Decision Support Systems 182 (2024) 114231

Contents lists available at ScienceDirect

Decision Support Systems

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

The information content of financial statement fraud risk: An ensemble learning approach

Wei Duan [a], Nan Hu [b], Fujing Xue [c,*]

[a] School of Management, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China
[b] School of Computing and Information Systems, Singapore Management University, 178902, Singapore
[c] Business School, Sun Yat-Sen University, Shenzhen 518107, China

ARTICLE INFO

Keywords:
Decision support systems
Ex-ante fraud risk
Ensemble learning
Feature engineering
Operational efficiency

ABSTRACT

This study aims to assess the financial statement fraud risk ex ante and empirically explore its information content to help improve decision-making and daily operations. We propose an ex-ante fraud risk index by adopting an ensemble learning approach and a theoretically grounded framework. Our ensemble learning model systematically examines the fraud process and deals effectively with the unique challenges in the financial fraud setting, which yields superior prediction performance. More importantly, we empirically examine the information content of our estimated ex-ante fraud risk from the perspective of operational efficiency. Our empirical results find that the estimated ex-ante fraud risk is negatively correlated with sustaining operational efficiency. This study redefines fraud detection as an ongoing endeavor rather than a retrospective event, thus enabling managers and stakeholders to reconsider their operation decisions and reshape their entire operation processes accordingly.

1. Introduction

As a start in daily operational decisions, the integrity and trustworthiness of information are often distorted by fraudulent activities, leading to efficiency reduction and risk increment [34]. Quite a few firms engage in fraud until recent years. A typical example is that, in 2020, Luckin Coffee was involved in financial statement fraud involving over US$300 million in fabricated sales—an event that prompted NASDAQ to delist the company and dismiss its CEO and COO, which impacted not only the company's operations but also its market reputation and investor trust [54]. In the age of digital transformation, fraud is even further spurred. A survey conducted by Deloitte Touche Tohmatsu and the Institute of Directors says that around 63% of independent directors believe that fraud cases will rise in the next two years [47]. Moreover, fraudulent financial information not only impacts the firm itself but also has a spillover effect on stakeholders, such as peers, suppliers, and regulators [8,44,45,59], especially when being used as an input for downstream decision-making [49], such as demand planning, performance forecasting, and regulatory monitoring. Therefore, it becomes crucial to evaluate firm's fraud risk ex ante and understand its information content.

Regulators and auditors have made significant efforts in monitoring financial statement fraud. Nevertheless, they cannot efficiently deal with fraud due to their inherent skill limitations and the costs of gathering and processing fraud-relevant information [24]. Furthermore, confirmation of fraud for regulators and auditors requires a large amount of time to obtain conclusive evidence. Usually, the damage has already been done when the conviction and punishment are completed. Therefore, the core problem is how to be aware of the financial statement fraud risks in a company before they are exposed. Once the fraud risk can be assessed ex ante, the entire decision and operation process can be reshaped and optimized, and companies can reconsider their choice of supply chain partners or investment objects to enhance stability. In brief, an ex-ante financial statements fraud risk index can bring a head start in operation and is urgently needed.

Previous studies that investigated financial fraud detection have proposed various models. Yet, due to challenges stemming from theoretical foundations, modeling techniques, feature selection, and other factors, some remaining problems, such as scalability [13], adaptability [1], and interpretability [6], suggest an ongoing need for refinement and advancement in the field, along with improved predictive performance. Furthermore, existing studies predominantly treat fraud as a retrospective event. Notably, there is a gap in the literature concerning the ex-ante estimation of fraud risk and the exploration of its informational

* Corresponding author.
E-mail addresses: duanwwei@stu.xjtu.edu.cn (W. Duan), nanhu@smu.edu.sg (N. Hu), xuefj@mail.sysu.edu.cn (F. Xue).

https://doi.org/10.1016/j.dss.2024.114231
Received 18 March 2023; Received in revised form 22 April 2024; Accepted 23 April 2024
Available online 27 April 2024

content.

With that in mind, in this study, we study the following research question (RQ). RQ (1): How can advanced modeling techniques, integrated with theoretical and novel features, produce a strong prediction on the ex-ante fraud risk? And RQ (2): What is the information content of the estimated fraud risk? Does it indicate lower efficiency?

To address our first question, we propose a novel model to detect corporate financial statement fraud based on theoretically grounded features and an ensemble learning approach. Features determine the upper limit of machine learning (ML) models. To increase the explanatory power of our ML model, we understand corporate fraud from a systemic perspective by regarding it as a process [4] and propose a framework with three dimensions of "why fraud," "how fraud," and "how fraud manifests." Drawing inspiration from Baucus [7], who dissected the antecedents of fraud into three factors, namely, pressure, opportunity, and predisposition, and building upon Misangyi et al.'s [43] insights into the significance of symbolic resources, particularly those linked to impression management in executing fraudulent schemes, our approach in the current study involves the comprehensive characterization of these dimensions. Such characterization is achieved through the extraction of both numerical and textual features, including thematic content through the latent Dirichlet allocation (LDA) algorithm.

Then we innovatively build our detection model based on ensemble learning, a well-respected approach that aligns closely with our research context. Ensemble learning harnesses the collective intelligence of multiple models to mitigate individual weaknesses and uncertainties as well as enhance overall predictive performance [48]. This collaborative strategy not only boosts accuracy but also promotes robustness and generalization across various datasets. The alignment lies in the synergy of individual models, thereby creating a dynamic and adaptive framework that excels in handling complex patterns, improving decision-making, and pushing the boundaries of what can be achieved in predictive analytics. Through the efforts made in the modeling process and sampling method, our ensemble learning approach deals effectively with the unique challenges involved in the financial fraud detection setting, such as the issue regarding unbalanced data issue and the confounding influence of extreme values. Our ensemble learning model, validated through 10-fold cross-validation, exhibits strong performance with an AUC of 0.734 and sensitivity of 0.754, representing a performance superiority relative to other state-of-the-art methods in the same context. After establishing the classification model, we then focus on the estimated firm-year fraud probability before classifying by the threshold value and take it as the ex-ante fraud risk index.

Upon completing the estimation of fraud risk (RQ1), we then move to RQ2 to understand the information content of the estimated fraud risk. Specifically, we investigate whether fraud risk is related to a firm's future operational efficiency. The impact of fraud risk on a firm's operational efficiency can be complex. On the one hand, fraudulent activities divert resources [32,60], disrupt standard procedures [17], and erode trust [56], hindering operational effectiveness. On the other hand, presenting false financial health may temporarily attract investment and lower capital costs [21], potentially boosting efficiency. Whether firms with high fraud risk will sustainably maintain low operational efficiency remains an empirical inquiry.

In this study, we use the generalized method of moments (GMM) estimation technique to analyze the data and find that fraud risk is negatively correlated with a firm's future operational efficiency. In summary, our study makes contributions in several important ways:

(1) First, our study makes a significant contribution to the field of fraud detection methodologies by introducing an ensemble learning model. Through the integration of theoretical foundations, advanced modeling techniques, and innovative feature extraction, our model makes substantial progress regarding the persistent challenges of scalability, adaptability, and interpretability. Moreover, our approach demonstrates performance superiority when compared with other state-of-the-art methods, showing its potential effectiveness in handling unique challenges within the financial fraud detection setting.
(2) Second, our study proposes an ex-ante financial statement fraud risk index, thus introducing a forward-looking perspective of assessing fraud risk rather than relying solely on retrospective assessments. Thus, compared with prior binary classification of fraud and non-fraud studies, our study allows for a more comprehensive understanding of varying levels of fraud risk.
(3) Third, our study investigates the information content of the fraud risk from the perspective of operational efficiency. Furthermore, we propose theoretical and logical support for the connection between fraud risk and operational efficiency and validate this relationship through empirical methods. In doing so, this paper sheds light on the risk management and operational efficiency literatures, thereby providing valuable insights through which organizations and individuals to optimize their decisions and enhance their operational practices.

2. Background and literature review

2.1. Consequences of fraud in operational management

Recent studies have pointed out that once a firm is accused of fraud, it usually suffers decreased market value [31,51], increased CEO turnover [28], and investment inefficiency [36]. A stream of literature also looked at the spillover effect of fraud on stakeholders. A firm's financial statement has been a common public information channel for stakeholders to retrieve integrated information [12,22]. When firms engage in financial statement fraud, the eroded information accumulates and propagates through the aggregation process. Based on such wrongful information, stakeholders are likely to make biased decisions, resulting in sub-optimal operational outcomes. For example, Beatty et al. [8] provided systematic evidence that fraudulent information leads to over-investment of industry peers. Yin et al. [59] also found that customers' fraudulent information distorts suppliers' investment decisions.

In addition, fraud also damages the economic system and social progress by implying perversion of order, ideal, and trust. For example, financial report restatements by major customers or industry peers tended to increase loan spreads for borrowers according to the study of Files and Gurun [25]. Scholars have also proposed a stigma effect that generates a negative spillover on industry peers, when one firm reveals financial misconduct [44,45]. Furthermore, when fraud becomes a lingering shadow, the level and cost of operational control are forced to increase, thus reducing productivity and decision-making efficiency across the entire sector [8]. However, prior studies have always taken fraud as a retrospective event. To date, no study has focused on the estimation of fraud risk ex ante and investigate its information content.

2.2. Big data analytics in predicting fraud

Answering the call for adaptive and evolutionary fraud detection techniques, many powerful models and more precise algorithms have been developed in the recent decade. For instance, Cecchini et al. [13] developed the SVM-FK model to beat the traditional logistic model for detecting financial frauds. Their SVM using the newly constructed financial kernel correctly labeled 80% of the fraudulent cases on a holdout set. In another example, Zhou and Kapoor [61] examined the effectiveness and limitations of common mining methods such as logistic regression, decision trees, neural networks, and Bayesian networks. Abbasi et al. [1] developed a novel meta-learning framework called MetaFraud to detect financial frauds. Their experimental results showed that the MetaFraud framework works with both legitimate and fraud sensitivity of over 80% for different stakeholder cost settings. Bao et al. [6] proposed a powerful ML method called "RUSboost", which is based

on raw financial data to predict frauds. Their results showed that RUSboost used the financial data more efficiently and outperformed the logistic regression method and SVM model in predicting frauds out of the sample.

The abovementioned algorithms based on big data have led to advancements in financial fraud prediction. However, such fraud detection models tend to be data-driven, and few studies have focused on the interpretability of [19] with a theoretical foundation. To address this gap, our study introduces an ensemble learning model that broadens the horizons of fraud detection methodologies. The proposed model provides a comprehensive approach by integrating the advantages of theoretical foundations, advanced modeling techniques, and innovative feature extraction methods. Furthermore, our study enriches the literature on addressing operation management questions through descriptive, predictive, and prescriptive big data analytics [15,57]. The current work also answers the call for improving data quality in big data analytics by enhancing interpretability and accessibility [5,20,26].

2.3. Fraud risk and operational efficiency

The impact of fraud risk on firm's future operational efficiency can be confounding. On the one hand, a firm's fraud risk is likely to negatively influence its operational efficiency. Such operational efficiency depends on a firm's resources, routines, and capabilities [37,46]. "Resources" encompass all assets that the firm can utilize for productive purposes [2,46]. From the resource-based view (RBV), firms can achieve a competitive advantage by leveraging a complex and unique set of resources that are rare, valuable, inimitable, and non-substitutable [18,29,37]. Firms engaging in fraudulent activities often allocate their resources inappropriately [32,60]. Covering up fraudulent activities requires extra effort, time, and resources that could have been utilized for production and operations. These resources are redirected, wasted, or misused, thereby depriving the firm of necessary support for its operational efficiency. "Routines" refer to the standard operating procedures and patterns of behaviors employed by a firm to achieve its desired outcomes through the use of resources [33]. To some extent, fraudulent activities may destroy these standard procedures. For example, to carry out its fraudulent activities, a firm may deviate from the established processes and procedures to conceal its actions, thus leading to reduced efficiency. Moreover, the internal control systems, as an important routine, may fall apart by employee collusion or management override, further deteriorating a firm's operational efficiency [17]. Meanwhile, "capabilities" represent the strength and expertise derived from a combination of interconnected routines, enabling a firm to carry out specific tasks proficiently [46]. The attainment of capability is established through the dynamic exchange of information and knowledge among individuals, customers, and various institutions [37]. In fraudulent firms, the intertwining of genuine and false information poses significant obstacles to accurate assessment and sound decision-making in daily operations. The lack of reliability in generated data and reports make it difficult for an organization to make informed decisions. Moreover, as an immoral behavior, fraud generates mutual distrust and fosters the culture of silence [56] within a company, significantly impeding the free exchange of information and knowledge among its members. As a result, creativity and innovation, which are the crucial determinants of capabilities and efficiency, are considerably suppressed.

On the other hand, a firm's fraud risk is also likely to positively influence its operational efficiency. In particular, a whitewashed financial statement can create a false impression of financial health and firm stability. It may give the illusion that the firm is performing well and has strong financial prospects, which can temporarily boost investor confidence and attract capital, thus leading to more investments and partnerships as well as lower capital costs [21] in the short term. If companies avoid being detected, their fraudulent activities may provide them with additional resources, potentially improving their operational efficiency. Consequently, the question of whether companies with a high likelihood of fraud risk will sustainably maintain low operational efficiency in the future remains an empirical inquiry.

3. Constructing the financial statement fraud risk index

3.1. Fraud sample

Fraud is a complex and extensive concept. In this work, we focus on financial statement fraud and correspondingly identify the precisely matched fraud sample. Our fraud sample comes from the China Stock Market and Accounting Research Database (CSMAR), which collects enforcement reports of violating firms provided by regulatory agencies, including the China Securities Regulatory Commission (CSRC), the Shanghai and Shenzhen stock exchanges, and the Ministry of Finance.

Our fraud sample covers A-share listed firms with financial statement frauds in China during the period from 2007 to 2018 (after eliminating financial industry firms). The period starts from 2007 because the Chinese accounting standards experienced considerable changes and effectively achieved international convergence during that year. Then, the sample ends in 2018 because of the lag between the occurrence of fraud and the announcement of the official enforcement actions disclosed in the market. According to the enforcement reports, on average, it takes 2.6 years for fraud to occur from occurrence up to the time it is discovered. Thus, we collect the enforcement reports for the period of January 2007 to May 2020. Following Lisic et al. [39], when a firm engaged in financial fraud for more than one year, we consider each year a fraud-year observation. Ultimately, our fraud sample included 395 firm-year observations.

To obtain a clean sample set of nonfraud companies, we first eliminate firm-year observations, of which the firms have committed financial statement fraud from 2007 to 2018 over the entire sample period. Second, we eliminate the firms that take the value of zero on financial statement fraud but have other irregularities. These firms are not entirely healthy firms and may be applicable to other forecasting models. Finally, our nonfraud sample included 22,976 firm-year observations. Table 1 shows the sample selection procedure followed for the fraud and nonfraud samples over the period 2007–2018. All the samples are firm-year observations.

3.2. Feature engineering and composition of predictive input features

"Feature engineering" refers to the process of extracting informative features from raw data to better represent the underlying interpretation of a phenomenon being investigated [35]. Specifically, feature engineering is an application of domain knowledge and theoretical

Table 1
Sample Selection Procedure.

Panel A. Sample selection of fraud sample	
Number of fraud samples	Number
Enforcement reports of violating firms by CSRC, stock exchanges and the Ministry of Finance	839
Less: samples beyond the period of 2007–2018	(241)
Less: unsatisfactory types of irregularities	(203)
Total	**395**

Panel B. Sample selection of non-fraud sample	
Number of non-fraud samples	Number
Firm-year observations satisfying that firms (non-financial firms) have A-shares traded on the Shanghai and Shenzhen stock exchanges from 2007 to 2018	29,132
Less: firm-year observations that firms had irregularities of any type in any year	(6,156)
Total	**22,976**

foundations to data science. In this study, we form our input features from the framework with three dimensions shown in Fig. 1: why, how, and the manifestation of fraud.

Our first dimension of features investigates why and under what circumstances companies are prone to fraud. Baucus [7] answered the question of why fraud occurs by identifying three cause factors that precede fraud: pressure, opportunity, and predisposition. Corporate performance can be attributed to certain combinations of strategy, structure, and environment, as well as different configurations of combined firm and environmental factors that lead to corporate illegality [7]. We generate a series of condition characteristics related to the internal control situations (e.g., board composition and management power) and external supervision environments (e.g., audit and analyst attention). Our model involves 16 informative firm condition characteristics as part of the input features. Detailed feature definitions and calculations are presented in Table 2.

Our second dimension of features is expected in terms of how corporate fraud occurs. The use of the LDA algorithm brings new insights into financial fraud detection by analyzing what managers discuss in the management's discussion and analysis (MD&A) section of a financial report [14,30,58]. Misangyi et al. [43] suggested the role of symbolic resources related to impression management in carrying out fraud schemes in operation. Hoberg and Lewis [30] found evidence indicating that fraudulent managers preach about their brilliant performance but disclose fewer details explaining the sources of their performance. This means that managers can strategically control topics or be evasive about critical financial information to achieve their fraud scheme. Meanwhile, greater linkages between financial fraud and managers' discussion are implicit and subtle. In relevant discussions, managers are likely to emphasize their fraud incentives, such as meeting requirements, reducing financing costs, and preserving company prestige. We can find evidence on how managers execute fraud based on the topics discussed. LDA, an unsupervised model developed by Blei et al. [10], can help quantify the extent to which each topic is discussed in individual MD&As. We set the number of topics to 90 as the optimal choice fitting our model. Details of the choice of optimal topic number are in the Appendix.

Our last dimension of features focuses on the manifestation of financial fraud, including both numerical and textual cues. Financial frauds always involve aggressive revenue recognition, false articulation, and abnormal financial index. The circumstance is complicated but traceable. Our model involves 17 informative financial metrics that monitor the red flags on accounts receivable, inventory, cash flow, employee, accrual, and market performance. To calculate these financial metrics, we obtain the original financial information before the restatement from the China Center for Economic Research (CCER) database, which retains the first version of financial statements released in April of each year. Table 2 defines each of the features outlined above.

However, quantitative financial information provides investors with an incomplete picture of a firm's economic situation without text information. Textual style facilitates the processing of quantitative information and generally informs the reader. Furthermore, changes in the textual style are often unconscious when fraud occurs, which leaves a new way to detect fraud. When textual data in financial statement are used in conjunction with traditional numerical variables, the accuracy of deep learning models for bankruptcy prediction can be further improved [42]. Churyk et al. [53] suggested that textual style in MD&A has an advantage over quantitative methods in determining deception promptly. Specifically, Loughran and McDonald [41] reported that negative and uncertain language is positively significantly associated with securities litigation involving suspected accounting misconduct. Prior studies also suggested that textual readability of corporate disclosure [27,40] is a great predictor of misreporting. Lo et al. [40] suggested that firms with lower textural readability in their MD&A are more likely to have managed earnings and misstatements. Goel and Gangolly [27] found that compared with nonfraudulent firms. Fraudulent firms attempt to be greater use of complex sentential structures and complicated words. In addition, Hoberg and Lewis [30] found that firm's disclosure similarity to the industry is also positively linked to the probability of fraudulent behavior. Therefore, we add these three main textual style metrics to our models, including readability, similarity, and tone. Detailed construction process is presented in the Appendix.

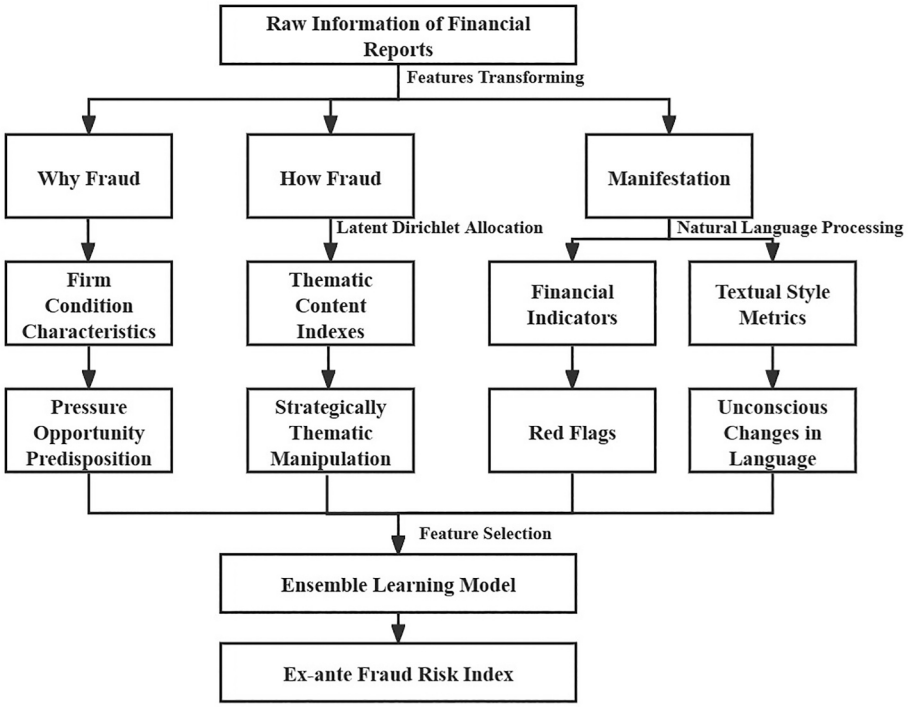

Fig. 1. Fraud Risk Assessment Framework.

Table 2
Feature definitions.

Feature	Definition
Panel A. Firm condition characteristics	
Asset	Natural logarithm of total assets.
Age	Natural logarithm of one plus firm age.
Audit	Dummy variable with a value of 1 if the firm has an unqualified audit opinion, otherwise 0.
Audit Firm	Dummy variable with a value of 1 if the firm switches the audit firm, otherwise 0.
BO	Percentage of the firm that board members own.
IO	Percentage of the firm that institutional investors own.
MO	Percentage of the firm that managers own.
CENT	Percentage of the firm that the first shareholder owns.
Analyst	Number of analysts following the firm.
Contract_am	Amounts of related-party transactions.
Contract_num	Number of related-party transactions.
HHI	Herfindahl-Hirschman Index.
SOE	Dummy variable with a value of 1 if a firm's ultimate controller is a government agency or government-owned entity, and zero otherwise.
Duality	Dummy variable with a value of 1 if chairperson of board holds managerial position CEO or president, otherwise 0.
SEOs	Dummy variable with a value of 1 if a firm has an average ROE of 6% to 7% over the last three years, which just meet the seasoned equity offering requirement.
Delisting	Dummy variable with a value of 1 if a firm has a ROE of 0% to 2% while the ROE for the previous two years is <0, which just meet the conditions of not being delisted.
Panel B. Financial metrics	
Accrual	Accrual earnings management calculated by cash flow model.
Abs_accrual	Absolute value of accrual earnings management.
Rem	Real earnings management calculated by cash flow model.
Rec_turn	Account Receivable Turnover.
Inv_turn	Inventory Turnover.
Cha_rec	Change in Accounts Receivable.
Cha_inv	Change in Inventory.
Cha_roa	Change in ROA.
Leverge	Total Liabilities / Total Assets.
Ab_marg	Gross margin - Average gross margin of the industry.
Cash_pro	Net Operating Cash Flow / Net Profit.
Soft	Total Assets - PP&E.
Sale_cash	Sales to Cash Flow Ratio.
BM	Book-to-Market.
PE	Price-Earnings Ratio.
Ab_emplnum	Percentage change in employees − Percentage change in assets.
Return	Market-adjusted Stock Return.
Panel C. Textual style metrics	
Tone_full	The linguistic feature of the full text of firm i's annual report in year t.
Read_full	The degree of comprehension about the full text of firm i's annual report in year t.
Sim_indu_full	The similarity between the full text of firm i's annual report in year t and that of other firms in the same industry in year t.
Sim_self_full	The similarity between the full text of firm i's annual report in year t and that of this firm in year t-1.
Tone_MDA	The linguistic feature of the MD&A section of firm i's annual report in year t.
Read_MDA	The degree of comprehension about the MD&A section of firm i's financial report in year t.
Sim_indu_MDA	The similarity between the MD&A section of firm i's annual report in year t and that of other firms in the same industry in year t.
Sim_self_MDA	The similarity between the MD&A section of firm i's annual report in year t and that of this firm in year t-1.

3.3. Bootstrap aggregating and balanced random forest

Ensemble learning methods, which combine predictions of each base learners through boosting, bagging, stacking, or other related approaches and have developed as the most powerful data mining technique and machine learning methods in recent years [9]. The ensemble learning approach is a multiple classifier system that can reduce model bias and variance and simultaneously improve its prediction performance. By combining multiple inducers, the ensemble learning model compensates the errors of single inducers (e.g., overfitting and local optima) and extends the search space to better fit the data space [48]. Specifically, we employ a bootstrap aggregating (bagging) method, Balanced Random Forest (BRF) [16], to build our detection model. Bagging is an ensemble technique for improving the robustness of forecasts, and the core idea of bagging is model averaging. Instead of choosing one estimator, bagging considers a set of estimators trained on the bootstrap samples and takes their average output, thus helping decrease the variance of an estimator. The bagging procedure is as follows:

1) Generate bootstrap sample $\{(y_1^s, x_1^s), \ldots, (x_N^s, y_N^s)\}$ via randomly drawing with replacement, with$s = 1, \ldots, S$.
2) Estimate $\widehat{f}_s(x)$ via minimizing the loss function.

$$\min_{\widehat{f}_s(x)} \sum_{i=1}^{N} \left(y_i^s - \widehat{f}_s(x_i^s)\right)^2. \tag{1}$$

3) Construct the bagging estimate combining all the estimated forecasts.

$$\widehat{f}(x)_{bagging} = \frac{1}{S}\sum_{s=1}^{S} \widehat{f}_s(x). \tag{2}$$

BRF is a modification of random forest (RF) [11], but unlike RF, BRF replaces the step of bootstrap sample generation for each tree with a more balanced method that is more suitable for unbalanced sample cases. The learning process of BRF operates in three steps:

1) For each tree, draw a bootstrap sample from the minority class. Randomly draw the same number of instances, with replacement, from the majority class.
2) At each node, instead of searching through all features for the optimal split, only search through a set of randomly selected features.
3) Aggregate the individual tree classifications and make the final prediction.

BRF is able to handle imbalanced data, reduce overfitting, preserve information, leverage ensemble learning, perform automatic feature selection, and maintain scalability and efficiency. These advantages collectively make BRF a strong candidate for addressing the challenges posed by imbalanced financial fraud datasets and ultimately improving fraud detection performance.

3.4. Model evaluation

3.4.1. Out-of-sample performance

Next, we examine the performance of the prediction model. By default, we use a threshold of 0.5 to classify samples into positive or negative classes, allowing us to calculate precision, sensitivity, and F1 score based on this classification. AUC remains unaffected by class imbalances and classification thresholds, which means it is a robust metric that facilitates high comparability across different models. By prioritizing AUC, we effectively assess the models' comprehensive discriminative power, making it especially relevant for real-world applications, wherein class proportions may be unequal and the threshold for classification is subject to each information user or system operator.

In the trade-off between precision and sensitivity, we prefer the model with higher sensitivity. Emphasizing sensitivity allows us to assess the model's ability to correctly capability of catching true violators, which is crucial in fraud detection scenarios. We believe type II errors are more serious because the consequences of missing a fraudulent financial report can be severe for both businesses and individual

users alike. In addition, what we identify is potential fraud candidate only, the impact of a false positive can typically be addressed through additional verification steps or investigations. We recommend several key strategies. First, robust validation processes, including cross-referencing with independent sources and thorough investigations, are crucial to minimize false positives. Second, using continuous fraud probability predictions allows for a more nuanced analysis and resource allocation, reducing intervention costs and fatigue. Third, creating transparent communication channels for flagged companies ensures they understand the detection criteria and can contest false positives. These measures enhance fraud detection's effectiveness and fairness, mitigating false positives' impact.

Fig. 2 presents the AUC-ROC curve, demonstrating the trade-off between sensitivity and specificity across a range of classification thresholds, with the area under the curve indicating the overall performance of the model. Our AUC-ROC curve exhibits a concave shape that ascends towards the upper-left corner, suggesting excellent discriminatory power of the model.

Table 3 shows the performance of our model under rigorous 10-fold cross-validation. The aggregated confusion matrix presented in Panel A provides a detailed breakdown of the model's binary classification performance. With 75.4% of fraudulent cases successfully detected, our model demonstrates a commendable capability in accurately flagging instances of fraud. Panel B reports the detail performance metrics derived from 10-fold cross-validation. The value of AUC is 0.734, suggesting that the model exhibits a satisfactory level of discrimination, effectively capturing the true positive rate while maintaining a low false positive rate across different classification thresholds. As discussed above, a sensitivity of 0.754 indicates that our model demonstrates a considerable accuracy in correctly identifying fraudulent cases. To make our model more convincing, we conducted an elaborate comparison of our BRF model with other state-of-the-art models commonly employed in financial fraud detection, including RUSBoost, XGBoost, SVM, long short-term memory (LSTM), and logistic regression. The aggregated results in Panel B of Table 3, all derived from 10-fold cross-validation, indicate that the our BRF model consistently outperforms other state-of-the-art models across multiple evaluation metrics.

Table 3
Model Performance.

Panel A. Confusion matrix

	Predicted value = 0	Predicted value = 1
True value = 0	71.6% (13,940)	28.4% (5522)
True value = 1	24.6% (86)	75.4% (264)

Panel B. Out-of-sample performance of different methods

Method	AUC	Sensitivity	Precision	F1score
BRF	0.734	0.754	0.045	0.085
RUSBoost	0.671	0.706	0.034	0.064
XGBoost	0.525	0.603	0.044	0.047
SVM	0.653	0.606	0.036	0.068
LSTM	0.714	0.644	0.046	0.086
Logistic	0.656	0.734	0.030	0.058

3.4.2. *Final selected features and their importance score*

Feature selection is crucial to ML, because appropriate feature screening can prevent dimensional disasters, reduce training time, enhance the generalization ability, reduce the overfitting, and enable

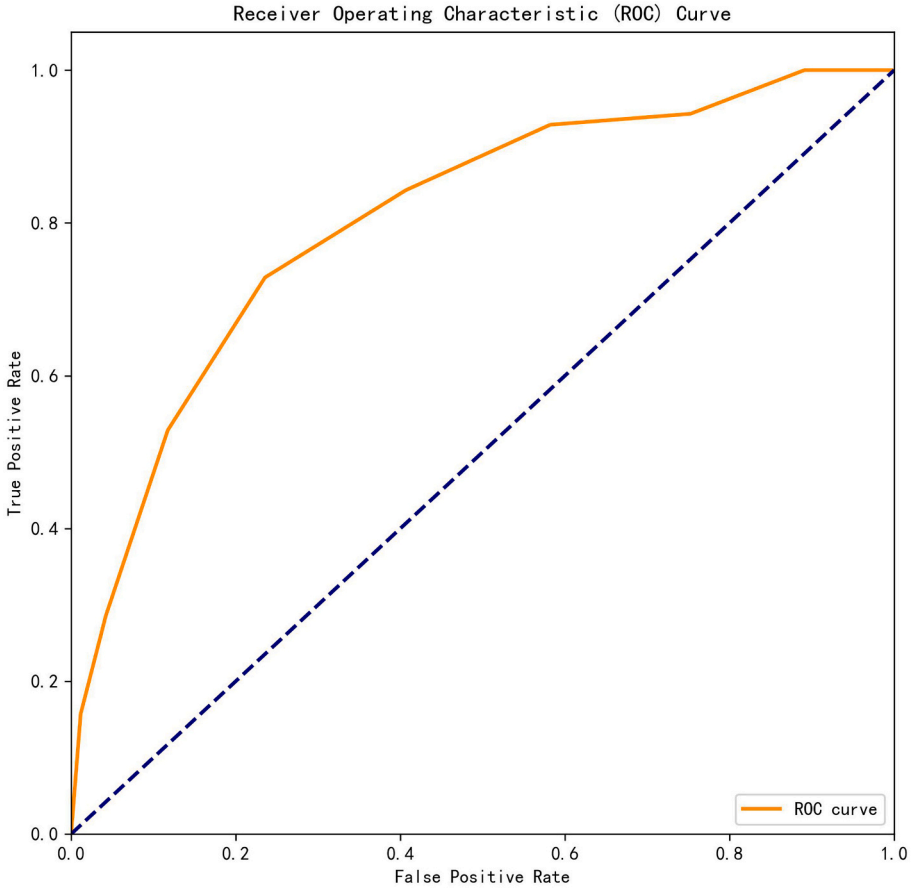

Fig. 2. AUC-ROC Curve.

the effect of the model to be close to its ceiling. Some effective methods have been proposed to help open the black box by estimating the importance of various features. We finalize our features and optimize the model based on the feature importance score, defined by Strobl et al. [52]. After normalizing all the importance scores of the 131 features, 46 significant features are chosen from 131 initial features. Table 4 reports the descriptive statistics on the importance of the selected 46 significant features for our ensemble learning model. The 46 significant features contain six firm condition characteristics, 16 financial metrics, 6 textual style metrics, and 18 thematic indices, indicating that features of different dimensions both play an important role in detecting financial statement fraud. The five most significant features are as follows: (1) proportion of net operating cash flow in net profit, (2) percentage of the firm that the first shareholder owns, (3) price-earnings ratio, (4) financial leverage, and (5) real earnings management.

We are also interested in determining what topics are related implicitly to financial statement fraud. Among the topics, 18 of the 90 topics play a role in the final model. Table 5 reports the list of highest weighted words in each of the top five topics with the highest importance score to understand the underlying content of each topic. These five topics are generally related to future uncertainty, production, new materials, performance comparison, and reorganization, respectively. These topics discussed in MD&A are informative in distinguishing among fraud firms. Table A2 in the Appendix presents details of the top five informative topics words in Chinese.

After establishing the classification model, we pay attention to the estimated firm-year fraud probability before performing classification using the threshold value and take this as the ex-ante fraud risk index.

4. Information content of the financial statement fraud risk

In this section, we investigate the information content of the estimated fraud risk from the perspective of operational efficiency. We examine whether the company with high fraud risk has been compensated or is in ongoing operational distress. Depending on resources, routine, and managerial capability, operational efficiency reflects the true operating condition and can be difficult to manipulate by fraud activities. We use the Stochastic Frontier Estimation (SFE) methodology to construct a measure of company operational efficiency [23,37,38]. SFE models the efficiency at which a company transforms various operational input resources into operational performance outcomes

Table 4
The most importance features in predicting financial fraud.

Rank	Selected feature	Importance score	Rank	Selected feature	Importance score
1	*Cash_pro*	0.0421	24	*Topic_9*	0.0199
2	*CENT*	0.0365	25	*Cha_inv*	0.0197
3	*PE*	0.0359	26	*Sale_cash*	0.0196
4	*Leverage*	0.0308	27	*Cha_rec*	0.0195
5	*Rem*	0.0292	28	*Accrual*	0.0191
6	*Topic_1*	0.0289	29	*Topic_10*	0.0190
7	*Tone_MDA*	0.0231	30	*Read_full*	0.0189
8	*Topic_2*	0.0223	31	*Sim_self_full*	0.0186
9	*Topic_3*	0.0222	32	*Topic_11*	0.0184
10	*Analyst*	0.0219	33	*Ab_emplnum*	0.0184
11	*Tone_full*	0.0218	34	*Age*	0.0184
12	*Topic_4*	0.0216	35	*Return*	0.0183
13	*Topic_5*	0.0216	36	*Topic_12*	0.0182
14	*Topic_6*	0.0214	37	*Topic_13*	0.0182
15	*Topic_7*	0.0212	38	*Topic_14*	0.0179
16	*Contract_am*	0.0207	39	*Ab_marg*	0.0178
17	*Sim_indu_MDA*	0.0204	40	*IO*	0.0175
18	*Soft*	0.0203	41	*Topic_15*	0.0174
19	*Cha_roa*	0.0203	42	*Topic_16*	0.0174
20	*Asset*	0.0202	43	*Topic_17*	0.0169
21	*Topic_8*	0.0200	44	*Topic_18*	0.0163
22	*Sim_self_MDA*	0.0199	45	*BM*	0.0161
23	*Sim_indu_full*	0.0199	46	*Abs_accrual*	0.0161

Table 5
The top five informative topics.

Topic	Representative words
Topic_1	risk, continually, future, possible, be faced with, change, market competition, policy, suppose, adverse impact
Topic_2	produce, reduce, price, capacity, optimize, output, purchase, raw materials, product structure, cost of production
Topic_3	materials, new material, environmental, application, material performance, high-performance, composite material, technology, develop, tombarthite
Topic_4	decrease, current period, same period last year, revenue, cash, decline, year-on-year, change, expense, increase
Topic_5	material assets reorganization, asset, issue, share, listed company, transaction, China Securities Regulatory Commission, reorganization, review, consideration

[38], thus capturing the idea of relative efficiency in transforming defined from the conventional OM perspective [37,50]. The corresponding SFE function can be specified as follows:

$$\begin{aligned} ln(Operating\ Income)_{ij} &= \beta_0 + \beta_1 ln(Number\ of\ Employees)_{ij} \\ &+ \beta_2 ln(Cost\ of\ Goods\ Sold)_{ij} \\ &+ \beta_3 ln(Capital\ Expenditure)_{ij} + v_{ij} - u_{ij}. \end{aligned} \tag{3}$$

where v_{ij} is the stochastic random error term and u_{ij} is the technical inefficiency term. We then estimated the operational efficiency of firm i in year t as follows:

$$Operational\ Efficiency_{ij} = 1 - \widehat{u}_{ij} \tag{4}$$

We use dynamic panel data models to examine the relationship between fraud risk and future operational efficiency. Following Lam et al. [37], we adopted the GMM estimation [3] with the standard lag of dependent variable over periods 1 to 3 as the GMM-type instruments. The specific model is as the following:

$$\begin{aligned} Efficiency_{i(t+1)} &= \beta_0 + \beta_1 LagEfficiency_{it} + \beta_2 FraudRisk_{it} + \beta_3 Age_{it} + \beta_4 Asset_{it} \\ &+ \beta_5 MktShare_{it} + \beta_6 Concentration_{it} + \beta_7 Foreign_{it} \\ &+ \beta_8 FCF_{it} + \beta_9 Error_{it} + IndustryFE + YearFE + \varepsilon_{it} \end{aligned} \tag{5}$$

where $Efficiency_{i(t+1)}$ refers to the measure of operational efficiency for firm i in year $t+1$, $Efficiency_{it}$ refers to one-year lagged dependent variables, and $FraudRisk_{it}$ refers to the probability of fraud predicted by our detection model for firm i in year t. We included one lag of the dependent variable to appropriately control for the path-dependent and persistent influence of operational efficiency over time [37,55]. Following Cheng et al. [17] and Demerjian et al. [20], we controlled the determinants of firm operational efficiency including life cycle (*Age*), firm size (*Asset*), market share (*MktShare*), industry diversification (*Concentration*), foreign operations (*Foreign*), and available cash (*FCF*). We also included the deviation of the predicted fraud risk from the true value (*Error*) to control the potential effect of prediction error. Table A3 presents detailed definitions. Table 6 presents the descriptive statistics for the full sample.

Table 7 presents the empirical results on the association between fraud risk and future operational efficiency. The coefficient on lagged dependent variables is positive and significant ($p < 0.01$), which confirms the efficiency persistence. After controlling the past performance and firm characteristics, we find that the coefficient on *FraudRisk* is negative and significant ($p < 0.01$) in year t + 1. We extend the test period to the second and third years and the results remain significant. Aside from GMM estimation, we also include the OLS estimation for the robust test, and we obtain similar empirical results. In accordance with the results, a high risk of fraud means firms will be in operational deficiency for a long time. In turn, such firms will fail to resolve their operational deficiencies through fraudulent activities. On the contrary, these firms appear to perpetuate a cycle of operational inefficiency due to the misallocation of resources, lack of organized operational

Table 6
Descriptive statistics for the full sample.

VarName	Mean	SD	P25	Median	P75	Min	Max
Efficiency	0.525	0.129	0.456	0.539	0.612	0.007	0.838
FraudRisk	0.451	0.104	0.377	0.446	0.519	0.106	0.894
Age	2.057	0.888	1.609	2.303	2.773	0.000	3.367
Asset	21.986	1.365	21.042	21.832	22.769	10.842	28.520
MktShare	0.027	0.073	0.002	0.005	0.019	0.000	1.000
Concentration	0.795	0.241	0.572	0.933	1.000	0.203	1.000
Foreign	0.431	0.495	0.000	0.000	1.000	0.000	1.000
FCF	0.220	0.414	0.000	0.000	0.000	0.000	1.000
Error	0.444	0.101	0.374	0.443	0.515	0.106	0.782

Table 7
Fraud risk and operational efficiency.

	GMM *Efficiency* $_{t+1}$	GMM *Efficiency* $_{t+2}$	GMM *Efficiency* $_{t+3}$	OLS *Efficiency* $_{t+1}$	OLS *Efficiency* $_{t+2}$	OLS *Efficiency* $_{t+3}$
FraudRisk	−0.267***	−0.191**	−0.170**	−0.025**	−0.045***	−0.016
	(−3.97)	(−2.25)	(−2.05)	(−2.01)	(−3.49)	(−1.17)
LagEfficiency	0.653***	0.617***	0.570***	0.697***	0.703***	0.702***
	(10.52)	(5.10)	(5.49)	(109.87)	(109.18)	(102.11)
Age	−0.001	0.030*	0.041**	−0.001	−0.002*	−0.001
	(−0.33)	(1.85)	(2.34)	(−0.64)	(−1.88)	(−0.51)
Asset	−0.015**	−0.045**	−0.048**	0.007***	0.006***	0.006***
	(−2.54)	(−2.02)	(−2.18)	(9.93)	(8.75)	(7.23)
MktShare	0.072*	0.271*	0.263**	−0.007	0.006	−0.005
	(1.88)	(1.93)	(2.03)	(−0.63)	(0.59)	(−0.47)
Concentration	−0.015	0.009	−0.002	0.004	0.009***	0.011***
	(−1.56)	(0.67)	(−0.16)	(1.42)	(2.92)	(3.42)
Foreign	0.006*	0.023**	0.048***	−0.002	−0.001	0.001
	(1.80)	(2.29)	(3.19)	(−1.33)	(−0.76)	(0.33)
FCF	0.015	0.041	0.090**	0.008***	0.009***	0.008***
	(0.59)	(1.10)	(2.37)	(5.07)	(5.22)	(3.91)
Error	−0.054	−0.144*	−0.166**	0.016	0.030**	−0.010
	(−0.71)	(−1.74)	(−2.05)	(1.21)	(2.26)	(−0.72)
Industry FE	Y	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y	Y
AR(1)	$p = 0.000$	p = 0.000	p = 0.000			
AR(2)	$p = 0.330$	$p = 0.668$	$p = 0.728$			
Sargan	$P = 0.111$	$p = 0.746$	$p = 0.115$			
Intercept	1.116***	1.238***	1.313***	0.019	0.033**	0.049***
	(3.95)	(3.24)	(3.10)	(1.20)	(2.04)	(2.75)
N	13,467	11,615	9837	14,609	12,996	10,927
adj. R^2				0.601	0.604	0.600

Notes: *p*-values are in parentheses. Coefficients marked with***, **, and * are significant at the 1%, 5%, and 10% level, respectively.

procedures, and impediments in the share of information.

5. Conclusion and implications

Financial statement fraud has always been a social hazard with far-reaching concerns. It threatens the growth of enterprises themselves, the interests of stakeholders, and ultimately the sustainable development of the capital market. Using an ensemble learning approach, we propose an ex-ante fraud risk index that can help improve the integrity and trustworthiness of the input financial information. We further examine the information content of our ex-ante fraud risk index from the perspective of operational efficiency using empirical evidence.

Our study has considerable theoretical implications. First, this work introduces a new theoretical insight beyond conventional paradigms. Departing from binary classifications, the study introduces the ex-ante financial statement fraud risk index, which is a forward-looking instrument that assesses varying levels of risk. This perspective redefines fraud detection as an ongoing endeavor rather than a retrospective event. Second, a significant theoretical innovation lies in a systematic framework that summarize the process of financial statement fraud. This framework integrates the numerical, textual, and thematic features, thus emphasizing the interconnectedness of information. Furthermore, the theoretical implications extend to the concept of operational efficiency. By discerning whether fraud poses a threat or a potential short-term advantage, this research not only reframes the interaction between fraud and operational efficiency but also offers theoretical and logical foundations for this relationship. This theory-based insight reveals how fraudulent endeavors divert resources from productive activities, disrupt established operational routines, and hinder the development of essential capabilities, thereby enriching our understanding of how fraud influences operational efficiency.

Aside from its theoretical implications outlined above, our study also has considerable practical implications for both managers and various stakeholders. First, for firms themselves, engaging in fraudulent activities not only poses legal and reputational risks but also leads to sustained operating difficulties. Thus, managers should be aware of the importance of maintaining a strong ethical culture, implementing robust internal controls, and promoting transparency and integrity throughout the organization. They must also prioritize the optimal allocation of resources for value creation, avoiding diversion towards fraudulent practices. Second, regulators can benefit from the assessment of the ex-ante fraud risk index, because it offers an improved monitoring mechanism and decision support system. By integrating this index into their oversight processes, regulators can effectively identify and prioritize companies at a higher risk of financial statement fraud. This proactive approach, in turn, enables them to allocate resources efficiently and

focus on entities requiring closer scrutiny, enhancing regulatory efficiency and effectiveness. Third, the proposed ex-ante fraud risk index provides valuable information for analysts, investors, and supply chain stakeholders to make informed decisions. Using this index, analysts can incorporate the fraud risk index into their analysis to evaluate the risk associated with investment opportunities. Investors can also use the index to assess the credibility of financial statements and scrutinize investment opportunities, making more informed investment decisions. Meanwhile, supply chain stakeholders can utilize the index to select reliable partners, reduce the risk of fraudulent entities in their supply chains, and make informed sourcing decisions. This anticipatory approach transforms decision-making from reactive to proactive, thus enhancing overall stability and minimizing disruptions caused by fraudulent activities.

Finally, our study has several possible extensions. First, our research focuses on financial statement fraud detection, which is just one type of fraud. Future research can extend our method to other fraudulent activities, such as investment scams and financial mis-selling. Second, for fraudulent companies, their executives might leave traces through other information channels, such as changing their voices or gestures in earnings calls. Future research can thus take advantage of other potential informative cues in fraud detection.

CRediT authorship contribution statement

Wei Duan: Writing – review & editing, Writing – original draft, Software, Project administration, Methodology, Formal analysis, Data curation, Conceptualization. **Nan Hu:** Writing – review & editing, Supervision, Resources, Methodology, Investigation, Conceptualization. **Fujing Xue:** Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

Government of Singapore Ministry of Education Grant numbers: 22-SIS-SMU-042; National Natural Science Foundation of China Grant numbers: 72172118.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.dss.2024.114231.

References

[1] A. Abbasi, C. Albrecht, A. Vance, J. Hansen, Metafraud: a Meta-learning framework for detecting financial fraud, MIS Q. 36 (4) (2012) 1293–1327.
[2] R. Amit, P.J. Schoemaker, Strategic assets and organizational rent, Strateg. Manag. J. 14 (1) (1993) 33–46.
[3] M. Arellano, S. Bond, Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations, Rev. Econ. Stud. 58 (2) (1991) 277–297.
[4] B.E. Ashforth, D.A. Gioia, S.L. Robinson, L.K. Trevino, Re-viewing organizational corruption - introduction, Acad. Manag. Rev. 33 (3) (2008) 670–684.
[5] B. Baesens, S. Höppner, T. Verdonck, Data engineering for fraud detection, Decis. Support. Syst. 150 (2021) 113492.
[6] Y. Bao, B. Ke, B. Li, Y.J. Yu, J. Zhang, Detecting accounting fraud in publicly traded US firms using a machine learning approach, J. Account. Res. 58 (1) (2020) 199–235.
[7] M.S. Baucus, Pressure, opportunity and predisposition - a multivariate model of corporate illegality, Aust. J. Manag. 20 (4) (1994) 699–721.
[8] A. Beatty, S. Liao, J.J. Yu, The spillover effect of fraudulent financial reporting on peer firms' investments, J. Account. Econ. 55 (2–3) (2013) 183–205.
[9] C.M. Bishop, N.M. Nasrabadi, Pattern Recognition and Machine Learning, 4(4), Springer, 2006.
[10] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (4–5) (2003) 993–1022.
[11] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[12] L.D. Brown, A.C. Call, M.B. Clement, N.Y. Sharp, Inside the "black box" of sell-side financial analysts, J. Account. Res. 53 (1) (2015) 1–47.
[13] M. Cecchini, H. Aytug, G.J. Koehler, P. Pathak, Detecting management fraud in public companies, Manag. Sci. 56 (7) (2010) 1146–1160.
[14] M. Cecchini, H. Aytug, G.J. Koehler, P. Pathak, Making words work: using financial text as a predictor of financial events, Decis. Support. Syst. 50 (1) (2010) 164–175.
[15] R.M. Chang, R.J. Kauffman, Y. Kwon, Understanding the paradigm shift to computational social science in the presence of big data, Decis. Support. Syst. 63 (2014) 67–80.
[16] C. Chen, A. Liaw, L. Breiman, Using random forest to learn imbalanced data, University of California, Berkeley 110 (1–12) (2004) 24.
[17] Q. Cheng, B.W. Goh, J.B. Kim, Internal control and operational efficiency, Contemp. Account. Res. 35 (2) (2018) 1102–1139.
[18] T.T. Coates, C.M. Mcdermott, An exploratory analysis of new competencies: a resource based view perspective, J. Oper. Manag. 20 (5) (2002) 435–450.
[19] K. Coussement, D.F. Benoit, Interpretable data science for decision making, Decis. Support. Syst. 150 (2021) 113664.
[20] P. Demerjian, B. Lev, S. Mcvay, Quantifying managerial ability: a new measure and validity tests, Manag. Sci. 58 (7) (2012) 1229–1248.
[21] S.G. Dimmock, W.C. Gerken, Predicting fraud by investment managers, J. Financ. Econ. 105 (1) (2012) 153–173.
[22] S. Duhadway, S. Talluri, W. Ho, T. Buckhoff, Light in dark places: the hidden world of supply chain fraud, IEEE Trans. Eng. Manag. 69 (4) (2020) 874–887.
[23] S. Dutta, O. Narasimhan, S. Rajiv, Conceptualizing and measuring capabilities: methodology and empirical application, Strateg. Manag. J. 26 (3) (2005) 277–285.
[24] A. Dyck, A. Morse, L. Zingales, Who blows the whistle on corporate fraud? J. Financ. 65 (6) (2010) 2213–2253.
[25] R. Files, U.G. Gurun, Lenders' response to peer and customer restatements, Contemp. Account. Res. 35 (1) (2018) 464–493.
[26] M. Ghasemaghaei, G. Calic, Can big data improve firm decision quality? The role of data quality and data diagnosticity, Decis. Support. Syst. 120 (2019) 38–49.
[27] S. Goel, J. Gangolly, Beyond the numbers: mining the annual reports for hidden cues indicative of financial statement fraud, Intell. Syst. Account. Finance Manag. 19 (2) (2012) 75–89.
[28] D. Gomulya, W. Boeker, Reassessing board member allegiance: CEO replacement following financial misconduct, Strateg. Manag. J. 37 (9) (2016) 1898–1918.
[29] M.A. Hitt, K. Xu, C.M. Carnes, Resource based theory in operations management research, J. Oper. Manag. 41 (2016) 77–94.
[30] G. Hoberg, C. Lewis, Do fraudulent firms produce abnormal disclosure? J. Corp. Finan. 43 (2017) 58–85.
[31] J.M. Karpoff, D.S. Lee, G.S. Martin, The cost to firms of cooking the books, J. Financ. Quant. Anal. 43 (3) (2008) 581–611.
[32] S. Kedia, T. Philippon, The economics of fraudulent accounting, Rev. Financ. Stud. 22 (6) (2009) 2169–2199.
[33] M. Ketokivi, R. Schroeder, Manufacturing practices, strategic fit and performance: a routine-based view, Int. J. Oper. Prod. Manag. 24 (2) (2004) 171–191.
[34] M. Kraus, S. Feuerriegel, Decision support from financial disclosures with deep neural networks and transfer learning, Decis. Support. Syst. 104 (2017) 38–48.
[35] N. Kumar, D. Venugopal, L.F. Qiu, S. Kumar, Detecting review manipulation on online platforms with hierarchical supervised learning, J. Manag. Inf. Syst. 35 (1) (2018) 350–380.
[36] P. Kumar, N. Langberg, Corporate fraud and investment distortions in efficient capital markets, RAND J. Econ. 40 (1) (2009) 144–172.
[37] H.K.S. Lam, A.C.L. Yeung, T.C.E. Cheng, The impact of firms' social media initiatives on operational efficiency and innovativeness, J. Oper. Manag. 47-48 (2016) 28–43.
[38] S.L. Li, J. Shang, S.A. Slaughter, Why do software firms fail? Capabilities, competitive actions, and firm survival in the software industry from 1995 to 2007, Inf. Syst. Res. 21 (3) (2010) 631–654.
[39] L.L. Lisic, S. Silveri, Y.H. Song, K. Wang, Accounting fraud, auditing, and the role of government sanctions in China, J. Bus. Res. 68 (6) (2015) 1186–1195.
[40] K. Lo, F. Ramos, R. Rogo, Earnings management and annual report readability, J. Account. Econ. 63 (1) (2017) 1–25.
[41] T. Loughran, B. Mcdonald, When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks, J. Financ. 66 (1) (2011) 35–65.
[42] F. Mai, S.N. Tian, C. Lee, L. Ma, Deep learning models for bankruptcy prediction using textual disclosures, Eur. J. Oper. Res. 274 (2) (2019) 743–758.
[43] V.F. Misangyi, G.R. Weaver, H. Elms, Ending corruption: the interplay among institutional logics, resources, and institutional entrepreneurs, Acad. Manag. Rev. 33 (3) (2008) 750–770.
[44] I. Naumovska, D. Lavie, When an industry peer is accused of financial misconduct: stigma versus competition effects on non-accused firms, Adm. Sci. Q. 66 (4) (2021) 1130–1172.
[45] S. Paruchuri, V.F. Misangyi, Investor perceptions of financial misconduct: the heterogeneous contamination of bystander firms, Acad. Manag. J. 58 (1) (2015) 169–194.

[46] D.X. Peng, R.G. Schroeder, R. Shah, Linking routines to operations capabilities: a new perspective, J. Oper. Manag. 26 (6) (2008) 730–748.
[47] Pti, Fraud Cases to Spike in Next Two Years; Cyber Crimes, Financial Statement Frauds to Dominate: Survey, Available at: https://www.cnbctv18.com/information-technology/fraud-cases-to-spike-in-next-two-years-cyber-crimes-financial-statement-frauds-to-dominate-survey-11030032.htm, 2021 (Accessed 7 October 2021).
[48] O. Sagi, L. Rokach, Ensemble learning: a survey, Wiley Interdiscip, Rev.: Data Min. Knowl. Discov. 8 (4) (2018).
[49] A. Scott, A.T. Balthrop, The consequences of self-reporting biases: evidence from the crash preventability program, J. Oper. Manag. 67 (5) (2021) 588–609.
[50] C. Serrano-Cinca, Y. Fuertes-Callén, C. Mar-Molinero, Measuring DEA efficiency in internet companies, Decis. Support. Syst. 38 (4) (2005) 557–573.
[51] A.J. Sharkey, Categories and organizational status: the role of industry status in the response to organizational deviance, Am. J. Sociol. 119 (5) (2014) 1380–1433.
[52] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, A. Zeileis, Conditional variable importance for random forests, BMC Bioinformat. 9 (2008) 1–11.
[53] N. Tatiana Churyk, C.-C. Lee, B.D. Clinton, Early detection of fraud: Evidence from restatements, in: Advances in Accounting Behavioral Research, Emerald Group Publishing Limited, 2009, pp. 25–40.
[54] U.S. Securities and Exchange Commission, Luckin Coffee Agrees to Pay $180 Million Penalty to Settle Accounting Fraud Charges, Available at: https://www.sec.gov/litigation/litreleases/lr-24987, 2020 (Accessed 16 December 2021).
[55] R. Vandaie, A. Zaheer, Alliance partners and firm capability: evidence from the motion picture industry, Organ. Sci. 26 (1) (2015) 22–36.
[56] P. Verhezen, Giving voice in a culture of silence. From a culture of compliance to a culture of integrity, J. Bus. Ethics 96 (2010) 187–206.
[57] G. Wang, A. Gunasekaran, E.W.T. Ngai, T. Papadopoulos, Big data analytics in logistics and supply chain management: certain investigations for research and applications, Int. J. Prod. Econ. 176 (2016) 98–110.
[58] Y. Wang, W. Xu, Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud, Decis. Support. Syst. 105 (2018) 87–95.
[59] C. Yin, X. Cheng, Y.N. Yang, D. Palmon, Do corporate frauds distort Suppliers' Investment decisions? J. Bus. Ethics 172 (1) (2021) 115–132.
[60] F. Yu, X. Yu, Corporate lobbying and fraud detection, J. Financ. Quant. Anal. 46 (6) (2011) 1865–1891.
[61] W. Zhou, G. Kapoor, Detecting evolutionary financial statement fraud, Decis. Support. Syst. 50 (3) (2011) 570–575.

Wei Duan is a Ph.D. student in Xi'an Jiaotong University. His research interests are information quality and behavioral accounting & finance. He is also interested in the application of textural analysis, artificial intelligence and large language model in management study.

Nan Hu is an associate professor of information systems in Singapore Management University. His education background in information systems, statistics, and accounting enables him to examine a broad range of questions that can aid C-suite executives in improving the management of information, especially during the unstructured data era. His recent research focuses on the value implications and market efficiency of unstructured information (e.g., corporate narrative disclosure, conference call transcripts, social media, online consumer reviews, audio, and video). Nan has published in *Production and Operations Management, MIS Quarterly, Journal of Management Information Systems, Decision Support Systems* and other journals.

Fujing Xue is an assistant professor at the Business School in Sun Yat-sen University. Her research interests are information quality, financial decision, and the application of textural analysis and machine learning in management study. Fujing's research has appeared at *International Journal of Contemporary Hospitality Management*, *Pacific-Basin Finance Journal*.