma	,	Parenthesis	()
Pause	、	Dash	|
Semicolon	;	Ellipsis	…
Colon	:	Title no.	︴
Single quotation mark	「」	Proper name mark	|
Double quotation mark	『』	Spacer	·
Question mark	?	Hyphen	~

fraudulent/non-fraudulent financial statement;
n is the total number of auditor's review reports/letters to the shareholders/financial news of fraudulent/non- fraudulent financial statements;
df is the number of document with a term *i* appearing in auditor's review reports/letters to the shareholders/financial news of fraudulent/non-fraudulent financial statements.

3.5. Indicator normalization

Various data sources involve different numerical intervals of retrieved indicators; therefore, the number of retrieved indicators is normalized (i.e., transforming indicator numbers into the interval 0–1) using Eq. (2) (Chen et al., 2017).

$$X_{normalization} = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{2}$$

where

x_i is the *i*-th indicator;
$X_{\min}$ is the minimum value for all indicators;
$x_{\max}$ is the maximum value for all indicators.

3.6. Financial indicator selection

Friedman (1991) introduced multivariate adaptive regression splines (MARS) as a statistical method for fitting the relationship between a set of input variables and dependent variables. MARS is a nonlinear and nonparametric regression method and is based on a divide-and-conquer strategy in which the training data sets are partitioned into separate regions, each of which is fitted individually. No specific assumption about the underlying functional relationship between the input variables and output is required (Friedman, 1991; Koc and Bozdogan, 2015; Zhang and Goh, 2016). The lack of assumptions about the underlying functional relationship is rather useful in an application using diverse data. Thus, this study employs an open MARS source code from Jekabsons (2010) to select and predict critical indicator variables related to operating revenue, earnings per share, and debt ratio, with the aim of establishing of a fraud detection model. Eq. (3) is a general model of MARS, and Eq. (4) is used for selecting spline basis function (BF) that substantially contributes to MARS model optimization.

$$f(x) = \alpha_0 + \sum_{m=1}^{M} \alpha_m \prod_{k=1}^{K_m} [S_{k,m}(x_{v(k,m)} - t_{k,m})] \tag{3}$$

where

α_0 and α_m are numbers;
M is the number of BF;
K_m is the number of nodes in the segmentation interval, which can be divided into linear regression with different slopes;
$S_{k,\ m}$ represents the direction (value $+1$showing the right and value -1 showing the left);
$v_{k,\ m}$ is the predictor variable;
$t_{k,\ m}$ represent values on the corresponding variables.

$$\mathrm{GCV(M)} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} [\mathrm{y_i} - \mathrm{f_M(x_i)}]^2 / \left[1 - \frac{\mathrm{C(M)}}{\mathrm{N}}\right]^2 \tag{4}$$

where

GCV(*M*) is residual mean of calculating valid BF;
y_i is an actual value;

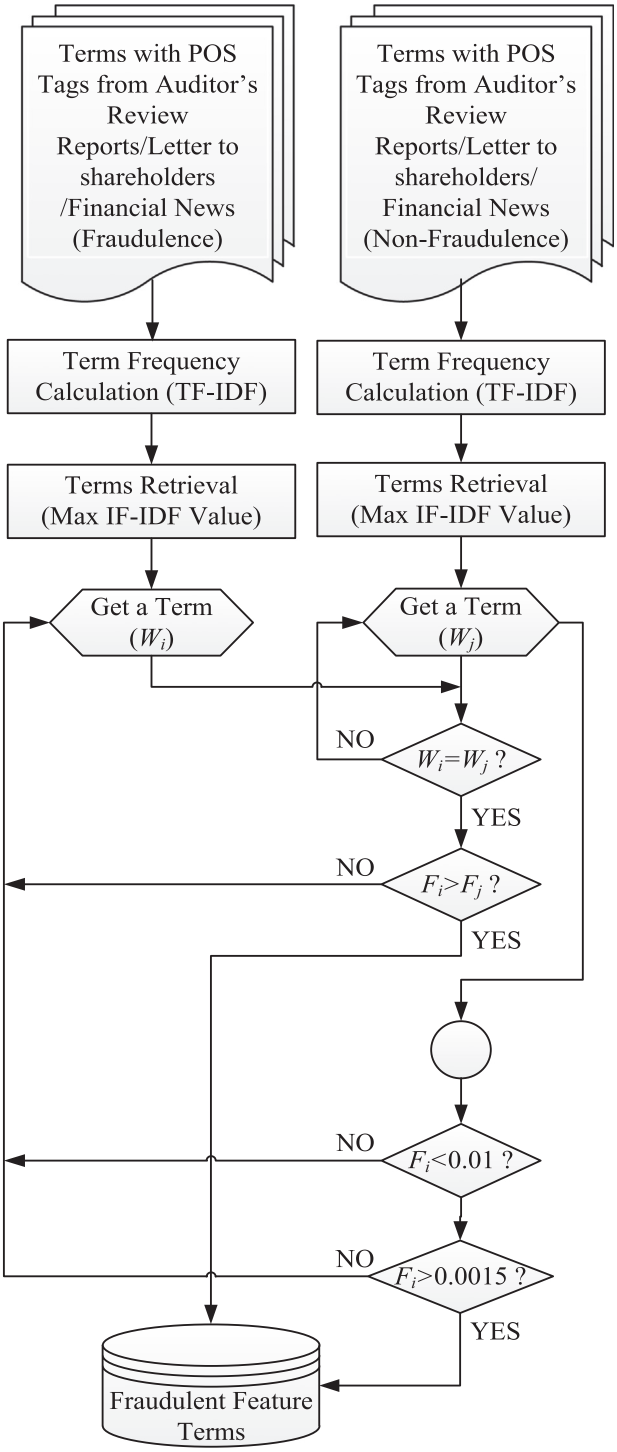
Terms with POS Tags from Auditor's Review Reports/Letter to shareholders /Financial News (Fraudulence)
Terms with POS Tags from Auditor's Review Reports/Letter to shareholders/ Financial News (Non-Fraudulence)
Term Frequency Calculation (TF-IDF)
Term Frequency Calculation (TF-IDF)
Terms Retrieval (Max IF-IDF Value)
Terms Retrieval (Max IF-IDF Value)
Get a Term (W_i)
Get a Term (W_j)
NO
W_i=W_j ?
YES
NO
F_i>F_j ?
YES
NO
F_i<0.01 ?
NO
F_i>0.0015 ?
YES
Fraudulent Feature Terms

Fig. 12. Algorithm for filtering fraudulent feature terms.

$f_M(x_i)$ is the predicted value;
C(M) is the product of valid BF parameters.

3.7. Predictive data retrieval

To acquire brokers' predictive data related to a company from the Taiwan Economic Journal (TEJ) database (*Taiwan Economic*

Journal, n.d.), the TEJ Smart Wizard is used, and settings for data class, name of database, corporate issue class, and industrial classification are used to retrieve gross profit, net profit, and net earnings per share predictive data. Fig. 13 presents the setting process used in the TEJ Smart Wizard.

3.8. Financial statement classification

An SVM is considered the most accurate classification model (Güraksın et al., 2014; Peng and Xu, 2013), and a QGA is widely used for adjusting and optimizing the parameters of classification models (Stern et al., 2006; Tsang et al., 2004). Thus, this study integrates an SVM and QGA to develop a classification algorithm for financial statement fraud detection (in FFS and non-FFS). The accuracy of the model is enhanced through selection and optimization of the SVM parameters using a QGA (Chen et al., 2017), as represented in Fig. 14. The relevant calculations are displayed in Eqs. (5)–(8).

$$D_{m+1} = M(q_i \times d_i) \tag{5}$$

$$F(d_i) = rank(D_{m+1}) \tag{6}$$

where

$F(d_i)$ denotes the fitness value;
D_m denotes the primal objective function;
q_i denotes the randomly selected fitness function in the optimal function sequence;
d_i denotes the randomly selected fitness function in all function sequences.

$$f(x) = sign\left(\sum_{i=1}^{n} \alpha_i y_i K(x, x_i) + b\right) \tag{7}$$

$$K(x, x_i) = \exp\left(-\frac{\|x - x_i\|^2}{2\sigma^2}\right) \tag{8}$$

where

$f(x)$ is the optimal decision function;
y is the class index of various indicators;
α is the Lagrange multiplier;
b is the offset value;
K is the RBF;
σ is the parameter of RBF.

An SVM is acquired after numerous iterations. Weight voting for the SVM is performed based on the weight used to generate the QGA-SVM model. Eq. (9) is the formula used in weight voting.

$$H(x) = arg\ max \sum_t \left(\ln \frac{1}{\beta_t}\right) h_t(x, y) \tag{9}$$

where

$H(x)$ denotes the class index of QGA-SVM;
$h_t(x, y)$ denotes the class index of SVM;
β_t denotes the weight of SVM.

Finally, the test dataset is input into the QGA-SVM model to determine the results of financial statement classification.

4. Demonstration and evaluation of the proposed approach

This section describes the use of Python 2.7 and Matlab R2014a to implement the techniques outlined in Section 3 for detecting fraud in the financial statements of business groups. Additionally, the feasibility and validity of the proposed approach are demonstrated using the financial statements of business groups in Taiwan. Furthermore, the detection accuracy is evaluated through comparison with other detection models to prove the effectiveness of the proposed approach.

Based on the claim-type of “false financial statements” listed within the Securities and Futures Investors Protection Center (referring to Securities and Exchange Act Article 20), 58 business groups that committed financial statement fraud between 2000 and 2014 are selected. For the same time range, 174 business groups that did not commit financial statement fraud are selected from the same industry as that of the fraudulent business groups. The selected business groups that did not commit fraud held similar total assets to those belonging to the fraudulent business group during the specified time range. Relevant data of both types of business

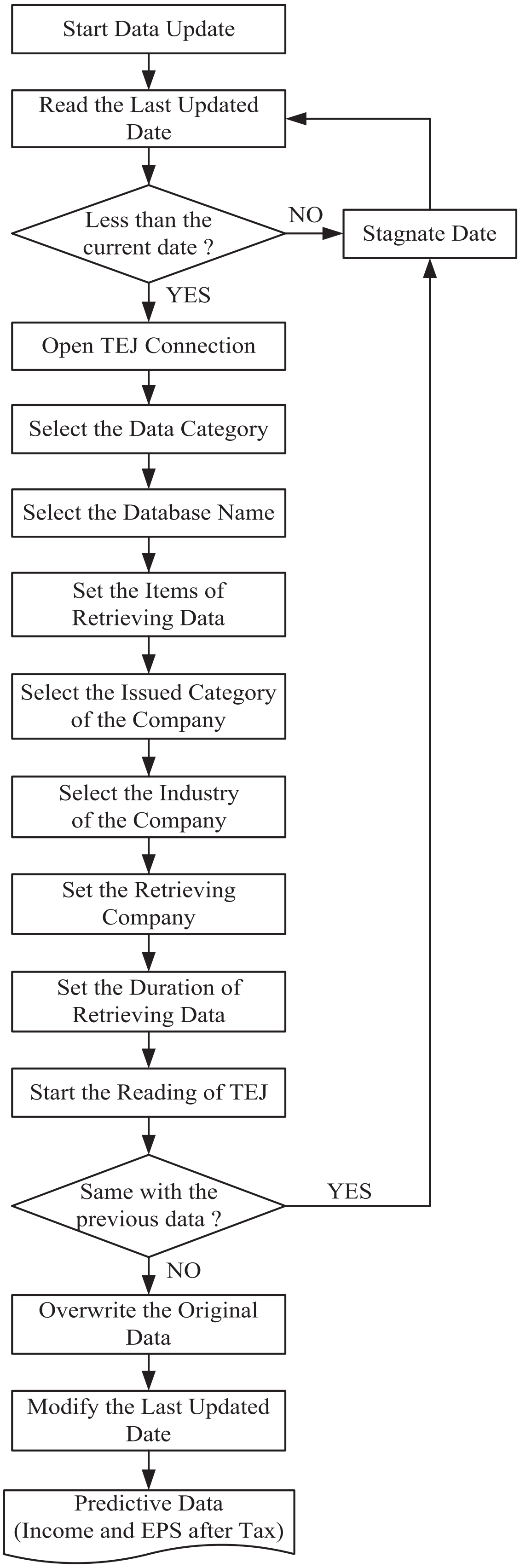
Start Data Update
Read the Last Updated Date
Less than the current date ?
NO
Stagnate Date
YES
Open TEJ Connection
Select the Data Category
Select the Database Name
Set the Items of Retrieving Data
Select the Issued Category of the Company
Select the Industry of the Company
Set the Retrieving Company
Set the Duration of Retrieving Data
Start the Reading of TEJ
Same with the previous data ?
YES
NO
Overwrite the Original Data
Modify the Last Updated Date
Predictive Data (Income and EPS after Tax)

Fig. 13. Setting procedure of TEJ Smart Wizard for retrieving predictive data.

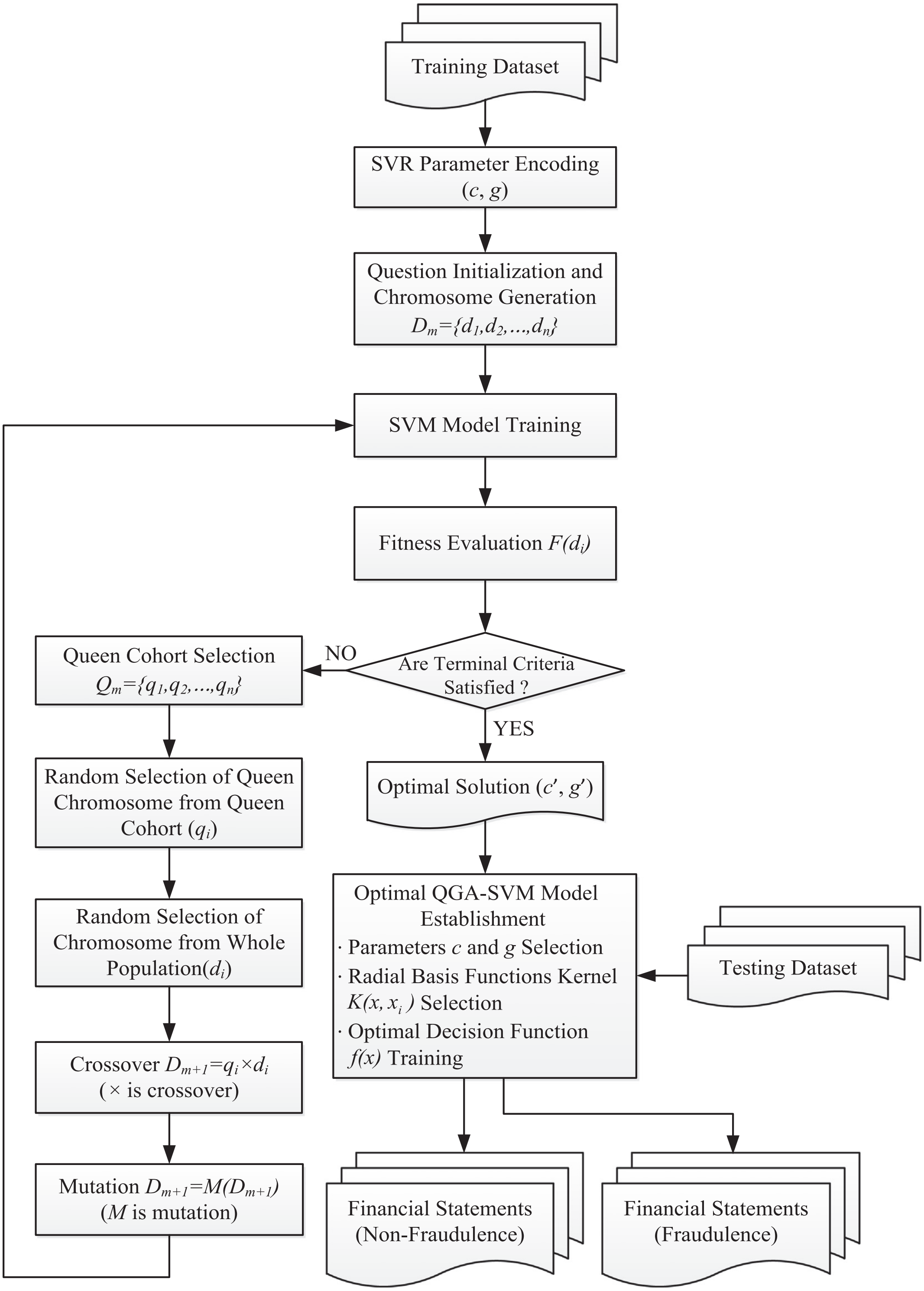
Training Dataset
SVR Parameter Encoding (c, g)
Question Initialization and Chromosome Generation $D_m=\{d_1,d_2,\ldots,d_n\}$
SVM Model Training
Fitness Evaluation $F(d_i)$
Are Terminal Criteria Satisfied ?
NO
YES
Queen Cohort Selection $Q_m=\{q_1,q_2,\ldots,q_n\}$
Random Selection of Queen Chromosome from Queen Cohort (q_i)
Random Selection of Chromosome from Whole Population(d_i)
Crossover $D_{m+1}=q_i\times d_i$ (× is crossover)
Mutation $D_{m+1}=M(D_{m+1})$ (M is mutation)
Optimal Solution (c', g')
Optimal QGA-SVM Model Establishment
· Parameters c and g Selection
· Radial Basis Functions Kernel $K(x, x_i)$ Selection
· Optimal Decision Function $f(x)$ Training
Testing Dataset
Financial Statements (Non-Fraudulence)
Financial Statements (Fraudulence)

Fig. 14. Algorithm for classifying financial statements.

Table 3
Partial letter to shareholders of INFODISC Technology Co., Ltd.

各位股東女士、先生們:
茲將本公司九十六年度整體營運狀況,報告如下:
一、96 年度營業報告書
(一) 營運成果
九十六年度公司營收淨額約為新台幣6.32 億元,較九十五年度減少約3 億元,衰退幅度約為33%。
二、研究發展狀況
(一)研發成果
本公司已成功研發下列預錄產品:
1. C-thru CD&DVD:
幾近透明,可直接看穿,以達到不同的視覺效果之光碟片。
2. DVD-black & white:
全白與全黑碟片,碟片表面可呈現特殊鏡面效果,以表現出十分高級的質感,並可在印刷時,做到同業所無法達到的滿版印刷,此項技術已獲得專利。
3. Colorful disc:
目前計有粉紅、螢光紅、橘、螢光黃、紫、深綠等六種顏色,於印刷時可產生特別的效果。
(Dear shareholders:
The overall operating conditions of the company in 2007 are reported as follows.
I. Business report in 2007
(1) Operating achievement
The net revenue in 2007 is about 0.632 billion NT dollars, which is 0.3 billion NT dollars less than it in 2006, with the decline rate about 33%.
II. Research and development conditions
(1) Research and development achievement
The company has successfully researched and developed the following pre-recorded products:
1. C-thru CD&DVD:
Such a CD is nearly transparent and could be directly seen through to achieve different visual effect.
2. DVD-black & white:
The surface of such black and white CDs present special mirror effect to perform extremely high quality and could be fully printed, which cannot be reached by other businesses in the same trade. The technology is patented.
3. Colorful disc:
There are six colors of pink, fluorescent red, orange, fluorescent yellow, purple, and dark green to appear special effect on printing.)

groups are then retrieved to demonstrate and evaluate the effectiveness of the approach proposed in this study.

In this section, the fraud type of business group financial statements "Irrational Balance Sheet through ECB" is given as an illustrative example to explain how the proposed approach can help readers of financial statements to identify the fraud occurred to the business group. Based on letters to shareholders, stock trading volume, debt structure indicators, corporate governance indicators, and the financial ratio of business groups in Taiwan retrieved from the TEJ database (*Taiwan Economic Journal*, n.d.) and the Taiwan Stock Exchange Corporation (*Taiwan Stock Exchange*, n.d.), the irrational balance sheet through ECB is detected as follows.

(1) Establish a library of fraudulent feature terms

According to the study proposed by Glancy and Yadav (2011), 17 letters to shareholders for irrational balance sheet through ECB and 51 letters to shareholders for non-irrational balance sheet through ECB are randomly collected and used to establish a library of fraudulent feature terms. Table 3 presents a partial letter to shareholders for the business group INFODISC Technology Co., Ltd, wherein an irrational balance sheet through ECB was identified.

Step 1. Preprocess data

Using the CKIP system, the letter to shareholders from INFODISC Technology Co., Ltd., is broken into sentences and words, as represented in Table 3. The POS of each word is then tagged. Table 4 illustrates partial results after pre-processing of the letter to shareholders.

Step 2. Combine term-pairs

Based on the results of data pre-processing listed in Table 4, some terms are combined as term-pairs, as depicted in Table 5.

Step 3. Delete unimportant terms

After pre-processing the letter to shareholders and combining terms into term-pairs, unimportant POS and punctuation marks are removed from the letter, as indicated in Table 6.

Step 4. Filter fraudulent feature terms

Terms in the letters to shareholders from 17 business groups with irrational balance sheet through ECB and 51 business groups without irrational balance sheet through ECB are determined by using Eq. (1) to identify terms with a high TF-IDF value. Partial

Table 4
Partial results of sentence breaking and POS tagging for the letter to shareholders.

各(every)(DET) 位(body)(M) 股東(shareholder)(N) 女士(Ms.)(N) 、(PAUSECATEGORY) 先生們(Mr.)(N): (COLONCATEGORY) 茲(hereby)(N) 將(will)(P) 本(the)(DET) 公司(company)(N) 九十六年度(Ninety-six years)(N) 整體(overall)(N) 營運(operating)(Nv) 狀況(status)(N),(COMMACATEGORY) 報告(report)(N) 如下(as follows)(Vt): (COLONCATEGORY)

一(first)(DET) 、(PAUSECATEGORY) 96 (ninety-six)(DET) 年度(year)(N) 營業(business)(Nv) 報告書(report)(N) ((PARENTHESISCATEGORY) 一(DET))(PARENTHESISCATEGORY) 營運(operational)(Nv) 成果(results)(N) 九十六年度(ninety-six years)(N) 公司(company)(N) 營收(revenue)(N) 淨額(net)(N) 約(approximately)(ADV) 為(is)(Vt) 新台幣(NT)(N) 6.32(DET) 億(billion)(DET) 元(dollar)(M),(COMMACATEGORY) 較(compare with)(P) 九十五年度(Ninety-five years) (N) 減少(reduce)(Vt) 約(about)(ADV) 3(DET) 億(billion)(DET) 元(dollar)(M),(COMMACATEGORY) 衰退(recession)(Vi) 幅度(range)(N) 約(approximately)(ADV) 為(is)(Vt) 33(DET) %(M) 。(PERIODCATEGORY)

二(DET) 、(PAUSECATEGORY) 研究(research)(Nv) 發展(development)(Nv) 狀況(status)(N)

((PARENTHESISCATEGORY) 一(DET))(PARENTHESISCATEGORY) 研發(R&D)(Nv) 成果(results)(N)

本(this)(DET) 公司(company)(N) 已(has)(ADV) 成功(successfully)(Vi) 研發(developed)(Vt) 下列(the following)(A) 預錄(pre-recorded)(Vt) 產品(products)(N): (COLONCATEGORY)

1(DET). (PERIODCATEGORY) C-thru(FW) CD&DVD(FW): (COLONCATEGORY) 幾近(almost)(Vt) 透明(transparent)(Vi),(COMMACATEGORY) 可(can be)(ADV) 直接(directly)(Vi) 看穿(seen through)(Vt),(COMMACATEGORY) 以(in order to)(C) 達到(achieve)(Vt) 不同(different)(Vi) 的(of)(T) 視覺(visual)(N) 效果(effects)(N) 之(of)(T) 光碟片(disc)(N) 。(PERIODCATEGORY)

2(DET).(PERIODCATEGORY) DVD-Black(FW) &(FW) White(FW): (COLONCATEGORY) 全(all)(DET) 白(white)(Vi) 與(and)(C) 全(all)(DET) 黑(black)(Vi) 碟片(disc)(N),(COMMACATEGORY) 碟片(disc)(N) 表面(surface)(N) 可(can)(ADV) 呈現(show)(Vt) 特殊(special)(Vi) 鏡面(mirror)(N) 效果(effect)(N),(COMMACATEGORY) 以(to)(P) 表現出(show)(Vt) 十分(very)(ADV) 高級(high)(Vi) 的(of)(T) 質感(quality texture)(N),(COMMACATEGORY) 並(and)(C) 可(can be)(ADV) 在(at)(P) 印刷(printing)(Nv) 時(time)(POST),(COMMACATEGORY) 做到(do)(Vt) 同業(same industry)(N) 所(that)(ADV) 無法(can not)(ADV) 達到(achieve)(Vt) 的(of)(T) 滿(full)(DET) 版(version)(N) 印刷(printing)(Nv),(COMMACATEGORY) 此(the)(DET) 項(item)(M) 技術(technology)(N) 已(has)(ADV) 獲得(obtained)(Vt) 專利(patent)(N) 。(PERIODCATEGORY)

3(DET). (PERIODCATEGORY) Colorful(FW) Disc(FW): (COLONCATEGORY) 目前(at present)(N) 計有(there are)(Vt) 粉紅(pink)(N) 、(PAUSECATEGORY) 螢光(fluorescent)(Vi) 紅(red)(Vi) 、(PAUSECATEGORY) 橘(orange) (N) 、(PAUSECATEGORY) 螢光(fluorescent) (Vi) 黃(yellow)(Vi) 、(PAUSECATEGORY) 紫(purple)(Vi) 、(PAUSECATEGORY) 深綠(dark green)(Vi) 等(etc.)(POST) 六(six)(DET) 種(type)(M) 顏色(colors)(N),(COMMACATEGORY) 於(in the)(P) 印刷(printing)(Nv) 時(time)(POST) 可(can)(ADV) 產生(produce)(Vt) 特別(special)(Vi) 的(of)(T) 效果(effects)(N) 。(PERIODCATEGORY)

Table 5
Results of term-pair combination for partial letter to shareholders.

Left term (LT)	Right term (RT)	Term-pair combination
營運(operational)(Nv)	狀況(status)(N)	營運狀況(operational status)(N)
營運(operational)(Nv)	成果(results)(N)	營運成果(operational results)(N)
研發(R&D)(Nv)	成果(results)(N)	研發成果(R&D results)(N)
營運(operational)(Nv)	計劃(plan)(N)	營運計劃(operational plan)(N)
經營(business)(Nv)	目標(objective)(N)	經營目標(business objective)(N)
管理(management)(Nv)	系統(system)(N)	管理系統(management system)(N)

Table 6
Results of unimportant term deletion for partial letter to shareholders.

股東女士先生們(Shareholders, ladies and gentlemen)(NNa) 茲(hereby)(Nd) 公司(company)(Nc) 年度(year)(Nd) 整體(overall)(Na) 營運狀況(operational status)(NNa) 報告(report)(Na) 如下(as follows)(VK) 年度(year)(Nd) 營運成果(operational results) (NNa) 年度(year)(Nd) 公司(company)(Nc) 新台幣(NT)(Na) 年度減少(reduce)(NdV) 約(approximately) (Da) 衰退幅度(recession range)(NNa) 研發成果(R&D results)(NNa) 公司(company)(Nc) 已成功(successfully)(DV) 研發(developed)(VC) 下列預錄產品(the following pre-recorded products)(NNa) 幾近(almost)(VJ) 透明(transparent)(VH) 可直接(directly)(DV) 看穿(seen through)(VJ) 達到(achieve)(VJ) 不同視覺效果光碟片(different visual effects of the disc)(NNa) 白(white)(VH) 黑碟片(black disc)(NNa) 碟片表面(disc surface)(NNa) 可呈現(can show)(DV) 特殊鏡面效果(special mirror effect)(NNa) 表現出(show)(VC) 分高級(very high quality)(DV) 質感(texture)(Na) 印刷(printing)(VC) 做到同業(same industry)(NNa) 無法達到(can not achieve)(DV) 版(version)(Na) 印刷(printing)(VC) 技術(technology)(Na) 已獲得(has obtained)(DV) 專利(patent)(Na) 目前(at present)(Nd) 計有(there are)(VJ) 粉紅螢光(pink fluorescent)(VHV) 紅橘(red orange)(NNa) 螢光黃(fluorescent yellow)(VHV) 紫深綠(purple dark green)(VHV) 顏色(color)(Na) 印刷(printing)(VC) 可產生(can produce)(DV) 特別效果(special effects)(NNa)

Table 7
TF-IDF values for partial terms from letters to shareholders (business groups for irrational balance sheet through ECB).

Item no.	Term	TF-TDF value	Item no.	Term	TF-TDF value
1	寒冬(slump)	0.005898481	11	擴展(extend)	0.002949240
2	衝擊(impact)	0.002488699	12	籠罩(envelop)	0.003698254
3	部門(department)	0.002932915	13	積壓(backlog)	0.005865829
4	疲弱(weak)	0.003737808	14	重整(restructure)	0.002527920
5	攪亂(disturb)	0.002514281	15	預算(budget)	0.003562004
6	客戶(customer)	0.003666922	16	甚深(very deep)	0.002514281
7	引發(trigger)	0.005028561	17	影響(influence)	0.003566700
8	訂單(order)	0.003527042	18	夾攻(attack)	0.005776612
9	支出(expenditure)	0.005898481	19	受限(limit)	0.003926798
10	醞釀(brew)	0.003737808	20	平息(calm down)	0.005216194

Table 8
TF-IDF values for partial terms from letters to shareholders (business groups for non-irrational balance sheet through ECB).

Item no.	Term	TF-TDF value	Item no.	Term	TF-TDF value
1	部門(department)	0.005197708	11	發揮(elaborate)	0.003444693
2	計畫(plan)	0.002475873	12	建立(construct)	0.003460348
3	受限(limit)	0.005498713	13	重整(restructure)	0.008417280
4	設計(design)	0.002479365	14	擴展(extend)	0.004385014
5	衝擊(impact)	0.008956202	15	衰退(decline)	0.003503609
6	影響(influence)	0.002462752	16	客戶(customer)	0.008272766
7	疲弱(weak)	0.005723750	17	支出(expenditure)	0.007724718
8	未來(future)	0.002499205	18	預算(budget)	0.001953896
9	籠罩(envelop)	0.005246771	19	變化(change)	0.003509546
10	模擬(simulation)	0.002559400	20	訂單(order)	0.003974506

results are displayed in Tables 7 and 8.

The terms in Table 7 are compared with those in Table 8. To acquire more representative fraudulent feature terms from Table 7, a simulated experiment is conducted, and the results indicate that (a) if terms in Table 7 appear in Table 8 but exhibit smaller TF-IDF values, the terms are removed from Table 7; and (b) if the TF-IDF values of the fraudulent terms for irrational balance sheet through ECB in Table 7 are larger than 0.01 or smaller than 0.0015, the fraudulent terms are also removed. All the remaining fraudulent feature terms are presented in Table 9.

(2) Normalize indicators

Table 9
Fraudulent feature terms for letters to shareholders (business groups for irrational balance sheet through ECB).

寒冬(slump)	期望(expect)	積壓(backlog)	蓬勃(flourishing)	深入(in depth)
攪亂(disturb)	景氣(boom)	甚深(very deep)	健全(sound)	立足(foothold)
引發(trigger)	支援(support)	夾攻(attack)	不同(different)	重心(focus)
醞釀(brew)	雜訊(noise)	平息(calm down)	要求(request)	改進(improve)
紛擾(trouble)	遞延(deferred)	分散(dispersion)	擴張(extend)	轉廠 (interplant transfer)
代工(OEM)	移動(move)	剝削(exploit)	實力(strength)	挹注(injection)
特殊(special)	佳境 (pleasant stages)	破壞(damag)	強烈(strongly)	突顯(highlight)
強化(consolidate)	回穩(stabilized)	最佳化(optimization)	追求(pursue)	疲弱(weak)
改善(improve)	尚達(not reach)	看穿(see through)	拓展(expand)	不利(unfavorable)
超頻(overclocking)	決議(resolution)	刺激(stimulate)	加強(strengthen)	議題(issue)
造成(cause)	保守(conservative)	改變(change)	原先(original)	避免(avoid)
推出(launch)	差距(gap)	可達(reach)	契機(opportunity)	優越(superior)
普及(universal)	龐大(huge)	投入(committed)	波動(fluctuation)	因應(in response)
提列(provision)	預估(forecast)	配合(cooperation)	情勢(situation)	事件(event)
調整(adjust)	競爭(competition)	降低(reduce)	著重(focus on)	搶救(rescue)
積極(active)	節省(economize)	爭取(strive for)	唯一(only)	提前(early)
整合(integrate)	困境(dilemma)	不足(insufficient)	面臨(face)	差異化(differentiation)
扭力(turning)	尋找(looking for)	低廉(low)	尚未(not yet)	抱持(hold)
不順 (not smooth)	謹慎(cautious)	激發(excitation)	擴展(extended)	不良(bad)
衰退(recession)	削價(underprice)	回溫(rebound)	穩固(firm)	營造(create)
穩定(stable)	上漲(rise)	逐步(step by step)	適時(timely)	拉低(pull down)
致力(dedicate)	盜版 (priate edition)	確保(ensure)	無擔(no burden)	加重(aggravate)
爆裂(burst)	協力 (unite efforts)	完備(complete)	遺憾(regret)	嚴重(serious)
顛覆(subvert)	飽和(saturation)	減少(reduce)	邁進 (stride forward)	解決(solve)
異業(cross-industry)	移往(shift)	不遺餘力 (spare no efforts)	預計(estimate)	全面(overall)
嚴格(strict)	急劇(sharp)	重新(anew)	困難(difficult)	私下(in private)
落差(deviation)	微利 (micro profit)	大致(roughly)	移轉(transfer)	必需(required)
優等(excellent)	居高不下 (remain high)	已深(deep)	共謀(conspiracy)	轉向(turn to)
重大(major)	下滑(decline)	控制(control)	拉開(enlarge)	不如 (not as good as)
合理(reasonable)	樂觀(optimistic)	檢討(review)	轉換(change)	衝擊(impact)
辛勞(toil)	引發(trigger)	爆發(break out)	償還(repay)	

Table 10
Partial results of stock trading volume, debt structure indicator, and debt credit rating normalization.

Business group name	Stock trading volume (Lot)	Normalized stock trading volume	Debt structure indicator	Normalized debt structure indicator	Debt credit rating	Normalized debt credit rating
ABIT	2109883	0.553028	0.578307	0.465626	6	0.555556
PROCOMP	1936825	0.507667	0.550137	0.429927	8	0.777778
XEPEX	1200233	0.314597	0.450513	0.303679	7	0.666667
KOLIN	2504295	0.656408	0.713819	0.637353	8	0.777778
BULL WILL Co.	2070419	0.542683	0.851222	0.811476	6	0.555556

The indicators for stock trading volume, debt structure, debt credit ratings, and corporate governance of business groups are normalized using Eq. (2). Tables 10 and 11 present the results of indicator normalization for five business groups.

(3) Select financial indicators

Critical indicators mostly relating to the debt ratio are selected from the financial statements of five business groups using Eqs. (3) and (4), as depicted in Table 12.

(4) Classify financial statements for Irrational Balance Sheet through ECB

The training dataset and testing dataset for classifying financial statements with irrational balance sheet through ECB are established according to the aforementioned experimental sample. The training dataset is sampled from data for 2000–2006 and comprises 20 financial statements with irrational balance sheet through ECB and 60 financial statements without irrational balance sheet through ECB. An additional testing dataset is sampled from data for 2007–2014 and comprises 38 financial statements with irrational balance sheet through ECB and 114 financial statements without irrational balance sheet through ECB.

The established datasets are input into the QGA-SVM classification model (Fig. 14). In training and testing this model, the relevant parameter settings are optimized, as listed in Table 13. Table 14 summarizes the results of classifying fraudulent and non-fraudulent financial statements in terms of irrational balance sheet through ECB. In clustering evaluations in the sign test, the *p*-values for 36 correct clusters of the 38 FFS and 106 correct clusters of the 114 non-FFS are 0.0069 and 0.0053, respectively. Testing of the QGA-SVM classification model demonstrates that it discriminates at the 0.01 level of significance for both fraudulent and non-fraudulent financial statements. The 99% confidence intervals for the mean of a sample randomly selected from fraudulent and non-fraudulent financial statements are (4, 34) and (10, 104).

The classification accuracy is compared with that of the following well-known classifiers: the decision tree C4.5 (J48 version with the minimum number of instances per leaf (default 2) and confidence factor for pruning (default = 0.25)), logistic regression (Broyden Fletcher Goldfarb Shanno learning algorithm), back-propagation neural network (BPN), k-nearest neighbors (KNN; value of k in KNN algorithm is usually determined by the rule $n\hat{}(1/2)$, where n is the number of features), genetic algorithms-support vector machine (GA-SVM), and particle swarm optimization-support vector machine (PSO-SVM). The results of the comparisons are represented in Table 15. The QGA-SVM model exhibits higher accuracy than the other six models.

The detection of exaggerated profit, undue deposit and debt, and misstated financial statements are also conducted, and results of classification accuracy comparisons against the six other models are listed in Tables 16–18, respectively. In evaluating the classification accuracy of these models, Type I and Type II error rates are considered. A Type I error occurs when a fraudulent financial statement is classified as non-fraudulent, whereas a Type II error occurs when a non-fraudulent financial statement is classified as fraudulent. Because of the difference in costs associated with choosing the right action (such as sample size) for Type I and Type II classification errors, these models have different costs of misclassification. Classifying a fraud company as non-fraud may lead to incorrect decisions, which may cause serious economic damage. The misclassification of a non-fraud company may cause additional investigations at the expense of the required time. In the experiments, the Type I and II error rates are lower for the proposed method.

5. Conclusions

This study considers diverse data used in finance and economics to develop an approach with which to precisely detect the financial statement fraud of business groups and thus reduce investment losses and risks and benefits investors and creditors. The main results and contributions of this study are summarized as follows.

(1) Fraud detection model for financial statements of business groups: Based on the concept of data diversity, a fraud detection model for the financial statements of business groups was designed. This model not only considers internal data such as auditors' review reports, financial ratios, and letters to shareholders but also external data: for example, financial news, stock trading volumes, security companies' predictive data, debt credit ratings, debt structure, and corporate governance. The proposed model may serve as a valuable reference model and can be applied to detect similar cases of fraud in other languages or may be applied in detection of other types of fraud, such as fraudulent collusion between employees and suppliers and sales fraud.

(2) Fraud detection method for financial statements of business groups: Based on the designed fraud detection model, a method for

Table 11
Partial results of corporate governance indicator normalization.

Business group name	Share collateralization by directors and supervisors (%)	Normalized share collateralization by directors and supervisors (%)	Share collateralization by directors (%)	Normalized share collateralization by directors (%)	Blockholder shareholding ratio (%)	Normalized blockholder shareholding ratio (%)
ABIT	35.24	0.3524	37.25	0.3725	3.74	0.086754813
PROCOMP	36.10	0.3610	41.08	0.4108	8.29	0.192298771
XEPEX	65.11	0.6511	65.11	0.6511	5.40	0.125260960
KOLIN	11.70	0.1170	12.28	0.1228	30.66	0.711203897
BULL WILL Co.	78.11	0.7811	76.29	0.7629	15.62	0.362328926

Table 12
Partial results of financial indicator selection.

Business group name	Cash reinvestment ratio	Income before tax/paid-in capital
ABIT	−0.81	−20.54
PROCOMP	0.73	−86.52
XEPEX	−23.95	−77.34
KOLIN	−4.28	−0.40
BULL WILL Co.	−12.87	24.16

Table 13
Parameter settings for the QGA-SVM model.

Parameter name	Value set
QGA population	20
QGA evolution	200
QGA threshold	0.9
c and *g* of SVM	Based on the results of QGA

Table 14
QGA-SVM testing results and detection at 0.01 significance level.

Testing sample	Total	Correctly identified	Incorrectly identified	*p*-Value	Detected at 0.01 level	
					Upper	Lower
Fraudulent financial statements	38	36	2	0.0069	34	4
Non-fraudulent financial statements	114	106	8	0.0053	104	10

Table 15
Classification accuracy comparison for irrational balance sheet through ECB.

Classifier	*c*	*g*	Fraud (%)	Non-fraud (%)	Total accuracy (%)
Decision tree C4.5	–	–	77.1	75.6	76.2
Logistic regression	–	–	89.8	87.1	88.4
BPN	–	–	85.8	81.4	83.5
KNN	–	–	85.3	81.7	83.4
GA-SVM	0.7272	1.5181	82.6	80.3	81.4
PSO-SVM	11.7667	2.5754	82.4	81.8	82.1
QGA-SVM	4.0705	1.7748	94.7	92.9	93.8

Table 16
Classification accuracy comparison for exaggerated profit.

Classifier	*c*	*g*	Fraud (%)	Non-fraud (%)	Total accuracy (%)
Decision tree C4.5	–	–	87.2	88.1	87.6
Logistic regression	–	–	72.4	70.0	71.0
BPN	–	–	79.0	75.6	77.5
KNN	–	–	83.4	81.8	82.7
GA-SVM	1.6104	3.2496	91.2	89.5	90.3
PSO-SVM	12.8717	5.8872	78.6	77.7	78.0
QGA-SVM	0.6655	2.0939	94.7	93.3	94.1

detecting fraud in the financial statements of business groups was developed. The proposed method can be utilized to support the development of Chinese text-based fraud detection systems. For example, it may be applied to annual reports and commercial emails.

(3) Fraud detection mechanism for financial statements of business groups: Based on the designed model and developed method, the fraud detection mechanism was implemented using Python 2.7 and Matlab R2014a. Business groups in Taiwan are employed as an example to demonstrate the feasibility and validity of the approach proposed in this study. This mechanism enhances the accuracy of fraud detection, fulfilling the aim of this study.

The results of this research facilitate the realization of fraud detection for the financial statements of business groups and the

Table 17
Classification accuracy comparison for undue deposit and debt.

Classifier	*c*	*g*	Fraud (%)	Non-fraud (%)	Total accuracy (%)
Decision tree C4.5	–	–	74.4	75.6	75.3
Logistic regression	–	–	88.7	86.1	87.5
BPN	–	–	89.4	85.9	87.5
KNN	–	–	81.3	79.5	80.2
GA-SVM	1.2745	1.0938	91.0	88.8	90.0
PSO-SVM	1.2027	7.3662	88.5	86.0	87.4
QGA-SVM	8.1577	3.3955	96.2	93.7	94.9

Table 18
Classification accuracy comparison for misstated financial statements.

Classifier	*c*	*g*	Fraud (%)	Non-fraud (%)	Total accuracy (%)
Decision tree C4.5	–	–	75.7	74.4	75.0
Logistic regression	–	–	83.3	81.1	82.3
BPN	–	–	79.3	75.5	77.6
KNN	–	–	82.2	78.7	80.4
GA-SVM	7.2040	0.1926	88.6	85.5	87.1
PSO-SVM	12.1507	0.1000	89.0	86.9	87.9
QGA-SVM	5.5839	0.5358	93.1	91.8	92.4

enhancement of fraud detection accuracy. These findings may be used to reduce investment losses and risks and enhance investment benefits for investors and creditors.

Compliance with ethical standards

(In case of Funding) Funding: This study was funded by Ministry of Science and Technology, R.O.C, Taiwan (Grant number MOST105-2410-H-327-014).

Conflict of Interest: Author A has received research grants from Company A. Author B has received a speaker honorarium from Company X and owns stock in Company Y. Author C is a member of committee Z. OR if no conflict exists: Author A declares that he/she has no conflict of interest. Author B declares that he/she has no conflict of interest.

(In case animals were involved) Ethical approval: All applicable international, national, and/or institutional guidelines for the care and use of animals were followed.

(And/or in case humans were involved) Ethical approval: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

(If articles do not contain studies with human participants or animals by any of the authors, please select one of the following statements) Ethical approval: This article does not contain any studies with human participants performed by any of the authors.

(Or) Ethical approval: This article does not contain any studies with animals performed by any of the authors.

(Or) Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors.

(In case humans are involved) Informed consent: Informed consent was obtained from all individual participants included in the study.

Acknowledgements

The authors would like to thank the Ministry of Science and Technology, Taiwan, for financially supporting this research under Contract No. MOST105-2410-H-327-014. Additionally, we deeply appreciate the editor, associate editor and reviewers for their constructive comments, suggestions, and proofreading on the work.

References

Chary, V.R.K., 2004. Ethics in Accounting. Global Cases and Experiences, Punjagutta. The ICFAI University Press, India.
Chen, J., Ye, T., Wang, H., 2015. Big data based fraud risk management at Alibaba. J. Financ. Data Sci. 1 (1), 1–10.
Chen, Y.J., Chen, Y.M., Lu, C.L., 2017. Enhancement of stock market forecasting using an improved fundamental analysis-based approach. Soft. Comput. 21 (13), 3735–3757.
http://ckipsvr.iis.sinica.edu.tw/, Chinese Knowledge and Information Processing.
Dutta, I., Dutta, S., Raahemi, B., 2017. Detecting financial restatements using data mining techniques. Expert Syst. Appl. 90, 374–393.
Friedman, J.H., 1991. Multivariate adaptive regression splines. Ann. Stat. 19 (1), 1–67.
Gepp, A., Linnenluecke, M.K., O'Neill, T.J., Smith, T., 2018. Big data techniques in auditing research and practice: current trends and future opportunities. J. Account.

Lit. 40, 102–115.

Glancy, F.H., Yadav, S.B., 2011. A computational model for financial reporting fraud detection. Decis. Support. Syst. 50 (3), 595–601.

Gray, G.L., Debreceny, R.S., 2014. A taxonomy to guide research on the application of data mining to fraud detection in financial statement audits. Int. J. Account. Inf. Syst. 15 (4), 357–380.

Güraksın, G.E., Haklıb, H., Uğuz, H., 2014. Support vector machines classification based on particle swarm optimization for bone age determination. Appl. Soft Comput. 24, 597–602.

Jekabsons, G., 2010. VariReg: a software tool for regression modelling using various modeling methods. Riga Technical Universityhttp://www.cs.rtu.lv/jekabsons/.

Jina, X., Wah, B.W., Cheng, X., Wang, Y., 2015. Significance and challenges of big data research. Big Data Res. 2 (2), 59–64.

Kaplan, R.S., Kiron, D., 2004. Accounting fraud at WorldCom. In: Harvard Business School Case 104-071.

Kirkos, E., Spathis, C., Manolopoulos, Y., 2007. Data mining techniques for the detection of fraudulent financial statements. Expert Syst. Appl. 32 (4), 995–1003.

Koc, E.K., Bozdogan, H., 2015. Model selection in multivariate adaptive regression splines (MARS) using information complexity as the fitness function. Mach. Learn. 101 (1–3), 35–58.

Nguyen, K., 2010. Financial Statement Fraud: Motives, Methods, Cases and Detection. Dissertation.com, Boca Raton, Florida, USA.

Peng, X., Xu, D., 2013. A twin-hypersphere support vector machine classifier and the fast learning algorithm. Inf. Sci. 221, 12–27.

Ravisankar, P., Ravi, V., Rao, G.R., Bose, I., 2011. Detection of financial statement fraud and feature selection using data mining techniques. Decis. Support. Syst. 50 (2), 491–500.

Stern, H., Chassidim, Y., Zofi, M., 2006. Multiagent visual area coverage using a new genetic algorithm selection scheme. Eur. J. Oper. Res. 175 (3), 1890–1907.

Suraj, S., Sesia, A., 2011. The crisis at Tyco - a director's perspective. In: Harvard Business School Case 111-035.

Swartz, M., Watkins, S., 2003. Power Failure: The Inside Story of the Collapse of Enron. Library of Congress Cataloging-in-Publication Data (USA, ISBN 0-385-50787-9).

http://www.tej.com.tw/twsite/, Taiwan Economic Journal.

http://www.twse.com.tw/ch/index.php, Taiwan Stock Exchange.

Ten Things About Financial Statement Fraud - Third Edition, a Review of SEC Enforcement Releases, 2000–2008. Deloitte Forensic Center, Deloitte.

Tian, X., Han, R., Wang, L., Lu, G., 2015. Latency critical big data computing in finance. J. Financ. Data Sci. 1 (1), 33–41.

Tsang, E., Yung, P., Li, J., 2004. EDDIE-automation, a decision support tool for financial forecasting. Decis. Support. Syst. 37 (4), 559–565.

Vernadat, F.B., 1996. Enterprise Modeling and Integration: Principles and Applications. Chapman & Hall, London, New York.

Zhang, W., Goh, A., 2016. Evaluating seismic liquefaction potential using multivariate adaptive regression splines and logistic regression. Geomech. Eng. 10 (3), 269–284.

Zhang, W., Yoshida, T., Tang, X., 2011. A comparative study of TF*IDF, LSI and multi-words for text classification. Expert Syst. Appl. 38 (3), 2758–2765.

Zhou, W., Kapoor, G., 2011. Detecting evolutionary financial statement fraud. Decis. Support. Syst. 50, 570–575.