# Non-financial Factors and Diversified Fraudulent Activities in China: A Machine Learning-based Approach

Yuchen ZHAO, Xiaogang BI*

Yuchen ZHAO: Ph.D. candidate in Nottingham University Business School China;

Tel: 13623700036;

E-mail: Yuchen.Zhao@nottingham.edu.cn.

Xiaogang BI:

Department of Finance and Accounting, Nottingham University Business School China, The University of Nottingham, 199 Taikang East Road, Ningbo, P. R. China, 315100;

E-mail: x.bi@nottingham.edu.cn.

This Version: September 2024

* Corresponding author.

# Non-financial Factors and Diversified Fraudulent Activities in China: A Machine Learning-based Approach


## Abstract

We develop a comprehensive framework for fraud prediction based on multi-dimensional financial and non-financial factors and employ two representative ensemble learning models—random forest and GBDT—to forecast the corporate violation tendency. Our analysis shows that the GBDT model adapting the boosting idea exhibits significant advantages in predicting corporate wrongdoing. We further find that non-financial factors are also important, while their relative importance varies when predicting different types of irregularities. This emphasises the key signalling roles played by non-financial variables, which have a significant impact on violation prediction, just as financial variables do. Our findings provide a new perspective and valuable reference for identifying the enterprise violation risk in addition to reliable support for the practical application of machine learning in corporate fraud detection in China.



**Keywords:** Corporate governance, machine learning, corporate misconduct, fraud prediction, corporate finance

## 1. Introduction

Corporate fraud belongs to corporate governance practices and is defined as illegal activities carried out by individuals or companies in a dishonest or unethical manner for profit. This topic has long attracted much attention and debate in corporate governance. Globally, the existence of irregularity has caused substantial economic losses as well as other devastating consequences, with severe negative impacts on investors and the capital market, and is viewed as a significant concern (Free and Murphy, 2015). Illegal and non-compliant behaviour of listed companies can, first of all, seriously damage the legitimate rights and interests of various groups of stakeholders. Corporate violations will reduce shareholder value (Karpoff et al., 2008; Khanna et al., 2015), lead to serious financial adverse consequences (Cumming et al., 2015b), and severely harm the firm's reputation (Greve et al., 2010). In addition to monetary and reputational losses, enterprises may suffer penalties from regulators and the marketplace for irregularities, such as civil and criminal penalties for enterprises or administrators, as well as a significant risk of losing future cash flows. Secondly, exposure to corporate wrongdoings can also hurt peer firms. Giannetti and Wang (2016) found that the probability of retail investors exiting the stock market increased by seven percentage points following the revelation of corporate breaches in the same state, suggesting that peer firms experience a negative spillover after the disclosure of a firm's violations (Gande and Lewis, 2009). Besides, the fraudulent behaviour of enterprises significantly and directly damages the interests of broad investors and has a destructive effect on their confidence in the securities market (Cumming et al., 2015a; Khanna et al., 2015). Moreover, the wrongdoing of listed firms disrupts the good order of the stock market, distorts the efficient resource allocation in the economy, increases the risk of market instability, and seriously impedes the healthy development of the entire capital market (Djankov et al., 2008; Kedia and Philippon, 2009; Hou and Moore, 2010). As a result, the

research of illicit behaviours by quoted companies has received close attention from regulators, investors, media, and academia worldwide.

In recent years, with the rapid growth of information technology, machine learning has gradually come to the fore, providing a new and efficient way to solve this challenge. As the core technology and important branch of artificial intelligence, machine learning, with its excellent data processing and prediction capabilities, has shown remarkable application potential in financial fraud detection. As pointed out by Hastie et al. (2009), machine learning algorithms can automatically learn the key features and patterns of fraudulent behaviours from massive data by training models, thus achieving effective recognition and early warning of potential fraud risks. This characteristic significantly enhances the accuracy and efficiency of detection, making machine learning play an increasingly greater role in financial fraud detection (Perols, 2011; Bao et al., 2020). What's more, machine learning can develop more precise fraud prediction models that benefit from its powerful generalisation ability, which makes it widely and flexibly applied in corporate financial fraud detection. Traditional supervised learning algorithms, such as the random forest and support vector machine, have been shown to perform well in predicting financial fraudulent practices (Purda and Skillicorn, 2015; Chen, 2020; Xu et al., 2022). With the continuous development of algorithms and the introduction of advanced technique, machine learning is gradually being combined with technologies like deep learning, big data analysis, and natural language processing. This provides us with detection models that can better adapt to complex and changeable cheating means and scenarios, and further enables comprehensive and in-depth analysis and prediction of corporate fraud, thereby contributing to more accurately warn of fraud activities.

In addition to adopting advanced methods and tools, an effective fraud prediction model needs multiple streams of information, which is critical for improving the performance of enterprise fraud detection. The selection of input variables has always been regarded as a core sector in the research of fraud prediction. Previous studies on corporate fraud detection mainly focused on modelling financial fraud risk using public financial data. Wei et al. (2017) and Bao et al. (2020) showed that using original financial variables extracted directly from financial statements instead of financial ratios to establish a model can improve the effectiveness of financial fraud detection to a certain extent. However, relying solely on financial data as the input variable often fails to fully reveal the complexity and diversity of corporate violations, and it also partly limits the predictive performance of the model. According to the fraud triangle theory, non-financial variables such as firm characteristics, executive characteristics, and corporate governance have been proven to produce a non-negligible effect in predicting corporate financial misconduct (Chen et al., 2016; Wu et al., 2016). Bertomeu et al. (2021) found that the importance of non-financial variables was almost twice that of financial variables when building a prediction model for financial misreporting. Along with the depth of research and the diversification of information channels, scholars have attempted to integrate relevant information from more dimensions into the fraud prediction model, including, but not limited to, textual factors (Hajek and Henriques, 2017), corporate culture (Biggerstaff et al., 2015; Liu, 2016), industry trends (Wang and Winton, 2021), and macroeconomic environment (Armstrong et al., 2012; Cumming et al., 2018a). These multi-resource information can provide more abundant features that helps the model recognize corporate fraud risks more precisely. But these non-financial factors have not been comprehensively and thoroughly investigated in the literature on using machine learning models for fraud detection.

This study focuses on predicting the fraud propensity of Chinese listed companies, while the China's capital market is chosen as the research background mainly for several reasons. Firstly, as one of the largest economies in the world, China ranks among the top in terms of capital market size and the number of quoted companies. At the same time, the high-speed development of Chinese market has led to prevalent and rampant corporate misconduct, especially the fraudulent motives and methods exhibit diversity and complexity among listed firms (Jiang et al., 2010; Conyon and He, 2016). In this context, the fraud of Chinese listed companies not only provides us with rich study sample, but also presents a broad space for exploration. In-depth research on these samples can help us analyse the motivations, potential risks, and existing problems of irregularities by Chinese enterprises, and further reveal the underlying reasons and mechanisms. Secondly, Chinese firms are situated in a unique economic, political, and cultural background, and these factors have a complicated and profound impact on their business behaviour and tendency to commit fraud. For example, in the period of economy transformation in China, listed companies, as an important component of national economic system, are faced with challenges brought by intensified market competition and increased regulatory requirements while pursuing benefits; the government's macroscopic economic regulation, industrial policies, and other political factors directly affect the enterprise business strategy and market performance; in addition, the specific traditional cultural background of China, especially the emphasis on collectivism and other values, has had a far-reaching influence on the corporate decision-making and governance mechanisms to some extent. Compared with relatively mature capital markets in developed countries, Chinese financial market has long been highly intervened by the government (Chen et al., 2006). Although the China Securities Regulatory Commission (CSRC) and government agencies are effective in issuing a series of laws and regulations, for instance, the "Code of Corporate Governance for Listed

Companies in China" which revised in 2018[1], in an effort to standardize market order, there are still noticeable problems such as incomplete protection mechanism, inadequate regulatory enforcement, and lack of information transparency (Hou and Moore, 2010; Wu et al., 2016). These issues have resulted in the widespread existence of corporate violations in recent years, which highlights the fact that corporate fraud detection remains an urgent and momentous problem to be solved in China. Thirdly, while machine learning technology has made impressive progress in the field of financial fraud detection, existing research mainly focuses on the capital markets of developed countries (Bao et al., 2020; Brown et al., 2020). Even with a few studies targeting the Chinese market, the relevant research is concentrated on financial fraud detection (Xu et al., 2022; Achakzai and Peng, 2023). The applicability and effectiveness of these advanced technologies in corporate fraud prediction is still an area to be further explored in emerging markets with relatively weak supervisory regime, especially in those with rapidly growing economy but lacking sound regulatory system. As one of the most representative developing countries in the world, the fraud problem faced by Chinese enterprises is not only unique but also reflective. Therefore, by using machine learning to predict the possibility of enterprises engaging in violations and to explore the motives and influencing factors behind the fraudulent behaviours, we can gain a more comprehensive understanding of the financial environment, supervisory system, and characteristics of irregularities in this emerging market, which in turn can provide experience and reference for other emerging economies. Last but not least, the availability and quality of data on Chinese listed companies is also a significant reason for choosing the Chinese market for this study. The financial data, market data and other related information of public companies in China have

[1] Code of Corporate Governance for Listed Companies in China, published by the China Securities Regulatory Commission and effective September 30th, 2018. https://www.gov.cn/gongbao/content/2019/content_5363087.htm

been greatly enriched and improved with the development of information technology. This change offers more extensive and high-quality data resources, and allows us to fully utilize these financial and non-financial information to build the fraud detection model so as to capture wrongdoings in listed firms more precisely.

We first conduct a comprehensive analysis of the causes of illegal behaviours by enterprises. Based on the GONE framework proposed by Bologna et al. (1993), we construct a set of 86 input variables in total from multiple dimensions such as information disclosure, ownership structure, and financial indices to thoroughly portray the potential risk of violation, and aim to provide a basis for building a well-performing fraud prediction model. Then, in order to explore the feasibility and validity of machine learning in detecting corporate misconduct in the Chinese market, two widely used and representative ensemble learning algorithms, namely the random forest (RF) model and the gradient boosted decision tree (GBDT) model, are selected to predict enterprise's tendency to violate regulations. The results show that the GBDT model, which is a boosting classifier, outperforms the RF model adapting the bagging idea in several performance evaluation metrics, especially in the F1-score and the AUC value (area under the ROC curve), GBDT has significant improvement. This result indicates that GBDT exhibits higher prediction performance while having advantages in avoiding false-positive and false-negative errors. Considering that different types of corporate irregularities are closely related to a variety of subjects and motives of fraud, and that the type of violation may affect the prediction ability of the model, we subdivide fraud cases into accounting irregularities, operation irregularities, and leadership irregularities, and re-evaluate the forecasting performance of RF and GBDT in various types of fraud, so as to fully assess the application potential of machine learning techniques for corporate fraud detection in China. Our results demonstrate that the performance of the GBDT

model is superior to the RF model in predicting these different types of fraud, especially in predicting accounting fraudulent behaviours. This finding further strengthens the feasibility of applying machine learning methods to detect corporate fraudulent activities in the Chinese market. In addition, for the purpose of more efficiently identifying the signal of violations by listed companies, we not only focus on the improvement of model accuracy, but also delve into the factors that play a key role in predicting corporate fraud through ranking the importance of input variables. We observe that *Exposure* variables occupies the significant status in corporate fraud prediction, followed by *Ownership Structure* variables, *Environmental* variables and *Financial Index* variables. This discovery emphasises the importance and necessity of including non-financial factors to enrich the input variables of the model, thereby enhancing the accuracy and reliability of forecasts. For predicting different types of wrongdoings, *Exposure* variables consistently play a more crucial part, while *Financial Index* variables and *Ownership Structure* variables reveal more prominent importance in forecasting operation violations and leadership violations, respectively. These results enhance the rigorous and scientific of this study, and also provide reliable references for developing more precise and targeted regulatory strategies. Finally, we conduct misconduct prediction for different ownership enterprises and firms with different information disclosure qualities, and also re-examine the predictive ability of the models after adding new *Exposure* variables and after using a replacement sample. The robust results obtained again strongly support for the effective application of machine learning in fraud detection of listed companies in China.

Our study has made valuable contributions in the following aspects. Firstly, we innovatively construct a multi-dimensional framework for corporate fraud prediction, consisting of financial and non-financial factors. This framework not only improves the forecast accuracy, but also

recognises key risk factors, making an incremental contribution to the relevant literature. There has been a wealth of research on corporate violations, but previous studies mainly rely on logistic regression to detect financial fraud and often focus only on single financial or non-financial characteristics, lacking a systematic analysis of the causes of committing fraud in enterprises. In contrast, this paper takes full consideration and integration of information from multiple sources, and utilizes machine learning algorithms, an emerging fraud detection method with great potential, to build the prediction model, which significantly improves the predictive performance of fraud detection and adds new aspects and depth to the literature in this research area. Secondly, this study provides a momentous extension to the research of diversified corporate fraud. We have noticed that most of the previous research on using machine learning to predict corporate misconduct has concentrated on financial fraud, without capturing the features of other types of fraudulent behaviours. To the best of our knowledge, this is the first paper to cover and distinguish the diverse types of corporate irregularities. Different from Xu et al. (2022), who classified cheating into more-serious and less-serious, we broaden our fraud sample and, for the first time, comprehensively predict and analyse the overall violation and various types of violations including accounting violations, adding new research content to the application of machine learning in the field of enterprise fraud detection. Meanwhile, we also carry out a detailed study and comparison around the relative importance of input variables in predicting different types of corporate fraud, which provides us with fresh insights into the practical application. This analysis not only contributes to reveal the causations and mechanisms of a variety of deception risks, but also lays a solid foundation for advancing the precision and reliability of predicting violations, thus providing strong theoretical support and practical guidance for timely preventing and controlling fraud in all forms.

Moreover, this study makes some contributions to the field of corporate fraud detection in the Chinese market. Currently, the research using machine learning to predict the tendency of enterprises to commit wrongdoings principally focus on capital markets in developed countries, while relatively few studies have been conducted for emerging markets. Our results demonstrate that machine learning models can effectively predict the illegal behaviour of Chinese listed companies, and also explore the inherent properties and mechanisms of fraudulent activities in depth, affording fundamental basis for building the more efficient and accurate prediction model. This provides supervision agencies and investors in the China's capital market with more comprehensive information and more powerful fraud detection tools to help them better identify and deal with potential fraud risks, thereby promoting the development of related research. The Chinese market is still facing challenges such as the incomplete protection mechanism, insufficient regulatory enforcement, and information opacity (Chen et al., 2006; Hou and Moore, 2010; Wu et al., 2016), and many developing economies are also confronted with similar problems. Therefore, it is of great practical significance to investigate the generalization of machine learning techniques in predicting fraudulent behaviours of Chinese enterprises, which not only aims to enhance the capability of corporate fraud detection in China, but also offers valuable experience for other emerging markets, especially for economies with high economic growth and weak institutional environment, in terms of early warning and detection strategy of corporate irregularities. Last but not least, our findings reflect theoretical depth and practice-oriented for effective warning and crackdown on corporate misconduct, and have significant implications for listed companies, regulators, investors and other market entities. For enterprises, this study emphasises the consequence of strengthening corporate governance structure and perfecting internal control mechanism. Businesses should pay attention to the quality of information disclosure, ensure the

authenticity, accuracy and integrity of information, so as to prevent and reduce the risk of non-compliance, which in turn promotes compliant operations and sets up a good enterprise image, thus attracting more investment and fund. For regulatory agencies, this paper highlights the prominent value of machine learning technology in corporate fraud detection, which enables them to develop more precise and targeted supervising strategies, thereby improving the efficiency and effectiveness of regulation, curbing the occurrence of violations, protecting the interests of vast investors, guaranteeing the fairness and prosperity of the market, and driving the stability and high-quality development of economics. At the same time, it will also help investors make more informed investment decisions, raise their confidence, and promote healthy competition and orderly operation of the capital market.

The rest of this paper is organised as follows. Section 2 details the research design, and gives an introduction to the input variables, machine learning models used in this study, as well as evaluation metrics. Section 3 mainly discusses the empirical results and robustness tests, and concludes in Section 4.

## 2. Data and Research Design

### 2.1 Data and Sample

The fraud data in this paper comes from the China Stock Market and Accounting Research (CSMAR) database, which comprehensively collects and summarises violation announcements issued by the CSRC, Shenzhen Stock Exchange, Shanghai Stock Exchange, Ministry of Finance, listed companies, and other organisations, and is therefore widely used in studies related to corporate misconduct in China (Chen et al., 2006; Firth et al., 2011; Song et al., 2014; Conyon and

He, 2016; Xu et al., 2022; among many others). This database provides detailed information on the occurrence time, violation types, and punishment measures for breaches included in each violation announcement. We construct the fraud variable based on the year in which the firm actually committed the irregularity, rather than the declared date. Specifically, if a public company engages in wrongdoings in year $t$ and the event is exposed by the CSRC or other regulatory authorities in year $t + n$, then year $t$ is set as the violation year that we need. This study uses a dummy variable *Fraud* as a proxy for whether an enterprise has committed fraud, which is assigned a value of 1 if the firm has taken place any type of violation during the year, and vice versa.

We collect violation announcements released by the CSRC and other agencies between January 1, 2000, and December 31, 2023, from which all Chinese A-share listed firms that have engaged in fraudulent behaviour between 2000 and 2022 are selected. Panel A of Table 1 shows the annual distribution of offending enterprises. On the timeline of the first decade (i.e., 2000 to 2010), with 2001 as the cut-off point, the proportion of non-compliant companies to total firms peaked in 2001 and gradually declined thereafter. This trend is largely synchronized with the evolution of regulatory regime of China's capital market in the new era. The Chinese securities market has only established a centralised and unified supervision system since 2000, marking the beginning of a more standardised and mature stage for China's capital market. With the establishment and continuous improvement of the regulatory system, the number and proportion of illegal enterprises decreased significantly after 2001, which proves the positive effect of the strengthened supervisory system in reducing irregularities. In addition, we observe a remarkable upward trend in the number of offending parties since 2007. This may have something to do with the market environment at that time. In 2007, China was currently in a period of rapid economic development, with a notable increase in the capital market activity, and investors' optimistic

expectations and speculative enthusiasm for the stock market reached a peak. Against this backdrop, enterprises faced greater market competition and performance pressure despite the existence of stringent regulations. Therefore, some companies had strong incentives to resort to illegal means to rapidly expand or gain more market share, or to cater for market speculation in order to obtain higher share prices and market values.

**[Table 1]**

In the second decade, starting from 2011, the proportion of fraudulent firms increased observably, exceeding 20% every year. This is closely correlated with the development of China's financial market. First of all, China's economy is at a critical period of transformation, where the adjustment of industrial structure and changes in market competitive patterns have put enormous operational pressure and survival challenges on both traditional and emerging industries, making it possible for enterprises to take illegal actions to maintain survival and pursue higher profits in response to market changes. Second, the rapid growth and innovation of the financial market have brought about a double-edged sword effect. The rise of Internet finance and other new industry forms, as well as the constant emergence of a variety of financial products and trading methods, have provided firms with more financing channels and investment opportunities, but at the same time, they have also brought higher risks. For example, some companies can take advantage of the opacity of the financial market to conduct illegal operations such as insider trading, or utilise sophisticated financial derivatives for arbitrage or regulatory evasion, leading to an increase in the number of fraud firms. Besides, although regulators have made great efforts to strengthen oversight, there are still certain gaps or loopholes in China's supervision system, which, coupled with the deficiencies in the number and capacity of regulatory authorities and relevant personnel, have resulted in the inability of regulatory efforts to fully cover the financial market, which has also led

to a raising in violations. In particular, the number and proportion of violating companies increased sharply from 2015 onwards. This is first and foremost related to the adaptation and monitoring challenges of the Chinese initial public offering (IPO) market after its re-opening in 2014. On the one hand, some enterprises may face enormous pressure on their operation in order to meet the listing requirements, thereby increasing the possibility of financial malpractice occurring (Zhou et al., 2021; Achakzai and Peng, 2023); on the other hand, the new regulations and oversight requirements have produced some adjustments to the enforcement of regulators, which can lead to insufficient supervision of violations. Moreover, the release of a series of policies, such as the "Guidelines for the Articles of Association of Listed Companies"[2] (revised in 2014, 2016, and subsequent years) and "Several Opinions on Further Promoting the Healthy Development of the Capital Market"[3], has promoted the enhancement of corporate governance and the further improvement of the supervision of China's capital market, thus reinforcing the disclosure intensity of, and the crackdown against, corporate fraud, which in turn results in an increase in the number of illegal enterprises. Among them, the number of non-compliant companies between 2019 and 2022 has instead shown a decline. This is firstly due to the synthetic action of various factors, such as the progress of regulatory technology and application, the enhancement of media monitoring and public participation, the market environment has been further improved, and the number of offending parties has been effectively controlled. Secondly, the outbreak of the COVID-19 epidemic had a severe impact on the global economy, especially the closure and other measures taken by the Chinese government for the sake of people's lives and safety, making China's capital

[2] Guidelines for the Articles of Association of Listed Companies, published by the China Securities Regulatory Commission and revised on October 20th, 2014. http://www.csrc.gov.cn/csrc/c101864/c1024608/content.shtml

[3] Several Opinions on Further Promoting the Healthy Development of the Capital Market, published by the State Council and effective on May 9th, 2014. https://www.gov.cn/zhengce/zhengceku/2014-05/09/content_8798.htm

market suffer a serious influence and its economy in a state of stagnation. In a situation where it is difficult to carry out business, the amount of irregularities, naturally, decreased significantly. Thirdly, fraud detection tends to be delayed. Because of information asymmetry and restrictions imposed on law enforcement agencies, many infractions cannot be detected in time, so fraudulent activities in recent years were incompletely covered.

Since corporate violations involve a variety of actors and the complexity of their fraudulent motives, in order to conduct a more comprehensive and in-depth study of the diversity of corporate misbehaviours, this paper, with reference to the authoritative definitions of corporate fraud provided by the CSRC and the CSMAR database, classifies listed companies' wrongdoings into the following four categories: accounting irregularities, operation irregularities, leadership irregularities, and other types of irregularities. Specifically, accounting malpractices are mainly related to the processing and disclosure of financial information, including but not limited to false financial information disclosure, significant omission of information disclosure, and general improper accounting treatment, etc., and these behaviours directly undermine the investors' right to know and the market fairness; operation violations are correlated with the compliance of daily business activities, mainly including the appropriation of corporate assets, illegal guarantees, and illegal production or sales, etc., and these behaviours not only damage the company's reputation and interests, but may also cause adverse effects on society; Leadership misconducts are mainly concerned with the behaviours of senior management personnel, including illegal trading of shares, embezzlement, and other misbehaviours that are in serious breach of the leaders' responsibilities and principle of good faith, causing huge losses to the enterprise and its shareholders; and other types of violations cover all breaches in addition to the three categories mentioned above. We introduce dummy variables *Accounting_Fraud*, *Operation_Fraud*, *Leadership_Fraud*, and

*Other_Fraud* to respectively represent whether accounting violations, operation violations, leadership violations, and other types of violations occur; namely, if the enterprise has engaged in a certain type of violation in that year, the corresponding dummy variable is assigned a value of 1, otherwise it is assigned a value of 0. The definitions of these variables are detailed in the Panel A of Appendix Table 1. Panel B of Table 1 displays the annual distribution of the different types of irregularities that occurred in Chinese A-share public companies from 2000 to 2022. It should be particularly emphasised that each violation announcement may not only involve a single infraction but contain multiple illegal actions. In other words, one listed company may commit multiple wrongdoings within the same fiscal year for different reasons or motivations. Hence, comparing the data on different types of breaches in Panel B with the data in Panel A, we find that the aggregate number of violations presented in the former and the number of firms involved are beyond the statistics of the latter.

It is clear from Panel B that accounting irregularities constitute the bulk of corporate fraud. Statistics show that on average, up to about 76% of offending companies engage in accounting malpractice each year, and the most frequent categories of these misbehaviours include deferred disclosure, material omissions, and false recordation (misleading statements). This is bound up with the current development stage and regulatory environment of the Chinese securities market. With the growing maturity of the stock market, the inadequacies in internal control and governance mechanisms of enterprises have brought to the forefront the challenges and problems they face in terms of information disclosure. This trend, on the other hand, reveals the major responsibility that regulators bear in disclosure supervision. The increasingly stringent rules and strengthened penalties imposed by supervision agencies on corporate disclosure have made it impossible for any attempt to seek improper benefits by means of accounting violations. As a result, most of the

existing literature focuses on the prediction of corporate financial fraud (for example, Chen, 2020; Huang et al., 2020; Xu et al., 2022). Although the number of financial misconducts dominates, operation irregularities and leadership irregularities cannot be ignored as well. According to statistics, about 37% of the offending enterprises are involved in these two types of fraudulent activities every year on average. Among them, illegal trading of shares and appropriation of corporate assets are the categories with the most occurrences of these two types of violations, respectively. It is worth noting that prior to 2007, the number of operation breaches was observably higher than that of leadership misbehaviours; however, the number of the latter showed a remarkable upward trend from 2007 onwards and surpassed the number of the former in 2008. This change is not coincidental, but is related to multiple reasons. Firstly, regulatory authorities have strengthened their oversight of corporate governance mechanisms after going through the initial stage of standardised construction in the Chinese market. With constant improvement of regulatory policies and strengthening of enforcement efforts, the awareness of corporate compliance has been notably enhanced, and meanwhile, the agency problem in enterprises has gradually reformed, which has to some extent curbed operational violations at the company level. Secondly, with the expansion of firm scale and the complexity of business, the decision-making power and influence of leaders in the operation process have been growing. Some directors may take advantage of their position and power for personal gain, carry out illegal operations such as insider trading and embezzle public funds. Furthermore, the social attention received by company leadership has signally increased with the rise of media monitoring and public opinion, and the public's supervision over executive behaviour has gradually enhanced, resulting in more exposure to irregularities at the leadership level. To sum up, despite the fact that the number of operation violations and leadership violations might be less than that of accounting fraud, their harmfulness

and adverse effects are equally serious, which will impair the reputation and revenues of the enterprise and also badly disrupt the fairness of the market. Therefore, these two types of corporate fraud deserve special focus and in-depth research.

Karpoff et al. (2017) pointed out that the inability to observe the full population of fraudulent activity is a particularly common type of error in identifying corporate wrongdoings. To address this issue as much as possible, we take the following measures in selecting the breach sample to reduce bias caused by undetected fraud. First of all, the sample time span of this study is from 2000 to 2022 to ensure the adequacy of fraud data. This broad time horizon helps us capture changes in corporate infractions under different market environments, thus improving the entireness and correctness of our research. Next, when setting up the fraud proxies, we only consider the actual occurrence of breaches by the company in a certain year, not the specific number of fraudulent activities. We are able to focus more on recognizing those firms that do indeed commit fraud in this way, reducing the impact of omitted events on the results, and enhancing the reliability of this study. Additionally, due to the inherent delay in fraud detection, we specifically exclude observations of non-offending companies in 2022 to further improve the sample's accuracy. All the data for financial and non-financial variables comes from the CSMAR database, and after data assignment and other processing, a total of 86 multi-dimensional variables are obtained as input variables. Since our research objective is to forecast a company's tendency to engage in illegal activities in the next year, all of the input variables are lagged by one period. After combining all the variables used in this study, a sample containing 52,459 firm-year observations is ultimately obtained. It should be noted that we refrain from doing any operations during processing that could alter the initial state of the data, such as applying Winsorize to continuous variables, in order to guarantee the authenticity of sample data. From the descriptive

statistics of the input variables presented in Table 2, we can find that the standard deviation of individual variables (e.g., *CapEx* and *MA_Intensity*) is relatively large, which is probable because of the presence of certain noise or outliers in the raw data. However, given that the actual situation and real data are unknown, we do not intervene with these variables to ensure the integrity and authenticity of the data. What's more, there are a certain number of missing values in the sample, and we adopt two common approaches to solve this problem. We first eliminate all observations with missing information and end up with a sample of 32,227 firm-year observations, which we refer to as the Type I sample; we next fill in the blanks using the mean of continuous variables or the median of dummy variables, and finally obtain a sample containing 52,459 firm-year observations, which we refer to as the Type II sample. The basic statistics of these two samples are displayed in Panel A and Panel B of Appendix Table 2, respectively. It can be found through comparison that the data distribution of the Type I sample is more reasonable, and the excessive standard deviation problem mentioned above can be effectively mitigated, but this method inevitably loses a large amount of information about input variables, which may not fully reflect the overall situation and lead to sample selection bias. In contrast, the Type II sample retains more valid information from the original data without changing the overall distribution of the raw data as much as possible, while this method of simply filling in missing values can pose a risk of not conforming to reality and thus introducing erroneous information. Therefore, this study comprehensively takes the research results of these two types of samples into account, and improves the precision and credibility of the study by supporting and complementing each other.

**[Table 2]**

### 2.2 Categorizing Input Variables

When delving into the potential factors affecting the occurrence of corporate misconduct, we should comprehensively analyse the causes of enterprise fraud tendencies. In view of this, this paper reviews the previous literature and draws on the core concept of the GONE framework (Bologua et al., 1993), constructing 86 indicators as input variables from eight dimensions, namely, equity structure, corporate governance, personality characteristics, information disclosure, financial factors, the environment and industry in which companies are located, and other aspects, in order to more accurately portray the potential risks of corporate fraudulent activities.

First of all, *Greed* plays an undeniable role as the intrinsic driving force behind corporate irregularities. A greedy mentality often drives individuals or organisations to ignore or even violate legal and ethical constraints in the pursuit of personal interests. In China, the highest operational decision-making authority of listed firms is usually in the hands of the chairman rather than the CEO (Chen et al., 2011; Feng and Johansson, 2018), which is somewhat different from countries such as the United States or Europe. So we select the age variables (*Chair_Age* and *Elder*) and the gender variable (*Chair_Gender*) of the chairman (Benmelech and Frydman, 2015; Feng and Johansson, 2018; Liu, 2018; He et al., 2020), taking into account the characteristics of Chinese enterprises, to explore the possible effects of his or her greediness on the company's ethical orientation. Additionally, the governance and management layers of public firms, including directors, supervisors, and executives, not only control the day-to-day operations and oversight of the enterprise, but are also responsible for making major decisions. The extreme desire and excessive pursuit of personal profits by these individuals tend to have a significant impact on corporate fraud risk. To more thoroughly capture the influence of greedy mentality on malpractices, we focus on the average age of directors, supervisors, and executives (*Average_Age*) and their educational and working backgrounds (*Fin_Back* and *Overseas_Back*), and also measure the

reasonableness of compensation design (*Salary* and *Salary_Gap*) and its ability to effectively facilitate the alignment of interests between them and shareholders (Firth et al., 2011; Haß et al., 2015; Li et al., 2016; Clifford et al., 2018).

Secondly, the incidence of corporate violations is frequently inextricably linked to the *Opportunity* to carry out misbehaviour. At the level of ownership structure, its rationality and integrity have a profound impact on corporate governance mechanisms and business decisions; while an unreasonable equity structure often provides space and conditions for internal personnel to engage in fraudulent activities. Therefore, we choose a total of 11 variables, including the shareholders' holdings (*Largest_Own*, *Herfindahl_5*, *Control_Own*, *Control_Dummy*, and *Central*) and the insiders' shareholdings (*Board_Own*, *Chair_Own*, *Employee_Own*, *Insider_Board*, *Insider_Spvisor*, and *Man_Own*), to comprehensively measure the properties of a company's ownership structure (He and Luo, 2018; Shi et al., 2020; Ning et al., 2021; Xiong et al., 2021; Yao et al., 2021). Besides, the actual controller plays a crucial role in the breach risk, as they may seize private gains through the controlling chain, thereby causing damage to firm value and shareholder interests. We employ the separation of two rights (*Differ*) to reflect the possibility of the actual controller obtaining private benefits through the controlling chain so as to quantify this risk (Yao et al., 2021). Meanwhile, we also consider their other features, such as foreign residency (*Foreign_Resid*) and control rights of multiple companies (*Multi_Control*), which can increase the risk of transferring property or circumventing legal sanctions (Chen et al., 2018). Besides, as an important external monitoring force, institutional investors' shareholding has an equally remarkable influence on the occurring of corporate infractions. We adopt two variables, the shareholding ratio of institutional investors (*Inst_Own*) and the shareholding situation of qualified foreign institutional investors (*QFII*), to investigate the external supervision role of

institutional investors in corporation governance (Liu, 2016; Harford et al., 2018; He et al., 2020; Gam et al., 2021).

In addition to the ownership structure, *Opportunity* factors also emphasise the effectiveness of an enterprise's internal administrative systems. A valid internal governance mechanism can ameliorate a good operating environment, reduce potential risks, and thus diminish the probability of non-compliance. We first introduce the ratio of shareholding between the top 2–5 major shareholders and the largest shareholder (*Balance*) to represent the balanced relationship among stockholders, so as to study the impact of equity balance structure on listed company governance (Mackenzie et al., 2013; Su et al., 2021). The board of directors occupies a core status in corporate government, and its composition and functioning have a notable effect on its administrative outcomes. Hence, five variables are selected as important indicators to measure the constitution and operational effectiveness of the board of directors (*Board_Size*, *Even_Board*, *Duality*, *Female_Board*, and *Gender_Diverse*), which can reveal the communication efficiency and decision quality within the board (Khanna et al., 2015; He and Luo, 2018; Liu, 2018; He et al., 2020; Sun et al., 2021a). Moreover, independent directors play a vital role in curbing the opportunistic behaviour of management. So we construct four variables (*Board_Ind*, *Qualified_IndBoard*, *Strong_Board*, and *Centrality*) to weigh the validity of board independence (Wang and Hsu, 2013; Kuang and Lee, 2017; Ullah et al., 2021). Apart from the above variables, we use the number of board meetings (*Board_Meeting*) and the number of general meetings (*Shareholder_Meeting*) to judge the active degree of internal governance (Shi et al., 2020; Gam et al., 2021; Su et al., 2021), and employ the establishment of audit committees (*Committee*) and the characteristics of the supervisory board (*Spvisor*) at the same time to show the firm's rigour in decision-making oversight and its transparency in terms of information disclosure (Chen et al.,

2018; Cumming et al., 2018a; Zaman et al., 2021). These factors are key positive signals released by enterprises wishing to demonstrate their good corporate governance.

Furthermore, *Need* is the direct cause of fraudulent behaviour, particularly when it comes to motives that are closely related to financial needs. Enterprises often encounter a variety of challenges during the operating process, such as performance pressure and fierce market competition. These pressures may force the management to generate impulses to enhance corporate performance or seek personal gain through illegal means, which in turn fuels the growth of fraud. In order to appraise a company's economic and operational conditions in a more comprehensive way, we select a total of 13 financial indicators from five aspects, which cover key areas such as solvency, profitability, market value, investment and research and development (R&D) activities, and financial risk. Specifically, in terms of debt paying ability, we use four variables that can reflect the short- and long-term solvency of the enterprise (*Cash*, *Leverage*, *Liquidity*, and *Tangibility*) to represent the financial stresses it faces (Ananchotikul et al., 2010; Kizilaslan and Man akyan Mathers, 2014; Byun et al., 2019; Zaman et al., 2021); for profitability, we adopt two variables (*EBIT* and *Growth*) that can measure the firm's operating status to denote its earnings level and market expansion ability (Khanna et al., 2015; Li et al., 2016; Kuang and Lee, 2017); in the aspect of market value, we choose two variables (*MTB* and *Q*) that showing the company's valuation in the market so as to reveal its profit potential and growth space (Yu and Yu, 2011; Haß et al., 2015; Chen et al., 2018); and for investment and R&D activities, we construct five relevant variables (*CapEx*, *External_Financing*, *MA_Intensity*, *R&D*, and *R&D_Dummy*) to disclose the strategic layout and development direction of the enterprise (Liu, 2016; Cline et al., 2018; Shi et al., 2020; Hagendorff et al., 2021; Wang and Winton, 2021); and another variable (*ZScore*) is used to measure the financial risks faced by enterprises (Li et al., 2016; Liu, 2018). It

should be noted that in existing research, the proportion of R&D expenditures to total assets is usually used as the measurement of corporate R&D activities. The direct use of R&D expenditures would, however, result in a significant quantity of missing data because complete statistics on R&D expenditures have only been documented since 2007, while the sample time span of this study is relatively large. To address this issue, we introduce a dummy variable *R&D_Dummy* to estimate the firm's R&D activities. This variable is able to reflect whether the enterprise has conducted R&D activities, although the intensity cannot be exactly measured. In the subsequent robustness tests, we further apply a continuous variable *R&D* to represent the R&D investment to ensure the accuracy and reliability of the research results.

Finally, *Exposure* is an important dimension for assessing whether a company's misbehaviour can be detected in time and punished accordingly. If the likelihood of fraudulent irregularities being discovered is low, or they will not receive severe punishment even if they are found, then the incentives for firms to violate regulations tend to be strengthened, thereby raising the fraud risk. Therefore, in order to provide an overall evaluation of the degree of exposure to malpractices, we select six key variables for measurement. First, the analyst coverage (*Analyst*) can reflect the level of attention paid by financial analysts to corporate behaviour, which directly affects the probability of wrongdoings being discovered (Hagendorff et al., 2021; Sun et al., 2021a; Xiong et al., 2021); we then consider the choice of audit authorities, using three variables–whether to hire one of the Big 4 international accounting firms or Big 10 domestic accounting firms (*Big4* and *Big10*), and whether to hire a foreign accounting firm (*Foreign_Audit*)–to measure the audit quality and the intensity of fraud monitoring (Firth et al., 2011; He et al., 2020; Shi et al., 2020); we also utilize whether the financial reporting has received a standard audit opinion (*Opinion*) and the assessment results of the company's disclosure quality (*Disclosure*) to further evaluate the

possibility of corporate fraud being exposed (Ning et al., 2021; Su et al., 2021). In addition to the traditional oversight by analysts and auditors, the media plays an increasingly crucial external monitoring role in the modern corporate governance system, and, in recent years, textual data, which provides qualitative information about strategic planning, market reactions, etc., has been widely used in the study of corporate non-compliance. Hence, we introduce two media-related variables (*News* and *Media_Tone*) to express the media's attention and supervision of enterprise behaviour (Davidson et al., 2015; Li et al., 2021b; Heese et al., 2022) and make full use of these multi-channel data. We simultaneously apply two textual feature variables relevant to the MD&A section (*Similarity* and *Emotion_Tone*) to capture potential fraud signals that might be hidden by management in their reports (Purda and Skillicorn, 2015).

Under the GONE framework, apart from the variables concerned with the four core factors mentioned above, this study also constructs a series of indicators from the external environment and industry characteristics of enterprises. Given the strong link between the regional economic development level and the corporate business circumstances, as well as the significant impact on the propensity to engage in violations, we select three variables based on the provincial data to measure the regional economy where the enterprise is located (*GDP_Growth*, *Personal_Income*, and *Unemployment*) (Wang and Hsu, 2013; Hagendorff et al., 2021; Heese et al., 2022). These variables can show the economic vitality and labour market conditions of the province in which the company is located, thereby implying the fraud risks that the firm may come with. Also, considering the potential threat of economic fluctuations to business survival, especially in extreme situations such as economic depression or pandemics like COVID-19, the motivations for enterprises to commit financial infractions or other irregularities may be enhanced in the face of financial distress. Karpoff (2021) pointed out that the COVID-19 pandemic has caused an

enormous shock to the global economy, and the incidence of fraudulent behaviour is likely to rise markedly in the coming years. So we employ two variables (*Crisis* and *COVID19*) to quantify the potential effects of these external economic pressures on firms' tendency to violate rules (Shi et al., 2020; Hagendorff et al., 2021; Karpoff, 2021). Moreover, Doidge et al. (2007) argued that in developing countries, country-level characteristics are often more important than firm-level mechanisms in solving corporate governance issues, which has far-reaching implications for corporate fraud. In light of this, we introduce a variable Anti_Corrupt, using the convening of the 18th National Congress of the Communist Party of China in late 2012 as the starting point of the anti-corruption campaign (Li et al., 2016; Lin et al., 2016; Su et al., 2021). As a vital exogenous variable of governance mechanisms and endogenous regulatory shock (Su et al., 2021), this variable helps us to explore how improvements in the political environment affect the fraudulent activities of enterprises.

At the industry level, there are remarkable differences in the business models, regulatory environments, and other aspects among various industries, which directly lead to obvious differences in the types and degrees of risk faced by each industry. To probe into the specific manifestation of industry effects in the illegal activities of individual enterprises (Wang and Winton, 2021), we use three variables to clearly define the industry category to which the enterprises belong (*Finance*, *High_Tech*, and *Service_Trade*), aiming to reveal the latent differences and features of the breaches of enterprises in the context of various industries (Liu, 2016; Kuang and Lee, 2017; Hagendorff et al., 2021). Certainly, the degree of industry clustering and competing, as an important part of market circumstances, has a notable influence on the cooperative and competitive relationships between firms, as well as on the violation tendency. To assess this impact, we employ two variables, *Aggregation* and *HHI*, to measure the concentration

and competing situation within the industry (Gelman et al., 2021; Ismail et al., 2021; Wang and Winton, 2021). Additionally, Wang et al. (2010) found that investors' beliefs about industry prospects significantly affect the incidence of corporate misbehaviour within that industry. According to this finding, we further apply three variables to evaluate the operational performance of companies in different industries (*Adj_ROA*, *IndustryQ*, and *IndustryQ_sq*), which represent the profitability of firms as well as the market expectation for their future growth potential (Khanna et al., 2015; Kuang and Lee, 2017; Cline et al., 2018; Shi et al., 2020).

Lastly, we include a set of key factors that impact corporate fraud tendencies but are difficult to classify explicitly under the GONE framework in the *Other* category, following Xu et al. (2022). To be specific, we begin with the basic profile of the company by selecting the following six variables: its age (*Firm_Age*), its scale (*Size*), whether it is a family business (*Family*), whether it is state-owned (*SOE*), its number of employees (*Employee*), and whether it is at risk of litigation (*Litigation*) (Firth et al., 2011; Li et al., 2016; Cline et al., 2018; Gam et al., 2021; Heese et al., 2022). These factors can demonstrate the enterprise's background, business scale, and legal risk, which makes them useful sources of information for figuring out how likely it is to commit fraud. What's more, in view of the crucial position of the stock market, we specifically introduce two variables related to corporate stock (*Stock_Return* and *Volume*) to measure the company's stock yield and market activity (Cline et al., 2018; Harford et al., 2018; Su et al., 2021), and then investigate whether there is a close correlation between these factors and corporate wrongdoings. To further expose the dynamic characteristics of corporate fraud, we also define a dummy variable, *Past_Fraud*, to indicate whether a firm has initiated any irregularities in the past two years (Ananchotikul et al., 2010; Byun et al., 2019). This helps us study whether the existence of past fraudulent activities has a persistent effect on the company's propensity to breach. Last but not

least, although variables such as *Crisis* and *IndustryQ* already incorporate time and industry impacts to some extent, we construct two fixed effects variables (*Province_Year* and *Industry_Year*) to control more precisely for province and industry shocks (Bertrand and Mullainathan, 2003; Davidson et al., 2015; Liu, 2016), thus ensuring that our findings are robust and reliable.

As mentioned earlier, this study aims to provide an early warning mechanism for investors and regulatory agencies regarding enterprises' tendency to commit violations. Therefore, we use these input variables of from current year to predict the likelihood that a firm will carry out fraudulent behaviour in the following year. The definitions of all the above input variables are detailed in Appendix Table 1, and the descriptive statistics of each variable can be found in Table 2. It should be noted that for ease of differentiation and understanding, we reserve statistics for continuous variables to two decimal places, while statistics for dummy variables are reported directly as 0 or 1.

### 2.3 Machine Learning Models

With the ongoing advancements of machine learning technology, a growing number of models have been applied in the field of fraud detection, offering solid technological tools. Numerous studies have amply demonstrated the remarkable effectiveness and precision of machine learning in forecasting corporate financial malpractices (Perols, 2011; Bao et al., 2020; among many others). Machine learning-based fraud prediction models are constantly being improved and upgraded in terms of accuracy and efficiency, given that machine learning algorithms are still being studied and optimised. The main purpose of this study is not to develop new models or adopt state-of-the-art techniques to enhance the ability to detect violations, but to explore the generalisability of machine learning methods for predicting enterprise misconduct in

the emerging market environment, which includes conducting in-depth analysis of different types of corporate irregularities and investigating the multiple factors (both financial and non-financial) that influence these breaches. Therefore, we select two representative machine learning models, namely random forest and gradient-boosted decision tree (GBDT). Both models are built based on decision trees and have been widely validated for their exceptional performance in the corporate fraud prediction field (see, for example, Purda and Skillicorn, 2015; Chen, 2020; Bertomeu et al., 2021). By using these two models, we expect to gain a more comprehensive understanding of the application effects of machine learning methods in the detection of corporate misbehaviour in developing markets and provide forceful support for enterprise risk management.

In this study, all the samples (including offending and non-offending samples) are randomly divided. 70% of the samples are used as the training set for training the machine learning models, and the automatic parameter tuning method, *GridSearchCV*, is employed for the parameter grid optimal search to search for the best-performing parameter combinations. The optimal model obtained is then placed into the test set composed of the remaining 30% of the samples to perform the prediction and evaluation. In comparison to the common practice of dividing datasets by year (Bao et al., 2020; Xu et al., 2022), this random partitioning method can reduce human intervention and distribution deviation, ensuring that both the training and testing sets adequately reflect the overall distribution of data, so that the model is allowed to learn more data features during the training process and improve its generalisation ability. In addition, this study adopts a rich set of factors affecting the likelihood of firms violating regulations as input variables, and two control variables, *Industry_Year* and *Province_Year*, are introduced to account for the potential time effects and capture, to some extent, the impact of time trends on the model's forecast. We also average the results of several trials to obtain the final results, where the combination of randomly

generated training and testing sets can consequently increase the efficiency of data utilisation and further guarantee the robustness of predictive outcomes. By using the above means, we can improve the models' forecast accuracy while also ensuring their validity and reliability.

By combining several single classification models, ensemble learning efficiently extracts and integrates the advantages of each classifier, thereby greatly enhancing prediction and robust performances. In the family of ensemble classifiers, random forest and GBDT are representative methods of bagging and boosting, respectively, with their unique ensemble strategies. The core of the bagging idea (i.e., bootstrap aggregating) is to generate multiple "virtual cubes" of the same size as the original data through random sampling with replacement, then use the same learning algorithm to construct a model on each virtual dataset, and finally combine the forecast results of all the models by means of averaging or voting to determine the ultimate predicting outcome. The theoretical basis of the random forest model is the classification and regression tree (CART). As a frequently-used decision tree algorithm, CART selects the optimal feature by calculating and minimising the Gini index, regards it as a node of the decision tree to partition the sample data, then divides the sub-nodes recursively until a complete CART classification tree is built. Random forest, an aggregation of independent CART classification trees, constructs a variety of CART classification trees based on the sample bootstrap method and summarises the prediction results of each tree. The final result can be obtained by averaging or voting, so as to output a strong classifier with excellent performance. Such a bagging strategy endows the random forest algorithm with remarkable out-of-sample prediction and generalisation capabilities.

Different from random forest, which increases the overall forecast accuracy by building multiple decision trees in parallel, the GBDT (gradient-boosted decision tree) algorithm continuous to optimise the performance of each decision tree through serial iteration and,

eventually, accumulates the results of all trees as the high-precision predicting output. This is connected to the boosting idea utilised by GBDT, the core of which lies in constantly re-sample data, constructing a series of models in a serial manner. Each new model aims to compensate for the deficiencies of the previous one, and these weak classifiers are ultimately combined to form a strong classifier, thus enhancing the whole performance. Regression trees are the weak classifiers in GBDT; they have the property that the value of each node is obtained by splitting a given feature as optimally as possible. And the optimal measure at this time is to minimise the mean squared error (MSE), allowing the most reliable branching evidence to be selected by minimising the MSE. During the training phase, GDBT builds a variety of regression trees along the gradient direction of the loss function, then combines them with certain weights to form a strong classifier. Among them, each tree learns the predicting results of all previous trees, as well as the cumulative error between the predicted and true values (that is, residuals). In other words, the iterative process of GBDT is essentially the course of continuously reducing residuals, which enables GBDT to make full use of the information in the data to achieve extremely high forecasting precision and strong robustness.

### 2.4 Performance Evaluation Metrics

This study aims to probe into corporate fraud prediction and therefore focuses on a main binary classification question, namely distinguishing between offending and non-offending firms. In view of this research objective, we are particularly concerned about the accuracy of identifying violation samples, as this is directly related to the timely detection and prevention of corporate wrongdoing. For this reason, we define as positive ($Y = 1$) the case in which "listed companies will commit irregularities in the next year" as a positive case and as negative ($Y = 0$) the scenario in which "listed companies will not commit any irregularities in the next year". We adopt the

confusion matrix as an important and essential tool for assessing the performance of classification models, which visualises the correspondence between predicted and true values. As shown in Table 3, the confusion matrix consists of four elements that represent the four possible classification outcomes: true positive (*TP*) indicates that the model correctly predicts fraudulent firms; false positive (*FP*) shows that the model incorrectly classifies the non-fraudulent enterprise to be illegal; false negative (*FN*) manifests that offending parties are wrongly recognised by the model to be non-fraudulent; and true negative (*TN*) represents the model's accurate identification of illegal companies.

**[Table 3]**

Explanations of evaluation metrics, which include Accuracy, Precision, F1 score, Matthew's correlation coefficient (MCC) and AUC_ROC, are provided in the Internet Appendix B of this paper.

## 3. Results and Discussion

### 3.1 Performance of Machine Learning Models

Table 4 shows the forecast performance of the two ensemble learning models, RF and GBDT, when using the whole sample of listed companies' violations. The evaluation metrics include the accuracy, precision, recall, F1-score, MCC, and AUC, and the results of both Type I and Type II samples are taken into account, aiming to comprehensively measure the classification ability of these two models, with all metrics averaged. The comparison reveals that the GBDT model demonstrates superior performance to the RF model in most of the key indicators. Taking the Type I sample as an example, the accuracy of RF and GBDT is relatively high, with values of 0.7647

and 0.7810, respectively, indicating that both are effective in distinguishing offending and non-offending samples. In terms of precision, the performance of the RF model is better than that of GBDT (0.7129 vs. 0.6900), implying that about 71% and 69% of the companies predicted to have a tendency to commit fraud by RF and GBDT, respectively, are truly offending parties. While, GBDT performs well in terms of recall, with a significantly higher recall rate (0.5532) than RF (0.4236). This suggests that GBDT is able to correctly identify about 55% of fraudulent companies, that is, it can more effectively recognize the firms that are truly violating and is more advantageous in reducing the underreporting of illegal enterprises. The reason for focusing on precision and recall is that these two important indexes are directly related to false-positive and false-negative errors, respectively, and both errors may, especially in the scenario of corporate fraud detection, lead to serious consequences. Firstly, the false-positive error means that a non-offending firm is wrongly identified as violating. This can require additional manual review to check whether the marked non-compliant enterprises have actually committed wrongdoings, and may also arise improper accusations against innocent companies that damaging their reputations and normal operations, which not only increases the burden on enterprises, but also triggers instability in the capital market. On the other hand, the false-negative error, i.e. incorrectly labelling a breaching firm as non-offending, indicates the omission of real violations, which can lead to serious economic and legal consequences. Specifically, the error cost of underreporting fraudulent companies may include the financial losses suffered by investors as a result of exposure to security risks, as well as the legal liabilities and reputational losses faced by regulators for failing to discover and penalise irregularities in a timely manner. Therefore, it is prominent and of great significance to reduce these two types of errors, with a particular emphasis on avoiding the false-negative error. The GBDT model achieves a high rate of precision, and also has a noticeable

increase in recall compared to the RF model (approximately 13%), suggesting that it not only can identify the violating enterprises more accurately, but also has the ability to cover more real fraudulent activities, thus significantly reducing the risks of false-positive and false-negative errors.

In addition to recall and precision, we provide a comprehensive comparison of models using the evaluation indicators of F1-score, MCC, and AUC. The F1-Score is the harmonic average of precision and recall, serving as a more intuitive metric for weighing these two indexes. The F1-Score value of GBDT (0.6141) is markedly higher than that of RF (0.5314), showing that GBDT can maintain preferable predictive performance while avoiding both false-positive and false-negative errors. Besides, MCC and AUC provide us with valuable references as more general and dependable measures of the model's binary classification ability. From the table, it can be seen that the MCC value of GBDT (0.4693) is greater than that of RF (0.4111), which once again proves that the GBDT model has better prediction performance. The AUC value of GBDT is 0.7195, indicating that the model has a probability of about 72% that the likelihood of a fraudulent enterprise being randomly selected being predicted as fraud is higher than that of a randomly selected non-fraudulent party being predicted as fraud. It is prominently higher than the AUC value of RF (0.6726), demonstrating that GBDT is able to better recognize both offending and non-offending businesses across all classification thresholds, further supporting the excellent performance of GBDT in corporate fraud detection. This study also explores the effect of different samples on the model's classification performance. We find little difference in the predictive performance of the models after comparing the results obtained using the Type I and Type II samples. And this suggests that although the features and limitations of these two types of samples may have an impact on the performance of the models, their predicting outcomes complement each other and jointly illustrate the effectiveness of RF and GBDT in predicting corporate

misconduct, and in particular, ensemble algorithms such as GBDT can provide the powerful tool and optimal detection effect for fraud detection of listed companies.

**[Table 4]**

### 3.2 Importance of Variables in Corporate Fraud Prediction

In predicting the possibility of fraudulent activities by quoted companies, this study constructs 82 input variables from 8 dimensions to capture the potential influences on a firm's propensity to breach across the board. Given the complex correlations that may exist between these influencing factors, it is particularly important to determine which variables play an even more critical role in predicting corporate violations—as recognizing these key variables can assist us in more quickly "targeting" enterprises that might be at risk of committing fraud according to their performance in these areas. Hence, to achieve this goal, we adopt the feature importance ranking of the RF model to quantify the degree to which different variables impact a firm's fraud tendency. This approach characterizes the relative importance of each variable by calculating and ranking its average contribution in trees, and the higher the ranking of variables in the results, the greater their impact on the propensity of enterprises to violate regulations (see Bertomeu et al. (2021)).

Panel A in Table 5 reports the relative importance of each input variable in the whole sample's Type I and Type II samples of all corporate irregularities. Combining the results of these two types of samples, we can first find that the variable *Past_Fraud* comes first and has the most influence. This result supports the significance of improving the accuracy of corporate fraud detection, as companies' illegal actions tend to be persistent and inertial. If a firm's breaches over the past few years have not been detected in time or severely punished, this may reduce the cost of violations to the firm, and its management may take greater chances with the consequences and

risks of wrongdoings, and can even develop a pattern or cultural atmosphere of fraudulent behaviours. So it is necessary to determine those companies that have engaged in violations in the past few years as the key regulatory objects. In addition, *Exposure* variables contribute more to predicting the likelihood of corporate misconduct than other input variables. Among them, *Disclosure* is the second most important variable for corporate fraud prediction in both Type I and Type II samples, and *Opinion* ranks among the top 10 variables in both types of samples, which is sufficient to prove the strong correlation between the quality and transparency of company information disclosure and its fraud risk. Enterprises with better evaluation results of disclosure quality and obtaining standard audit opinions on financial reports send a positive signal to the market that their internal mechanisms are relatively complete, and their financial status and operating achievements are true and reliable; thus, such firms are more likely to comply with laws, regulations, and accounting standards. On the contrary, if a company's disclosure quality is poor or its financial reporting does not receive a standard audit opinion, meaning that the firm more tends to experience problems such as internal control vulnerabilities or mismanagement, which in turn increases the likelihood of violations. It is not surprising that variables related to the corporate conditions such as *EBIT* and *ZScore*, as well as variables related to the ownership structure such as *Herfindahl_5*, *Inst_Own*, *Largest_Own*, and *Control_Own*, all rank among the top 20 in these two types of samples. Companies with high profitability and low financial risk often have greater economic impetuses and more resources to abide by relevant regulations and avoid adverse consequences such as reputation losses that may arise from non-compliance; while a low level of profitability and high financial risk may force businesses to resort to illegal ways like false statements and insider trading to obtain funds due to financial pressures. From the perspective of equity structure, changes in the share proportion of each shareholder will affect the corporate

governance structure, monitoring mechanism, and consistency of interests, thereby significantly influencing the risk of violations, and meanwhile, the external supervision role of institutional investors should not be ignored.

**[Table 5]**

We also observe that the importance rankings of input variables related to the region and industry in which the enterprise is located, such as *Adj_ROA*, *GDP_Growth*, and *Personal_Income*, are high in both samples. At the regional level, a higher GDP growth rate and per capita income usually imply that the company is located in a province with a prosperous economy and strong market demand, which helps it expand the business and enhance profitability, thus reducing fraud risk; and contrarily in regions where economic growth slows down or falls into recession, or where per capita income is relatively low, enterprises may carry a greater risk of operating pressure and market competition, leading to their tendency to adopt low-cost and high-efficiency business strategies, and even engage in non-compliant behaviours, so as to maintain survival or pursue higher profits, which can be accompanied by a much higher risk of irregularities. For the industry dimension, a company's relative position and earning power in its industry can notably affect its probability of committing fraud. If its performance is significantly lower than the industry average level, it will be up against considerable profitability and competitive pressures, and its risk of violations may increase accordingly. Therefore, the impact of these industry-level and province-level variables on the enterprise is of equal interest when assessing its breach risk. Furthermore, it is particularly noteworthy that *Anti_Corrupt* is one of the top 5 most crucial variables in both samples. This variable, which represents the anti-corruption campaign undertaken in China, well emphasises the significance of strengthening the legal environment and regulatory efforts from the national level in improving the corporate governance environment. Under such circumstances,

companies face higher legal risks and more stringent legal sanctions, while their governance structure has been reformed, and there is also a shift in corporate culture and values towards one that focuses more on integrity and compliance management. Under the combined influence of these factors, the likelihood of firms engaging in fraudulent activities can be effectively reduced.

Panel B in Table 5 summarises the importance of each category of input variables. The result shows that, in addition to the *Other* category containing key influencing variables such as *Past_Fraud* and *SOE*, *Exposure* is the foremost category, accounting for 14.39% and 10.92% of the aggregate importance in the Type I and Type II samples, respectively. This indicates that the business information transparency plays a distinctly important role in identifying potential corporate violations. A highly transparent enterprise can better accept external oversight and evaluation, reducing the fraud risk and also helping to promote the firm's long-term development. Next are the *Ownership Structure* category (taking the Type I sample as an example, 10.61%; the same below), *Environmental* category (10.39%), *Financial Index* category (10.53%), and *Industry* category (6.37%). The stability and rationality of equity structure are the core content of good corporate governance, which is crucial to the future development of enterprises; the external environment and industry characteristics of a company are closely related to its business strategy, and also have a vital impact on its healthy development and violation risk. These non-financial factors occupy a significant position in the research of corporate violation detection (Bertomeu et al., 2021), and by complementing them with financial indicators can the firm's operational status and fraud risk be more comprehensively assessed. Finally, the importance of *Corporate Governance* category (2.35%) and *Personnel Characteristic* category (0.95%) in predicting corporate misconduct is relatively lower, but this does not mean that the variables under these two

categories have no substantial influence on the propensity of enterprises to breach. In fact, these complex and diverse variables still hold great potential for corporate fraud forecasts.

### 3.3 Different Types of Corporate Violations

Existing research has mostly focused on the effectiveness of machine learning in detecting accounting fraud in enterprises. Although financial improprieties do constitute the main body of corporate fraud, considering that various types of violations often involve different motives and manifestations, the detection of diversified fraud is necessary. This is helpful to get a whole picture of the influencing factors of different types of corporate irregularities and carry out more precise fraud risk prevention and control. Therefore, this study re-examines the predictive performance of machine learning models in the accounting fraud sample, the operation fraud sample, and the leadership fraud sample, and Panels A, B, and C of Table 6 report the outcomes of these three types of violations, respectively. For most of the crucial evaluation metrics, the GBDT model outperforms the RF model in both Type I and Type II samples, whether the models predict accounting cheating, operation cheating, or leadership cheating. This is consistent with the forecasting performance of the whole fraud sample (as shown in Table 4), again supporting the superior ability of the GBDT algorithm to improve the validity and accuracy of fraud detection. Taking the prediction results of GBDT in the Type I sample as an instance, it can be firstly found that the accounting fraud sample has the best performance among the three, which can achieve a high precision rate (0.6632) and recall rate (0.3860), and its F1-score and MCC value are 0.4880 and 0.4145, respectively. The results indicate that GBDT has good predictive ability in detecting accounting irregularities, and can effectively avoid false-positive and false-negative errors to a certain extent. Besides, the AUC value predicted by GBDT can reach 0.6668, further illustrating the excellent performance of the GBDT model in recognizing accounting malpractices in

enterprises. These results once again provide empirical support for the validity of machine learning in predicting corporate accounting fraud, and moreover, given the prevalence and harmfulness of accounting fraudulent activities, the precise forecast of accounting irregularities is of great significance in protecting investors' benefits and safeguarding market order.

**[Table 6]**

Compared with accounting irregularities, machine learning models do not perform so well in predicting operation and leadership irregularities. Despite the GBDT model has a high precision in forecasting these two types of violations, the recall rates obtained are less than satisfactory (0.0692 for operation violations and 0.1149 for leadership violations). This is largely owing to the fact that the motivations and forms of the other two types of fraud are usually more complex than those of accounting fraud, which makes these fraudulent behaviours more insidious and tough to be discovered in time, leading to greater difficulty in collecting relevant data and limitations on sample size and quality, thus influencing the predictive effect. So MCC offers a more stable assessment for these two types of fraud samples with an imbalanced distribution. As can be seen from Panels B and C, the MCC values of operation and leadership breaches are 0.1801 and 0.2605, respectively. This shows there is a certain degree of consistency between the classification results and the real situation in these two types of misconduct samples, suggesting that the GBDT model can effectively predict corporate fraud, though it cannot achieve fairly good performance. Their AUC values, which are 0.5330 and 0.5552, respectively, obviously both outperform the random guess (i.e., 0.5), providing more evidence for the feasibility of machine learning techniques used in the fraud detection of publicly traded companies, as well as suggesting that the improvement of prediction accuracy of operation and leadership violations deserves more attention.

Next, we also rank the feature importance of the factors affecting these different kinds of corporate wrongdoings to be aware of variables taking a more important role in predicting specific types of illegal practices so that we are able to analyse the motives and manifestations of various irregularities in a targeted manner, as well as develop efficacious preventive measures and optimise regulatory strategies. Table 7 compares the importance of each category of input variables when predicting these three types of breaches, and the relative magnitude of each input variable is presented in Appendix Table 3. Firstly, the significance ordering of input variables in the accounting misconduct prediction is not much different from that of the whole fraud sample shown in Table 5. Specifically, *Exposure* remains the paramount category (accounting for 17.09% and 14.60% of the aggregate materiality in the Type I and Type II samples, respectively), and the variables under the categories of *Financial Index*, *Ownership Structure*, and *Industry* have a profound impact on the detection of corporate financial malpractices as well. This again demonstrates the universality of accounting violations among listed companies, and they are mainly influenced by the drive for economic benefits and other factors, which are comparatively easy to identify and quantify, resulting in a higher predicting accuracy of accounting irregularities.

**[Table 7]**

Secondly, for the operation fraud detection, the *Financial Index* becomes the most important category (accounting for 19.53% and 15.10% of total significance in the Type I and Type II samples, respectively), followed by *Exposure* variables. Financial factors such as *ZScore* and *Cash* rank among the top 10 in both samples (see Panel B of Appendix Table 3). This is principally because operation violations are often closely related to the daily business activities of enterprises, while these activities are directly reflected in their financial conditions. For instance, fictitious transactions, misappropriation of funds, and other operating infractions may lead to abnormal

changes in a firm's cash holdings. Therefore, changes in financial indicators can serve as an immediate warning signal of operation irregularities, prompting regulatory agencies to pay timely attention to potential operation wrongdoings. In contrast, accounting fraud is more involved in the exactitude and transparency of financial statements and might not necessarily be directly embodied in the financial indices of day-to-day operations. In addition, the *Corporate Governance* category has remarkably increased its proportion in overall importance, especially the significance ranking of variables like *Shareholder_Meeting* and *Board_Meeting* has risen quite a few, which concerned with the immediate impact of corporate governance structure on business decision-making. The structure of company governing determines the efficient degree of the allocation of power and supervision mechanism within an enterprise. A sound corporate governance can ensure the scientific, rational and transparent nature of operating decisions, thereby reducing the likelihood of operation breaches occurring. As key signals reflecting the validity of corporate governance structure, the frequency of board meetings and shareholder meetings gives expression to the attention and participation of directors and shareholders in the firm's management status, and will directly affect the corporate business behaviours and decision-making quality, thus influencing the risk of operating violations. Moreover, it is worth noting that the variable *HHI* is one of the crucial factors for forecasting operation irregularities (ranking 13th and 5th in materiality in the Type I and Type II samples, respectively). For the prediction of operating misconduct, the variation in HHI not only reflects the intensity of market competition, but may also image changes in corporate behaviour. In highly concentrated markets, firms might face fewer competitors and greater management risks, as the business decisions of a few enterprises that dominate the market can have a significant impact on the entire market. As a result, companies are more inclined and more likely to gain market share or maintain market position through committing operation violations.

Finally, *Ownership Structure* variables contribute more to the prediction of leadership irregularities than the input variables of other categories, and this category accounts for 24.70% and 19.33% of the total importance in the Type I and Type II samples, respectively. The results of Panel C in Appendix Table 3 show that *Board_Own*, *Man_Own*, *Chair_Own* and other variables related to equity structure are all among the top in both samples. The ownership structure dictates the power distribution and incentive mechanisms within the company. When corporate leaders, including the management, chairman, and board members, hold a high proportion of shares, they are highly likely to pursue the corporate short-term performance and personal profits and resort to non-compliance in spite of a close tie between their personal interests and the firm's long-term interests, because the short-term gains from illegal actions may far exceed the risks and costs they face. At the same time, as the core decision makers, leaders with an excessively high shareholding ratio may weaken the constraint ability of the internal monitoring mechanism on them, thus increasing the risk of leadership violating regulations. Hence, the significance of the *Ownership Structure* category in predicting leadership violations has observably increased. The next categories with higher relative importance are *Environmental* (accounting for 15.49% and 20.73% of the aggregate materiality in the Type I and Type II samples, respectively; the same below) and *Industry* (15.40% and 15.69%). Among them, the variables *Personal_Income* and *HHI* rank first and second in both samples, and taking the Type I sample as an example, their significance is as high as 8.49% and 7.21%, respectively. The region and industry where the enterprise is located not only have a remarkable impact on the leadership's business strategies and decision-making styles, but may also affect their psychological state and moral concepts, undoubtedly influencing their propensity to violate the rules. For instance, the per capita income level in this region is closely related to the leader's economic situation, and a lower individual income can make the

leader, under financial pressure or driven by greed, more susceptible to the temptation to rapidly increase personal wealth through any irregularities. While in a fiercely competitive market, leaders may face even greater performance stresses and market risks, leading them to take a number of illegal competitive measures to cope with these pressures and keep the company's market standing. Therefore, *Environmental* and *Industry* variables have higher importance in forecasting leadership fraud compared to accounting irregularities and operation irregularities. Furthermore, it should be noted that the importance of the *Exposure* category has relatively decreased (5.25% and 4.19%). Unlike accounting and operation violations, which usually have a direct correlation with corporate transparency and compliance, leadership wrongdoings involve issues such as improper exercise of power, leakage and abuse of inside information, illegal transactions, and so on, and these behaviours tend to be more concealed and difficult to detect through normal disclosure channels. And that's why the ability of *Exposure* variables to predict leadership irregularities is limited.

### 3.4 Additional Analyses

#### 3.4.1 Different Ownership Enterprises

With the rapid development of China's capital market and the deepening of reform and opening-up, there have been a growing number of non-SOEs that can more quickly adapt to the demands and changes of the market. However, SOEs are still a force to be reckoned with due to their significant market position and strong competitiveness. SOEs and non-SOEs, as the two main components of the Chinese market economic system, have differences in terms of resource allocation, business objectives, market competition, and other aspects. Specifically, SOEs mostly enjoy direct government support and more preferential policies and have relatively easy access to resources, making them take part in market competition with lower costs and higher efficiency.

Meanwhile, SOEs are often subject to some kind of government regulation and intervention in resource distribution, which requires them, in addition to pursuing economic benefits, to emphasise social effects and public interests, as well as bear social responsibilities and public functions such as providing public services and promoting social stability. Hence, their business targets and strategic positioning are more diversified and complex. Non-SOEs, by contrast, rely more on market mechanisms to obtain the necessary resources, facing more tremendous pressure and risk of market competition; and in the allocation of resources, non-SOEs focus more on economic benefits and market demand, and constantly improve their profitability and market competitiveness with the primary management goal of maximising economic profits and shareholder wealth. These differences determine their various roles and functions in the national economy, and also enable them to exhibit disparate characteristics and performances during their operating processes. In this study, the variable *SOE* is almost always one of the top 10 input variables in terms of relative importance ranking, whether in the whole sample of corporate breaches or in samples of different types of violations (see Table 5 and Appendix Table 3 for details), indicating a strong correlation between the system of ownership and corporate fraud risk. Therefore, we want to further explore how SOEs and non-SOEs perform in predicting fraudulent behaviours.

Table 8 reports the predictive performance of the RF and GBDT models in both SOE and non-SOE samples. Consistent with the previous results, for the majority of evaluation metrics, GBDT demonstrates superior performance than RF in forecasting the enterprise's propensity to commit irregularities, and the prediction outcome in the Type II sample (as shown in Panel B) is slightly better than that in the Type I sample (see Panel A), but the difference is tiny. So we take the classification results of the GBDT model in the Type II sample as an example to compare the

fraud forecasts of enterprises with different ownership. First, the prediction of infractions in non-SOEs can achieve better performance than in SOEs in terms of precision and recall. In particular, a recall rate of 0.6051 suggests that GBDT is able to correctly recognize about 60% of fraudulent non-SOEs, which increases the probability of precisely identifying illegal SOEs by approximately 20%! This shows that machine learning can reduce underreporting to a greater extent and significantly mitigate the risk of false-negative error when predicting irregularities in non-SOEs. The F1-score of malpractice prediction for non-SOEs (0.6499) is also remarkably larger than that for SOEs (0.5061), implying that GBDT can not only reach higher precision and recall when recognizing misconducts in non-SOEs but also realise a preferred trade-off between the two, resulting in better performance. Moreover, both the MCC and AUC values of the non-SOE sample (0.4880 and 0.7347, respectively) are greater than those of the SOE sample (0.4366 and 0.6782, respectively), and this once again supports the GBDT model's good performance in fraud detection of non-SOEs.

**[Table 8]**

This result is interesting but not particularly surprising. SOEs usually bear more social responsibilities and expectations, and are also subject to more stringent government supervision and auditing. Such a high-intensity regulatory environment will cause SOEs and their executives to preserve the reputation and political image of enterprises and individuals, pay attention to their behavioural compliance, thus reducing the risk of violations. While non-SOEs are more likely to engage in fraudulent activities under the incentive of individual performances and short-term profits compared to SOEs, this can also be supported by evidence from the number of fraud samples in these two types of enterprises (4,155 for SOEs vs. 9,297 for non-SOEs). Notably, the relative importance of *Exposure* variables such as *Disclosure* and *Opinion* in forecasting the fraud

tendency of SOEs has signally weakened and even dropped out of the top 10, whereas their materiality in predicting breaches by non-SOEs aligns with previous outcomes. This might be related to the direct or indirect government intervention in the information disclosure of SOEs. For reasons such as protecting national security or stabilising economy, it is possible that the government has restricted or delayed the disclosure of information correlated with SOEs, which, coupled with the stricter disclosure requirements and more cumbersome review mechanisms and processes faced by SOEs, can reduce corporate transparency, leading to an attenuated relationship between the quality of information disclosure in SOEs and their fraud risks. Overall, the differences in misconduct forecast performance of enterprises with different ownerships manifest that adopting machine learning technology for fraud detection among Chinese listed firms, especially non-SOEs, is more feasible and effective. Regulatory authorities and investors should take the ownership of companies into account and focus on their information disclosure, financial status, market behaviours, and other respects more targeted, in order to guard against possible violations.

### 3.4.2 Different Information Disclosure Qualities

The quality of corporate disclosure directly concerns the image and fame of the enterprise, as well as being crucial for investors. High-quality information revelation can enhance the transparency of business, which helps investors better assess its value and risk, thereby boosting their confidence and making informed investment decisions, and which also facilitates the firm itself to establish a favourable image and attract higher investments and funds. Of course, if the quality of information disclosure is low, the company may intentionally conceal or distort material facts to mislead the public. And this not only impairs the interests and confidence of investors, but also undermines the fairness and stability of the capital market. In this study, the input variable

*Disclosure* has a greater impact in corporate fraud prediction than other variables, meaning that the discrepancy of disclosure quality can remarkably affect the likelihood of companies engaging in irregularities. Therefore, we divide the whole sample into four groups according to the evaluation results of corporate information disclosure (i.e., excellent, good, qualified, and unqualified or result missing), and re-examine the effectiveness of machine learning models in predicting violations in these four groups. Panels A and B of Table 9 present the prediction performance of machine learning models in the Type I and Type II samples of these four categories, respectively, and it can be found that the classification results of the models in both samples are similar without vast differences. So we take the outcomes in the Type I sample (as shown in Panel A) as an instance to compare the cheating prediction of enterprises with different disclosure qualities.

**[Table 9]**

We can first see that the RF and GBDT models perform the worst in forecasting wrongdoings in the "Excellent" group among the four groups. GBDT can only correctly identify 20.57% of fraudulent companies whose information disclosure evaluation results are excellent, and its F1-score only reaches 0.2865. But this is inseparable from the small sample size of violations in this group (907 firm-year observations), which also reflects from the side that high-quality disclosure enable to reduce the possibility of corporate fraud. The high-quality disclosure of information not only enhances an enterprise's reputation and credibility, but also contributes investors and other stakeholders to have a more comprehensive understanding of the business situation, and such high level of transparency helps to lessen information asymmetry and the probability of illegal behaviour occurring. When such firms face performance pressure, they are more likely to solve problems through compliance rather than resorting to fraudulent means so as to avoid reputational

damage. In addition, the AUC values of RF and GBDT (0.5501 and 0.5766, respectively) indicate that the predictive performance of these models is better than that of random guess, which still supports the feasibility of applying machine learning in the corporate violation prediction. While in the other three groups, RF and GBDT display obvious improvement in recognizing fraudulent enterprises. Taking the "Unqualified or Missing" group as an example, the GBDT model performs slightly better in prediction than RF, with its precision and recall reaching 0.6888 and 0.5507, F1-score and AUC values of 0.6121 and 0.7277, respectively. This result demonstrates that the models can effectively detect fraud in enterprises with unsatisfactory or missing results of disclosure assessment, and reduce both false-positive and false-negative errors simultaneously to some extent. This implies that, on the one hand, firms with inferior disclosure quality might be more prone to issues such as loopholes in internal control or poor management, resulting in an increased likelihood of irregularities; on the other hand, the problem of incomplete or inaccurate information disclosure can lead to doubts among regulatory agencies and investors about the operation of companies and strengthen the supervision of these firms, which in turn raises the possibility of corporate malpractices being discovered.

Machine learning has shown outstanding forecasting performance in the "Qualified" group, achieving high precision and recall rates in identifying violating businesses among those with qualified disclosure evaluation results. Both the RF and GBDT models can accurately recognize about 76% of offending parties, and at the same time, approximately 68% of the non-compliant firms they predict are truly fraudulent. The models' ability to forecast corporate skulduggeries with high precision and markedly reduce the underreporting of illegal enterprises is closely related to the high proportion of fraud samples in this group, which is 48%. An interesting phenomenon is that, compared to companies whose disclosure assessment results are unqualified or missing, firms

with good and qualified evaluation results have a higher incidence of fraud. And this might be due to a combined effect of multiple factors. First of all, enterprises with evaluation results of "Good" and "Qualified" may loosen their emphasis on the authenticity of information disclosure because their disclosure quality has already met the regulatory requirements, and there may also be certain problems in internal control. These issues can lead firms and internal personnel to take advantage of the grey zone in supervision standards to commit infractions. While for enterprises with "Unqualified or Missing" results, they may pay more attention to improving their internal governance mechanisms to meet the proposed rectification requirements, as they will be under the focus of the regulatory authorities, thus reducing the fraud risk to a certain extent. What's more, the "Good" and "Qualified" assessment results can confront enterprises with greater performance pressure and market competition pressure, especially those with "Qualified" results. These companies have a strong incentive to enhance short-term performance and satisfy market expectations in order not to be judged as unqualified, thereby increasing the likelihood of non-compliance. In the "Qualified" group, the predictive performance of the RF model is comparable to that of the GBDT model, and even exhibits better stability. This manifests that RF and GBDT each have their advantages in different situations. Specifically, in the case of balanced sample categories, the prediction results of each tree in the RF can reflect the real distribution of data exactly, coupled with the model's robustness to noise and outliers, making the integrated overall predicting outcomes more reliable and stable; whereas regarding imbalanced categories, since GBDT calculates the gradient of the current loss function in each iteration and trains a new decision tree on this basis, it can focus more on the category with fewer samples (i.e., the violating samples), which improves its prediction capability for offending enterprises.

Finally, it is worth noting that, given equal sample sizes of firms with "Good" and "Unqualified or Missing" assessments, the latter group is observably superior at forecasting irregularities than the former group, despite the fact that there are more fraudulent companies in the former group than in the latter group (5,779 for "Good" vs. 4,188 for "Unqualified or Missing"). From the classification results of the GBDT model, we can see that the model is able to more effectively distinguish true illegal enterprises among the firms whose disclosure evaluation results are unqualified or missing, and significantly reduce the fraud cases failed to report (especially the recall rate has increased by 8.36%). And compared to recognizing misconducts among companies with good disclosure assessment results, GBDT has better predictive performance in the "Unqualified or Missing" group, with notably higher F1-score (0.6121), MCC value (0.4907), and AUC value (0.7277) than those of the former group (0.5449, 0.3896, and 0.6754, respectively). All these results suggest that machine learning models perform better in detecting breaches in firms with poor disclosure quality, which again supports the existence of trust traps in regulation. Enterprises with good disclosure evaluations often draw more trust and attention from supervisory authorities and investors, and this trust may lead regulators to relax their vigilance when assessing these businesses, overlooking some of the potential fraud risks. Meanwhile, these companies can use complicated financial structures and other ways to conceal their illegal behaviours, making it difficult to detect fraudulent means. In contrast, enterprises with poor evaluation results are more likely to alert investors and regulatory agencies, so their offending activities have a higher probability of being discovered. Therefore, it is necessary to further raise the quality and transparency of information disclosure by listed companies, and regulators should also strengthen the supervision and censorship of enterprises, in particular paying special attention to those

obtained “Good” and “Qualified” results in disclosure quality assessment so as to detect latent cheating risks.

### 3.5 Seriality of Corporate Fraud

Corporate violations often has seriality. As we have found, the relative importance of the variable *Past_Fraud* is far ahead of other factors affecting the propensity of corporate fraud; in other words, enterprises that have engaged in irregularities are more inclined to continue to commit fraudulent activities. This phenomenon can also be confirmed by the sample size of this study, with the proportion of firms that will violate regulations in the next year among those that committed fraud in the current year (0.6286) far exceeding the proportion among those that did not engage in breaches that year (0.1570). There is a strong link between this phenomenon and the misconduct environment formed by enterprises and the lack of external supervision. Due to the fact that fraudulent acts tend to bring immediate benefits to businesses and individuals, if regulators fail to timely detect corporate wrongdoings or impose insufficient penalties for violations, the fraud cost can be reduced. And this may lead enterprises or their management to ignore the negative impact of malpractices and continue to gain more profits through cheating means, thereby increasing the likelihood of further violations in the following years. This, once again, highlights the significance and necessity of improving the accuracy and timeliness of corporate fraud detection.

Considering the possible influence of the continuity of corporate non-compliance, we group the our sample according to the number of irregularities that occurred in the same year, and then re-examine the predictive performance of the two ensemble learning models to serve as a robustness test. The number of companies that did not commit any violations that year is much

higher than those with fraudulent practices, constituting the bulk of the whole sample (about 2/3). Table 10 shows that the predicted results are similar to previous findings. Taking the GBDT model's performance in the Type I sample as an example (see Panel A), when identifying enterprises with a tendency to violate from those that have not engaged in misconducts in that year, although it is necessary to improve the coverage for recognizing real non-compliant firms (with the recall of 0.2829), the AUC value, which is the overall performance evaluation metric, has a good outcome (0.6293). This suggests that the GBDT model still possesses superior power in predicting discontinuous fraudulent behaviours, and also proves the reliability and consistency of our findings. Besides, we report results of subsamples with the number of violations of 1 and more than 1, with the two subsamples being equally large. The latter have a larger number of fraud samples than the former, apparently again providing support for the seriality of corporate irregularities. It can be seen that in the classification outcomes of the Type I samples of these two groups, the models can achieve high precision and high recall at the same time, indicating that the fraud detection model can contributed more to the reduction of false-positive and false-negative errors when predicting the possibility of fraud in enterprises with violations in that year, and especially has excellent forecast performance when conducting fraud detection in companies with multiple breaches (the F1-scores of RF and GBDT are 0.8203 and 0.8140, respectively). Certainly, both RF and GBDT exhibit similarly superior performance in these two groups, further supporting the respective strengths of the two ensemble methods for various samples (i.e., category-balanced and category-unbalanced). To sum up, the results, consistent with previous ones, demonstrate the effectiveness and superiority of machine learning models in predicting infractions of Chinese listed companies.

**[Table 10]**

## 4. Conclusion

Corporate fraudulent activities not only imply huge economic losses and reputational damage to enterprises, but also a serious infringement on the interests of the broad investors, weakening their confidence in the capital market, and at the same time, poses a severe threat to the healthy development of the market by disrupting its good order. That's why the exact prediction and timely detection of violations have become an indispensable part of corporate governance practices. However, corporate fraud detection has always been a challenging task. With the vigorous development of machine learning technology, the prediction model of financial malpractices based on machine learning algorithms has gradually gained widespread attention and been effectively applied in developed markets, while its application in developing countries is still in the exploratory stage. Therefore, investigating the validity of machine learning techniques in predicting irregularities by Chinese listed companies has great practical significance and is crucial for maintaining the healthy development of emerging markets. This study first systematically reviews the influencing factors of corporate fraud, and determines a more comprehensive set of variables related to breaches in conjunction with the GONE framework. Based on these financial and non-financial input variables, we select two representative ensemble learning algorithms, RF and GBDT, to build the fraud prediction model. Through the empirical research, we find that the GBDT model performs better in predicting misconducts in Chinese enterprises, especially observably outperforming the RF model in key metrics such as recall, which indicates that GBDT is able to identify real offending firms more efficiently and precisely and reduce the risk of underreporting. Besides, we use these two models to make predictions about different types of corporate violations. The results show that, compared to operation irregularities and leadership

irregularities, both RF and GBDT exhibit better performance in predicting accounting irregularities, and GBDT still has a more remarkable predictive performance. This finding not only provides renewed empirical support for the effectiveness of machine learning in detecting financial infractions, but reveals the importance of improving the accuracy of diversified fraud detection for more targeted breach risk prevention and control. In addition, we explore the impact of different ownership systems and different disclosure qualities on the likelihood of enterprises committing fraudulent behaviours, and find that the state-owned ownership and high-quality disclosure can help reduce the fraud risk. So supervision agencies should strengthen the regulation and censorship of non-SOEs as well as enterprises with good and qualified disclosure quality evaluation results. To ensure the robustness and reliability of our findings, taking the possible effects of the seriality of wrongdoings and the proxy for fraud into account, we introduce new *Exposure* variables and use replacement samples for testing. The consistent research results demonstrate the stability of this study and further prove the applying value and dependability of machine learning models in the field of predicting irregularities by Chinese public companies.

Improving the precision of corporate fraud prediction is not only about recognizing potential misconducts more accurately, but more momentously about constructing a strong deterrence mechanism through strengthening the forecast ability, so as to notably inhibit the occurrence of future violating behaviours (Karpoff, 2021). The achievement of this goal also relies on the accuracy and timeliness of early warning signals, which can provide regulators and investors with targeted monitoring strategies, thereby enhancing the efficiency and effectiveness of fraud detection. Therefore, this study ranks the relative importance of input variables in order to exactly identify the risk factors that have a greater influence on predicting corporate misbehaviour. The result shows that *Exposure* variables play a major role in forecasting breaches, and in particular,

the quality of information disclosure (*Disclosure*) contributes more than other input variables. This finding emphasises the significance and necessity of establishing a more transparent information environment. Moreover, *Ownership Structure* variables and *Environmental* variables present almost equal aggregate materiality as *Financial Index* variables, which again proves the core position of non-financial factors in corporate violation forecast. These multi-type variables complement each other and can collectively form a comprehensive framework for predicting corporate non-compliance. For diversified fraud prediction, we find that the importance of these influencing factors varies slightly in different types of fraudulent practices. Specifically, *Exposure*, *Ownership Structure*, and *Financial Index* variables continue to play a more vital role in the detection of accounting malpractices; the *Financial Index* category outweighs the *Exposure* category in predicting operation violations, and the aggregate materiality of *Ownership Structure* variables significantly increases; while in the prediction of leadership irregularities, *Ownership Structure* variables are more influential than other categories of input variables. Investors and regulatory authorities ought to, based on the above findings, pay full attention to different warning indicators when assessing the risk of corporate infractions, thus raising the prediction accuracy of various types of fraudulent activities in a more targeted manner. Meanwhile, we also note the relatively low materiality of the *Personnel Characteristic* category; variables such as the age and gender of the chairman (*Chair_Age*, *Elder*, and *Chair_Gender*), as well as the educational and work backgrounds of directors, supervisors, and executives (*Fin_Back* and *Overseas_Back*), have a less remarkable role in predicting the propensity of enterprises to violate regulations. This may be because these variables are more difficult to measure than other input variables, which causes the models to underestimate or fail to adequately capture their impact on corporate wrongdoing. Nevertheless, this does not imply that these variables have no bearing on detecting corporate fraud.

Future research could further explore other variables under *Personnel Characteristic* to fully evaluate the influence of this category on violation prediction. Overall, the findings of this study provide in-depth insights into the important factors that affect corporate breaches and suggest that the fraud detection model should include these key variables to optimise predictive performance. And this has crucial guiding significance for perfecting the governance mechanism of Chinese listed companies, enhancing the supervision level in the China's capital market and reinforcing the effectiveness of fraud detection system.

This study, of course, has several limitations. Firstly, in the selection process of input variables, despite our best efforts to consider a wide range of variables, our sample is subjected to the availability of data and cannot cover all the potential influencing variables, which may lead to bias in the results. Future research could explore how to effectively integrate multi-dimensional features and multi-source data, in order to more comprehensively and deeply reveal the complexity of corporate violations and improve the precision of fraud prediction. Secondly, this study uses the methods of elimination and simple filling to obtain the Type I and Type II samples during sample processing. These two types of samples can provide available and complementary results, but their shortcomings also limit the predictive power of models to a certain extent. Future studies may consider optimising sample processing strategies, for example, filling the blanks in the sample with data closer to the real information as much as possible while ensuring a reasonable distribution, thereby improving the forecast accuracy. What's more, refining the fraud prediction model is, no doubt, a substantial aspect of future research. The aim of this study is not to adopt the latest technologies or develop new algorithms to promote the model's performance. This study, however, from a new perspective, takes a more comprehensive set of factors affecting corporate non-compliance as input variables and applies valid machine learning models in the field of

diversified fraud prediction for Chinese quoted companies, demonstrating the generalisability of machine learning to corporate fraud detection in emerging markets. Although we are not able to attempt all machine learning methods, the robust results of this study make us have good reasons to believe that, corporate misconducts in the Chinese market can be prevented and combated with the power of machine learning techniques. And future research can further expand and deepen on this basis, optimising and ameliorating machine learning algorithms to achieve the best prediction. Finally, it is still worthwhile to continue research on how to better address the imbalance issue, so as to recognize all types of fraudulent behaviours more correctly.

**Table 1: Distribution of Chinese A-Share Listed Companies Involved in Violations by Year over 2000-2022**

This table shows the annual occurrence of violations by Chinese A-share listed companies between 2000 and 2022. The data covers the actual years and types of violations by listed firms in the violation announcements that issued by the China Securities Regulatory Commission, stock exchanges and other institutions from January 1, 2000, to December 31, 2023, drawn from the CSMAR database. Panel A shows the overall picture of listed companies' infractions, mainly the frequency and proportion of offending parties for each year. Panel B reports the annual distribution of different types of fraud committed in listed companies, including misbehaviors related to accounting irregularities, operation irregularities, leadership irregularities, and other types of irregularities, detailing the number of these types of violations and the number of companies involved. The specific types of violations are defined in Appendix Table 1.

**Panel A: Distribution of Fraud Firms**

| Year | Number of Listed Firms | Number of Fraud Firms | Percentage (%) | Number of Violations |
|---|---|---|---|---|
| 2000 | 1,061 | 147 | 13.85 | 169 |
| 2001 | 1,142 | 234 | 20.49 | 317 |
| 2002 | 1,199 | 186 | 15.51 | 246 |
| 2003 | 1,269 | 163 | 12.84 | 230 |
| 2004 | 1,355 | 172 | 12.69 | 240 |
| 2005 | 1,351 | 148 | 10.95 | 210 |
| 2006 | 1,434 | 167 | 11.65 | 239 |
| 2007 | 1,548 | 239 | 15.44 | 320 |
| 2008 | 1,602 | 275 | 17.17 | 371 |
| 2009 | 1,751 | 337 | 19.25 | 469 |
| 2010 | 2,106 | 352 | 16.71 | 496 |
| 2011 | 2,341 | 502 | 21.44 | 767 |
| 2012 | 2,472 | 598 | 24.19 | 958 |
| 2013 | 2,515 | 606 | 24.10 | 1,009 |
| 2014 | 2,631 | 606 | 23.03 | 1,107 |
| 2015 | 2,823 | 820 | 29.05 | 1,103 |
| 2016 | 3,263 | 959 | 29.39 | 2,085 |
| 2017 | 3,648 | 1,066 | 29.22 | 2,567 |
| 2018 | 3,750 | 1,182 | 31.52 | 3,062 |
| 2019 | 3,960 | 1,122 | 28.33 | 2,945 |
| 2020 | 4,413 | 1,059 | 24.00 | 2,545 |
| 2021 | 4,835 | 1,116 | 23.08 | 2,372 |
| 2022 | 5,134 | 975 | 19.00 | 1,804 |

**Panel B: Distribution of Fraud Firms with Different Types of Violations**

| Year | Violation Type | Number of Firms Involved in This Type | Number of Violations of This Type | Percentage to Total Number of Fraud Firms (%) |
|---|---|---|---|---|
| 2000 | Accounting Irregularities | 129 | 145 | 87.76 |
| | Operation Irregularities | 25 | 27 | 17.01 |
| | Leadership Irregularities | 7 | 7 | 4.76 |
| | Other Types | 16 | 16 | 10.88 |
| 2001 | Accounting Irregularities | 205 | 273 | 87.61 |
| | Operation Irregularities | 49 | 54 | 20.94 |
| | Leadership Irregularities | 15 | 15 | 6.41 |
| | Other Types | 28 | 28 | 11.97 |
| 2002 | Accounting Irregularities | 168 | 222 | 90.32 |
| | Operation Irregularities | 42 | 45 | 22.58 |
| | Leadership Irregularities | 10 | 10 | 5.38 |
| | Other Types | 8 | 8 | 4.30 |

| | | | | |
|---|---|---|---|---|
| 2003 | Accounting Irregularities | 149 | 205 | 91.41 |
| | Operation Irregularities | 47 | 51 | 28.83 |
| | Leadership Irregularities | 8 | 8 | 4.91 |
| | Other Types | 11 | 12 | 6.75 |
| 2004 | Accounting Irregularities | 159 | 221 | 92.44 |
| | Operation Irregularities | 62 | 73 | 36.05 |
| | Leadership Irregularities | 3 | 3 | 1.74 |
| | Other Types | 11 | 11 | 6.40 |
| 2005 | Accounting Irregularities | 123 | 180 | 83.11 |
| | Operation Irregularities | 43 | 49 | 29.05 |
| | Leadership Irregularities | 4 | 4 | 2.70 |
| | Other Types | 17 | 18 | 11.49 |
| 2006 | Accounting Irregularities | 132 | 192 | 79.04 |
| | Operation Irregularities | 43 | 52 | 25.75 |
| | Leadership Irregularities | 13 | 13 | 7.78 |
| | Other Types | 21 | 21 | 12.57 |
| 2007 | Accounting Irregularities | 164 | 222 | 68.62 |
| | Operation Irregularities | 41 | 48 | 17.15 |
| | Leadership Irregularities | 38 | 41 | 15.90 |
| | Other Types | 39 | 40 | 16.32 |
| 2008 | Accounting Irregularities | 182 | 243 | 66.18 |
| | Operation Irregularities | 45 | 48 | 16.36 |
| | Leadership Irregularities | 55 | 62 | 20.00 |
| | Other Types | 51 | 53 | 18.55 |
| 2009 | Accounting Irregularities | 224 | 302 | 66.47 |
| | Operation Irregularities | 48 | 49 | 14.24 |
| | Leadership Irregularities | 78 | 86 | 23.15 |
| | Other Types | 64 | 67 | 18.99 |
| 2010 | Accounting Irregularities | 233 | 330 | 66.19 |
| | Operation Irregularities | 46 | 54 | 13.07 |
| | Leadership Irregularities | 70 | 78 | 19.89 |
| | Other Types | 63 | 68 | 17.90 |
| 2011 | Accounting Irregularities | 343 | 522 | 68.33 |
| | Operation Irregularities | 70 | 85 | 13.94 |
| | Leadership Irregularities | 88 | 102 | 17.53 |
| | Other Types | 106 | 118 | 21.12 |
| 2012 | Accounting Irregularities | 418 | 687 | 69.90 |
| | Operation Irregularities | 84 | 98 | 14.05 |
| | Leadership Irregularities | 95 | 111 | 15.89 |
| | Other Types | 135 | 145 | 22.58 |
| 2013 | Accounting Irregularities | 438 | 728 | 72.28 |
| | Operation Irregularities | 83 | 107 | 13.70 |
| | Leadership Irregularities | 142 | 165 | 23.43 |
| | Other Types | 101 | 118 | 16.67 |
| 2014 | Accounting Irregularities | 447 | 791 | 73.76 |
| | Operation Irregularities | 85 | 109 | 14.03 |
| | Leadership Irregularities | 133 | 178 | 21.95 |
| | Other Types | 118 | 153 | 19.47 |
| 2015 | Accounting Irregularities | 603 | 1,094 | 73.54 |

| | | | | |
|---|---|---|---|---|
| | Operation Irregularities | 108 | 166 | 13.17 |
| | Leadership Irregularities | 241 | 327 | 29.39 |
| | Other Types | 151 | 205 | 18.41 |
| 2016 | Accounting Irregularities | 718 | 1,486 | 74.87 |
| | Operation Irregularities | 170 | 281 | 17.73 |
| | Leadership Irregularities | 252 | 340 | 26.28 |
| | Other Types | 188 | 255 | 19.60 |
| 2017 | Accounting Irregularities | 815 | 1,867 | 76.45 |
| | Operation Irregularities | 209 | 432 | 19.61 |
| | Leadership Irregularities | 275 | 382 | 25.80 |
| | Other Types | 224 | 321 | 21.01 |
| 2018 | Accounting Irregularities | 910 | 2,170 | 76.99 |
| | Operation Irregularities | 263 | 599 | 22.25 |
| | Leadership Irregularities | 315 | 432 | 26.65 |
| | Other Types | 338 | 479 | 28.60 |
| 2019 | Accounting Irregularities | 864 | 2,054 | 77.01 |
| | Operation Irregularities | 267 | 556 | 23.80 |
| | Leadership Irregularities | 300 | 440 | 26.74 |
| | Other Types | 370 | 519 | 32.98 |
| 2020 | Accounting Irregularities | 820 | 1,800 | 77.43 |
| | Operation Irregularities | 223 | 396 | 21.06 |
| | Leadership Irregularities | 286 | 428 | 27.01 |
| | Other Types | 301 | 397 | 28.42 |
| 2021 | Accounting Irregularities | 802 | 1,574 | 71.86 |
| | Operation Irregularities | 133 | 223 | 11.92 |
| | Leadership Irregularities | 363 | 545 | 32.53 |
| | Other Types | 283 | 391 | 25.36 |
| 2022 | Accounting Irregularities | 645 | 1,082 | 66.15 |
| | Operation Irregularities | 61 | 90 | 6.26 |
| | Leadership Irregularities | 314 | 455 | 32.21 |
| | Other Types | 252 | 355 | 25.85 |

## Table 2: Descriptive Statistics

This table reports descriptive statistical information for all input variables used in the study, including variables related to corporate ownership structure, corporate governance, characteristics of relevant personnel (directors, supervisors, and executives), information exposure, as well as variables related to corporate financial characteristics, industry characteristics, and environmental characteristics. It is to be noted that for ease of distinction, continuous variable statistics are retained to two decimal places, while statistics for dummy variables are only retained to two decimal places of the mean and standard deviation, with the rest still reported as 0 or 1. All variables are defined in Appendix Table 1.

| | N | Mean | S.D. | Min | Median | Max |
|---|---|---|---|---|---|---|
| Ownership Structure Variables | | | | | | |
| *Board_Own* | 49,648 | 10.77 | 18.43 | 0.00 | 0.03 | 131.42 |
| *Chair_Own* | 44,882 | 7.59 | 13.73 | 0.00 | 0.00 | 80.01 |
| *Control_Own* | 47,241 | 36.91 | 15.72 | 0.00 | 35.05 | 155.19 |
| *Control_Dummy* | 47,241 | 0.75 | 0.43 | 0 | 1 | 1 |
| *Central* | 47,242 | 0.21 | 0.40 | 0 | 0 | 1 |
| *Employee_Own* | 48,184 | 0.01 | 0.40 | 0.00 | 0.00 | 29.04 |
| *Inst_Own* | 50,049 | 46.21 | 26.10 | 0.00 | 48.66 | 157.08 |
| *QFII* | 52,459 | 0.10 | 0.29 | 0 | 0 | 1 |
| *Largest_Own* | 48,185 | 34.95 | 15.44 | 0.29 | 32.41 | 89.99 |

| | | | | | | |
|---|---|---|---|---|---|---|
| *Herfindahl_5* | 48,185 | 0.17 | 0.12 | 0.00 | 0.13 | 0.81 |
| *Insider_Board* | 51,587 | 22.77 | 21.98 | 0.00 | 18.18 | 100.00 |
| *Insider_Spvisor* | 51,585 | 18.74 | 27.29 | 0.00 | 0.00 | 100.00 |
| *Man_Own* | 49,515 | 6.18 | 13.36 | 0.00 | 0.01 | 84.33 |
| *Differ* | 45,945 | 4.77 | 7.83 | -80.58 | 0.00 | 60.32 |
| *Foreign_Resid* | 45,967 | 0.03 | 0.16 | 0 | 0 | 1 |
| *Multi_Control* | 47,206 | 0.31 | 0.46 | 0 | 0 | 1 |
| Corporate Governance Variables | | | | | | |
| *Balance* | 48,185 | 0.73 | 0.63 | 0.00 | 0.56 | 4.00 |
| *Board_Size* | 51,587 | 10.06 | 2.70 | 4.00 | 9.00 | 31.00 |
| *Even_Board* | 51,587 | 0.33 | 0.47 | 0 | 0 | 1 |
| *Board_Ind* | 51,587 | 35.76 | 10.05 | 0.00 | 33.33 | 80.00 |
| *Qualified_IndBoard* | 51,587 | 0.81 | 0.39 | 0 | 1 | 1 |
| *Strong_Board* | 51,587 | 0.05 | 0.21 | 0 | 0 | 1 |
| *Centrality* | 44,646 | 0.36 | 0.36 | 0.00 | 0.33 | 4.00 |
| *Board_Meeting* | 50,485 | 9.33 | 4.01 | 0.00 | 9.00 | 58.00 |
| *Shareholder_Meeting* | 50,522 | 3.12 | 1.75 | 0.00 | 3.00 | 33.00 |
| *Committee* | 50,637 | 3.58 | 1.25 | 0.00 | 4.00 | 9.00 |
| *Duality* | 45,828 | 0.27 | 0.44 | 0 | 0 | 1 |
| *Female_Board* | 51,587 | 13.60 | 12.16 | 0.00 | 11.11 | 83.33 |
| *Gender_Diverse* | 51,587 | 17.71 | 11.24 | 0.00 | 16.00 | 80.00 |
| *Spvisor* | 51,587 | 4.25 | 1.66 | 0.00 | 4.00 | 25.00 |
| Personnel Characteristic Variables | | | | | | |
| *Chair_Age* | 51,303 | 52.26 | 7.54 | 23.00 | 52.00 | 87.00 |
| *Elder* | 51,303 | 0.23 | 0.42 | 0 | 0 | 1 |
| *Chair_Gender* | 51,313 | 0.05 | 0.22 | 0 | 0 | 1 |
| *Salary* | 49,427 | 5.43 | 1.50 | 0.00 | 5.79 | 7.68 |
| *Salary_Gap* | 46,182 | 5.24 | 0.64 | 0.00 | 5.27 | 7.45 |
| *Average_Age* | 51,579 | 48.41 | 3.51 | 34.08 | 48.50 | 64.24 |
| *Fin_Back* | 51,580 | 0.57 | 0.50 | 0 | 1 | 1 |
| *Overseas_Back* | 51,580 | 0.46 | 0.50 | 0 | 0 | 1 |
| Exposure Variables | | | | | | |
| *Analyst* | 52,459 | 5.71 | 9.11 | 0.00 | 1.00 | 80.00 |
| *Big4* | 51,594 | 0.06 | 0.24 | 0 | 0 | 1 |
| *Big10* | 51,594 | 0.45 | 0.50 | 0 | 0 | 1 |
| *Foreign_Audit* | 49,847 | 0.04 | 0.20 | 0 | 0 | 1 |
| *Opinion* | 51,593 | 0.06 | 0.24 | 0 | 0 | 1 |
| *Disclosure* | 51,398 | 1.33 | 1.12 | 0 | 2 | 4 |
| *News* | 51,398 | 1.03 | 4.23 | 0.00 | 0.00 | 248.00 |
| *Media_Tone* | 51,398 | 12.04 | 30.26 | 0.00 | 0.00 | 100.00 |
| *Similarity* | 33,994 | 93.40 | 5.06 | 4.82 | 94.75 | 99.76 |
| *Emotion_Tone* | 37,045 | 0.27 | 0.13 | -1.00 | 0.28 | 0.80 |
| Financial Index Variables | | | | | | |
| *Cash* | 52,318 | 0.17 | 0.38 | -0.16 | 0.13 | 71.55 |
| *CapEx* | 52,251 | 5.21 | 5.51 | -11.27 | 3.52 | 74.88 |
| *EBIT* | 51,715 | -1.80 | 933.25 | -168,951.20 | 0.15 | 117,524.63 |
| *External_Financing* | 52,442 | 0.19 | 1.00 | -48.87 | 0.05 | 105.24 |
| *Growth* | 52,096 | 3.52 | 593.80 | -1.80 | 0.12 | 134,607.06 |
| *Leverage* | 52,448 | 0.51 | 4.16 | -0.19 | 0.44 | 877.26 |
| *Liquidity* | 51,604 | 2.47 | 3.80 | -5.13 | 1.55 | 204.74 |
| *MA_Intensity* | 52,459 | 158.30 | 30,160.40 | 0.00 | 0.00 | 6,891,949.88 |
| *MTB* | 49,885 | 4.10 | 69.55 | -9,101.88 | 3.10 | 2,815.19 |
| *Q* | 49,885 | 2.48 | 67.46 | 0.61 | 1.54 | 14,810.31 |
| *R&D* | 32,069 | 2.39 | 2.98 | 0.00 | 1.94 | 269.60 |
| *R&D_Dummy* | 52,459 | 0.61 | 0.49 | 0 | 1 | 1 |
| *Tangibility* | 52,448 | 0.94 | 0.09 | 0.06 | 0.96 | 1.03 |
| *ZScore* | 44,837 | 6.16 | 333.42 | -11,027.60 | 2.88 | 69,667.94 |
| Environmental Variables | | | | | | |
| *GDP_Growth* | 52,459 | 11.26 | 6.33 | -25.02 | 10.95 | 49.69 |
| *Personal_Income* | 52,459 | 4.45 | 0.28 | 3.67 | 4.51 | 4.92 |
| *Unemployment* | 52,411 | 3.16 | 0.79 | 0.80 | 3.20 | 6.50 |
| *Crisis* | 52,459 | 0.09 | 0.29 | 0 | 0 | 1 |
| *Anti_Corrupt* | 52,459 | 0.61 | 0.49 | 0 | 1 | 1 |
| *COVID19* | 52,459 | 0.18 | 0.38 | 0 | 0 | 1 |
| Industry Variables | | | | | | |
| *Adj_ROA* | 52,439 | 0.41 | 103.15 | -2,146.18 | 0.00 | 23,509.67 |
| *Aggregation* | 52,459 | 1.11 | 1.04 | 0.01 | 0.74 | 5.05 |
| *HHI* | 52,455 | 0.09 | 0.11 | 0.01 | 0.05 | 1.00 |
| *IndustryQ* | 52,450 | 1.56 | 1.04 | 0.00 | 1.29 | 14.19 |
| *IndustryQ_sq* | 52,450 | 3.51 | 6.21 | 0.00 | 1.65 | 201.30 |
| *Finance* | 52,459 | 0.02 | 0.14 | 0 | 0 | 1 |
| *High_Tech* | 52,459 | 0.36 | 0.48 | 0 | 0 | 1 |
| *Service_Trade* | 52,459 | 0.26 | 0.44 | 0 | 0 | 1 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Other Variables | | | | | | |
| *Firm_Age* | 52,459 | 15.71 | 6.71 | 0.00 | 15.00 | 121.00 |
| *Size* | 52,375 | 5.51 | 0.66 | 0.71 | 5.42 | 9.55 |
| *Family* | 49,057 | 0.53 | 0.50 | 0 | 1 | 1 |
| *SOE* | 51,706 | 0.43 | 0.50 | 0 | 0 | 1 |
| *Employee* | 51,472 | 3.26 | 0.60 | 0.00 | 3.25 | 5.74 |
| *Litigation* | 52,459 | 0.21 | 0.41 | 0 | 0 | 1 |
| *Past_Fraud* | 52,459 | 0.34 | 0.47 | 0 | 0 | 1 |
| *Stock_Return* | 46,892 | 0.21 | 0.76 | -0.92 | 0.00 | 21.53 |
| *Volume* | 50,796 | 11.34 | 25.14 | 0.00 | 5.21 | 1,235.91 |
| *Industry_Year* | 52,459 | 0.72 | 1.23 | 0.00 | 0.41 | 16.75 |
| *Province_Year* | 52,459 | 0.72 | 1.21 | 0.00 | 0.45 | 13.27 |

### Table 3: Confusion Matrix

| | | **Actual Values** | |
|---|---|---|---|
| | | Positive (Y = 1) | Negative (Y = 0) |
| **Predicted Values** | Positive (Y = 1) | True Positive (*TP*) | False Positive (*FP*) |
| | Negative (Y = 0) | False Negative (*FN*) | True Negative (*TN*) |

### Table 4: Predictive Performance

This table reports the predictive performance of the likelihood of a company engaging in any violations in the next year using the Random Forest (RF) and Gradient-Boosted Decision Tree (GBDT) models applied to the Type I and Type II samples in our study. The evaluation metrics featured in the table include the Accuracy, Precision, Recall, F1-Score, MCC (Matthew's Correlation Coefficient), and AUC (Area under the ROC) score. The results show as mean ± standard deviation.

| | | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **MCC** | **AUC** |
|---|---|---|---|---|---|---|---|
| RF | Type I Sample | **0.7647** ± 0.0051 | **0.7129** ± 0.0108 | **0.4236** ± 0.0122 | **0.5314** ± 0.0106 | **0.4111** ± 0.0114 | **0.6726** ± 0.0060 |
| | Type II Sample | **0.7913** ± 0.0040 | **0.7064** ± 0.0094 | **0.4342** ± 0.0079 | **0.5378** ± 0.0063 | **0.4332** ± 0.0080 | **0.6821** ± 0.0040 |
| GBDT | Type I Sample | **0.7810** ± 0.0041 | **0.6900** ± 0.0110 | **0.5532** ± 0.0080 | **0.6141** ± 0.0068 | **0.4693** ± 0.0094 | **0.7195** ± 0.0046 |
| | Type II Sample | **0.8034** ± 0.0030 | **0.6856** ± 0.0106 | **0.5484** ± 0.0060 | **0.6094** ± 0.0066 | **0.4855** ± 0.0087 | **0.7254** ± 0.0035 |

### Table 5: Relative Variable Importance

This table reports the importance of the input variables based on the random forest model, presenting the importance of variables that affect the likelihood of a company engaging in any violations in the next year in the Type I (left) and Type II (right) samples, respectively. Panel A reports the importance of each input variable, and Panel B reports the importance of each category of the input variables. All variables are defined in Appendix Table 1.

| **Type I Sample** | | | | **Type II Sample** | | | |
|---|---|---|---|---|---|---|---|
| ***Panel A: Importance of Each Input Variable*** | | | | | | | |
| | **Category** | **Importance** | **Cumulative Importance** | | **Category** | **Importance** | **Cumulative Importance** |
| *Past_Fraud* | Other | 30.7709% | 30.7709% | *Past_Fraud* | Other | 39.1533% | 39.1533% |
| *Disclosure* | Exposure | 9.0722% | 39.8431% | *Disclosure* | Exposure | 7.7920% | 46.9453% |
| *Adj_ROA* | Industry | 5.2672% | 45.1103% | *Anti_Corrupt* | Environmental | 4.4003% | 51.3456% |
| *Anti_Corrupt* | Environmental | 5.1708% | 50.2811% | *Adj_ROA* | Industry | 4.3028% | 55.6484% |
| *EBIT* | Financial Index | 4.6236% | 54.9047% | *Province_Year* | Other | 4.2856% | 59.9340% |
| *Opinion* | Exposure | 4.5010% | 59.4057% | *Herfindahl_5* | Ownership Structure | 3.4826% | 63.4166% |
| *Family* | Other | 3.7738% | 63.1795% | *Personal_Income* | Environmental | 3.4660% | 66.8826% |
| *SOE* | Other | 3.1579% | 66.3374% | *SOE* | Other | 3.3391% | 70.2217% |

| | | | |
|---|---|---|---|
| *Province_Year* | Other | 3.0532% | 69.3906% |
| *GDP_Growth* | Environmental | 2.0963% | 71.4869% |
| *Personal_Income* | Environmental | 1.9803% | 73.4672% |
| *Herfindahl_5* | Ownership Structure | 1.9012% | 75.3684% |
| *Inst_Own* | Ownership Structure | 1.8081% | 77.1765% |
| *Control_Own* | Ownership Structure | 1.7453% | 78.9218% |
| *Largest_Own* | Ownership Structure | 1.5380% | 80.4598% |
| *Industry_Year* | Other | 1.4386% | 81.8984% |
| *Board_Own* | Ownership Structure | 1.3033% | 83.2017% |
| *Cash* | Financial Index | 1.2927% | 84.4944% |
| *Shareholder_Meeting* | Corporate Governance | 1.0796% | 85.5740% |
| *ZScore* | Financial Index | 0.9494% | 86.5234% |
| *Volume* | Other | 0.8747% | 87.3981% |
| *Analyst* | Exposure | 0.7183% | 88.1164% |
| *Central* | Ownership Structure | 0.6526% | 88.7690% |
| *Leverage* | Financial Index | 0.6138% | 89.3828% |
| *MTB* | Financial Index | 0.5942% | 89.9770% |
| *Crisis* | Environmental | 0.5111% | 90.4881% |
| *HHI* | Industry | 0.4880% | 90.9761% |
| *Growth* | Financial Index | 0.4802% | 91.4563% |
| *External_Financing* | Financial Index | 0.4705% | 91.9268% |
| *Chair_Own* | Ownership Structure | 0.4649% | 92.3917% |
| *Tangibility* | Financial Index | 0.4214% | 92.8131% |
| *Man_Own* | Ownership Structure | 0.4184% | 93.2315% |
| *Balance* | Corporate Governance | 0.3790% | 93.6105% |
| *Board_Meeting* | Corporate Governance | 0.3636% | 93.9741% |
| *Size* | Other | 0.3591% | 94.3332% |
| *Liquidity* | Financial Index | 0.3579% | 94.6911% |
| *Employee* | Other | 0.3413% | 95.0324% |
| *Stock_Return* | Other | 0.3267% | 95.3591% |
| *COVID19* | Environmental | 0.3221% | 95.6812% |
| *Q* | Financial Index | 0.3099% | 95.9911% |
| *Unemployment* | Environmental | 0.3093% | 96.3004% |
| *Multi_Control* | Ownership Structure | 0.3013% | 96.6017% |
| *Aggregation* | Industry | 0.2687% | 96.8704% |
| *Salary* | Personnel Characteristic | 0.2671% | 97.1375% |
| *Average_Age* | Personnel Characteristic | 0.2420% | 97.3795% |
| *Salary_Gap* | Personnel Characteristic | 0.2384% | 97.6179% |
| *CapEx* | Financial Index | 0.2082% | 97.8261% |
| *Firm_Age* | Other | 0.2031% | 98.0292% |
| *Control_Dummy* | Ownership Structure | 0.1787% | 98.2079% |
| *IndustryQ_sq* | Industry | 0.1717% | 98.3796% |
| *MA_Intensity* | Financial Index | 0.1700% | 98.5496% |
| *IndustryQ* | Industry | 0.1471% | 98.6967% |
| *Differ* | Ownership Structure | 0.1377% | 98.8344% |
| *Chair_Age* | Personnel Characteristic | 0.1320% | 98.9664% |
| *Gender_Diverse* | Corporate Governance | 0.1229% | 99.0893% |
| *Board_Size* | Corporate Governance | 0.1169% | 99.2062% |
| *Litigation* | Other | 0.1069% | 99.3131% |
| *Insider_Board* | Ownership Structure | 0.0876% | 99.4007% |
| *Female_Board* | Corporate Governance | 0.0744% | 99.4751% |
| *Big4* | Exposure | 0.0678% | 99.5429% |
| *Spvisor* | Corporate Governance | 0.0548% | 99.5977% |
| *Insider_Spvisor* | Ownership Structure | 0.0545% | 99.6522% |
| *Centrality* | Corporate Governance | 0.0542% | 99.7064% |
| *Board_Ind* | Corporate Governance | 0.0488% | 99.7552% |
| *R&D_Dummy* | Financial Index | 0.0407% | 99.7959% |
| *Fin_Back* | Personnel Characteristic | 0.0348% | 99.8307% |
| *Committee* | Corporate Governance | 0.0345% | 99.8652% |
| *Overseas_Back* | Personnel Characteristic | 0.0255% | 99.8907% |
| *Big10* | Exposure | 0.0218% | 99.9125% |
| *Service_Trade* | Industry | 0.0159% | 99.9284% |
| *Duality* | Corporate Governance | 0.0133% | 99.9417% |
| *High_Tech* | Industry | 0.0106% | 99.9523% |
| *Foreign_Audit* | Exposure | 0.0088% | 99.9611% |
| *QFII* | Ownership Structure | 0.0079% | 99.9690% |
| *Qualified_IndBoard* | Corporate Governance | 0.0067% | 99.9757% |
| *Chair_Gender* | Personnel Characteristic | 0.0058% | 99.9815% |
| *Foreign_Resid* | Ownership Structure | 0.0052% | 99.9867% |
| *Elder* | Personnel Characteristic | 0.0044% | 99.9911% |
| *Even_Board* | Corporate Governance | 0.0038% | 99.9949% |
| *Finance* | Industry | 0.0024% | 99.9973% |
| *Employee_Own* | Ownership Structure | 0.0016% | 99.9989% |
| *Strong_Board* | Corporate Governance | 0.0011% | 100.0000% |

| | | | |
|---|---|---|---|
| *Opinion* | Exposure | 2.9120% | 73.1337% |
| *Industry_Year* | Other | 2.7018% | 75.8355% |
| *Family* | Other | 1.9992% | 77.8347% |
| *GDP_Growth* | Environmental | 1.9230% | 79.7577% |
| *Largest_Own* | Ownership Structure | 1.8123% | 81.5700% |
| *Control_Own* | Ownership Structure | 1.6465% | 83.2165% |
| *EBIT* | Financial Index | 1.5029% | 84.7194% |
| *Shareholder_Meeting* | Corporate Governance | 1.1677% | 85.8871% |
| *Inst_Own* | Ownership Structure | 0.9229% | 86.8100% |
| *Tangibility* | Financial Index | 0.7212% | 87.5312% |
| *ZScore* | Financial Index | 0.6694% | 88.2006% |
| *Central* | Ownership Structure | 0.6645% | 88.8651% |
| *Board_Own* | Ownership Structure | 0.6164% | 89.4815% |
| *Volume* | Other | 0.5888% | 90.0703% |
| *Salary* | Personnel Characteristic | 0.5703% | 90.6406% |
| *Cash* | Financial Index | 0.5688% | 91.2094% |
| *Board_Meeting* | Corporate Governance | 0.5559% | 91.7653% |
| *R&D_Dummy* | Financial Index | 0.4471% | 92.2124% |
| *Growth* | Financial Index | 0.4347% | 92.6471% |
| *Committee* | Corporate Governance | 0.4334% | 93.0805% |
| *MTB* | Financial Index | 0.3951% | 93.4756% |
| *Control_Dummy* | Ownership Structure | 0.3818% | 93.8574% |
| *Leverage* | Financial Index | 0.3642% | 94.2216% |
| *Employee_Own* | Ownership Structure | 0.3583% | 94.5799% |
| *HHI* | Industry | 0.3439% | 94.9238% |
| *Size* | Other | 0.3362% | 95.2600% |
| *Man_Own* | Ownership Structure | 0.2976% | 95.5576% |
| *Stock_Return* | Other | 0.2721% | 95.8297% |
| *Unemployment* | Environmental | 0.2661% | 96.0958% |
| *Aggregation* | Industry | 0.2585% | 96.3543% |
| *Liquidity* | Financial Index | 0.2550% | 96.6093% |
| *COVID19* | Environmental | 0.2463% | 96.8556% |
| *Salary_Gap* | Personnel Characteristic | 0.2219% | 97.0775% |
| *Employee* | Other | 0.2204% | 97.2979% |
| *Q* | Financial Index | 0.2156% | 97.5135% |
| *Firm_Age* | Other | 0.2134% | 97.7269% |
| *Analyst* | Exposure | 0.1796% | 97.9065% |
| *Chair_Own* | Ownership Structure | 0.1738% | 98.0803% |
| *Balance* | Corporate Governance | 0.1711% | 98.2514% |
| *External_Financing* | Financial Index | 0.1707% | 98.4221% |
| *Average_Age* | Personnel Characteristic | 0.1593% | 98.5814% |
| *Multi_Control* | Ownership Structure | 0.1266% | 98.7080% |
| *IndustryQ_sq* | Industry | 0.1175% | 98.8255% |
| *CapEx* | Financial Index | 0.1107% | 98.9362% |
| *Insider_Board* | Ownership Structure | 0.1087% | 99.0449% |
| *IndustryQ* | Industry | 0.0974% | 99.1423% |
| *Fin_Back* | Personnel Characteristic | 0.0880% | 99.2303% |
| *Differ* | Ownership Structure | 0.0846% | 99.3149% |
| *Chair_Age* | Personnel Characteristic | 0.0828% | 99.3977% |
| *Litigation* | Other | 0.0748% | 99.4725% |
| *MA_Intensity* | Financial Index | 0.0690% | 99.5415% |
| *Crisis* | Environmental | 0.0584% | 99.5999% |
| *Board_Size* | Corporate Governance | 0.0554% | 99.6553% |
| *Board_Ind* | Corporate Governance | 0.0523% | 99.7076% |
| *Gender_Diverse* | Corporate Governance | 0.0518% | 99.7594% |
| *Female_Board* | Corporate Governance | 0.0447% | 99.8041% |
| *Centrality* | Corporate Governance | 0.0372% | 99.8413% |
| *Insider_Spvisor* | Ownership Structure | 0.0278% | 99.8691% |
| *Big4* | Exposure | 0.0219% | 99.8910% |
| *Spvisor* | Corporate Governance | 0.0211% | 99.9121% |
| *Overseas_Back* | Personnel Characteristic | 0.0204% | 99.9325% |
| *Big10* | Exposure | 0.0110% | 99.9435% |
| *High_Tech* | Industry | 0.0090% | 99.9525% |
| *Duality* | Corporate Governance | 0.0088% | 99.9613% |
| *Service_Trade* | Industry | 0.0077% | 99.9690% |
| *Foreign_Audit* | Exposure | 0.0067% | 99.9757% |
| *Even_Board* | Corporate Governance | 0.0050% | 99.9807% |
| *Elder* | Personnel Characteristic | 0.0044% | 99.9851% |
| *Finance* | Industry | 0.0042% | 99.9893% |
| *Strong_Board* | Corporate Governance | 0.0034% | 99.9927% |
| *Qualified_IndBoard* | Corporate Governance | 0.0033% | 99.9960% |
| *QFII* | Ownership Structure | 0.0023% | 99.9983% |
| *Chair_Gender* | Personnel Characteristic | 0.0015% | 99.9998% |
| *Foreign_Resid* | Ownership Structure | 0.0002% | 100.0000% |

| *Panel B: Importance of Each Category of Input Variables* | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Ownership Structure | Corporate Governance | Personnel Characteristic | Exposure | Financial Index | Environmental | Industry | Other |
| Type I Sample | 10.6063% | 2.3536% | 0.9500% | 14.3899% | 10.5325% | 10.3899% | 6.3716% | 44.4062% |
| Type II Sample | 10.7069% | 2.6111% | 1.1486% | 10.9232% | 5.9244% | 10.3601% | 5.1410% | 53.1847% |

### Table 6: Predictive Performance for Different Types of Violations

This table reports the predictive performance of the Random Forest (RF) and Gradient-Boosted Decision Tree (GBDT) models applied to the Type I and Type II samples in our study. The evaluation metrics featured in the table include the Accuracy, Precision, Recall, F1-Score, MCC (Matthew's Correlation Coefficient), and AUC (Area under the ROC) score. Panels A - C compare the predictive performance of the likelihood of a company engaging in different types of violations in the next year, namely accounting irregularities, operation irregularities, and leadership irregularities, respectively. The results show as mean ± standard deviation.

| | | Accuracy | Precision | Recall | F1-Score | MCC | AUC |
|---|---|---|---|---|---|---|---|
| ***Panel A: Accounting Fraud*** | | | | | | | |
| RF | Type I Sample | **0.8238** ± 0.0014 | **0.7239** ± 0.0280 | **0.2702** ± 0.0092 | **0.3935** ± 0.0060 | **0.3672** ± 0.0062 | **0.6212** ± 0.0027 |
| | Type II Sample | **0.8365** ± 0.0038 | **0.7112** ± 0.0224 | **0.2634** ± 0.0106 | **0.3844** ± 0.0146 | **0.3640** ± 0.0176 | **0.6188** ± 0.0062 |
| GBDT | Type I Sample | **0.8288** ± 0.0015 | **0.6632** ± 0.0108 | **0.3860** ± 0.0065 | **0.4880** ± 0.0050 | **0.4145** ± 0.0050 | **0.6668** ± 0.0028 |
| | Type II Sample | **0.8378** ± 0.0044 | **0.6392** ± 0.0170 | **0.3742** ± 0.0129 | **0.4721** ± 0.0149 | **0.4031** ± 0.0174 | **0.6617** ± 0.0075 |
| ***Panel B: Operation Fraud*** | | | | | | | |
| RF | Type I Sample | **0.9500** ± 0.0014 | **1.0000** ± 0.0000 | **0.0035** ± 0.0014 | **0.0069** ± 0.0028 | **0.0566** ± 0.0117 | **0.5017** ± 0.0007 |
| | Type II Sample | **0.9566** ± 0.0007 | **1.0000** ± 0.0000 | **0.0040** ± 0.0012 | **0.0080** ± 0.0024 | **0.0617** ± 0.0093 | **0.5020** ± 0.0006 |
| GBDT | Type I Sample | **0.9503** ± 0.0013 | **0.5391** ± 0.0236 | **0.0692** ± 0.0051 | **0.1227** ± 0.0098 | **0.1801** ± 0.0225 | **0.5330** ± 0.0029 |
| | Type II Sample | **0.9562** ± 0.0009 | **0.4842** ± 0.0224 | **0.0675** ± 0.0123 | **0.1184** ± 0.0197 | **0.1685** ± 0.0204 | **0.5321** ± 0.0060 |
| ***Panel C: Leadership Fraud*** | | | | | | | |
| RF | Type I Sample | **0.9262** ± 0.0027 | **0.9568** ± 0.0391 | **0.0517** ± 0.0118 | **0.0982** ± 0.0212 | **0.2119** ± 0.0229 | **0.5258** ± 0.0058 |
| | Type II Sample | **0.9374** ± 0.0037 | **0.9560** ± 0.0007 | **0.0462** ± 0.0034 | **0.0882** ± 0.0061 | **0.2029** ± 0.0079 | **0.5230** ± 0.0017 |
| GBDT | Type I Sample | **0.9272** ± 0.0039 | **0.6841** ± 0.0477 | **0.1149** ± 0.0160 | **0.1967** ± 0.0257 | **0.2605** ± 0.0320 | **0.5552** ± 0.0084 |
| | Type II Sample | **0.9390** ± 0.0035 | **0.7212** ± 0.0254 | **0.1123** ± 0.0116 | **0.1943** ± 0.0164 | **0.2687** ± 0.0093 | **0.5546** ± 0.0055 |

### Table 7: Relative Variable Importance for Different Types of Violations

This table reports the importance of the input variables based on the random forest model, presenting the importance of variables for each category of accounting irregularities, operation irregularities, and leadership irregularities (i.e. variables related to corporate ownership structure, corporate governance, characteristics of relevant personnel (directors, supervisors, and executives), information exposure, corporate financial characteristics, industry characteristics, environmental characteristics as well as other relevant variables). Panels A and B report the importance of each category of input variables in the Type I and Type II samples, respectively. The importance of each input variable can be found in Appendix Table 3, and all variables are defined in Appendix Table 1.

| | Ownership Structure | Corporate Governance | Personnel Characteristic | Exposure | Financial Index | Environmental | Industry | Other |
|---|---|---|---|---|---|---|---|---|
| ***Panel A: Type I Sample*** | | | | | | | | |
| Accounting Fraud | 8.6960% | 2.8092% | 1.1953% | 17.0899% | 12.3546% | 4.8520% | 8.0796% | 44.9234% |
| Operation Fraud | 10.9186% | 8.0708% | 3.2995% | 14.4722% | 19.5280% | 3.3997% | 8.4248% | 31.8864% |
| Leadership Fraud | 24.6955% | 5.9531% | 2.9896% | 5.2466% | 12.3728% | 15.4867% | 15.3953% | 17.8604% |
| ***Panel B: Type II Sample*** | | | | | | | | |
| Accounting Fraud | 8.9218% | 2.1209% | 1.2334% | 14.5957% | 8.8068% | 5.4201% | 7.2471% | 51.6542% |
| Operation Fraud | 8.8053% | 6.4062% | 2.9991% | 13.0006% | 15.1042% | 4.5121% | 9.1856% | 39.9869% |
| Leadership Fraud | 19.3261% | 5.0182% | 3.5158% | 4.1906% | 9.2305% | 20.7339% | 15.6935% | 22.2914% |

### Table 8: Comparison of the Predictive Performance by Ownership Property

This table compares the predictive performance of listed companies with different ownership properties (i.e. state-owned and non-state-owned enterprises) for the likelihood of any violations occurring in the next year. The evaluation metrics featured in the table include the Accuracy, Precision, Recall, F1-Score, MCC (Matthew's Correlation Coefficient), and AUC (Area under the ROC) score. Panels A and B report the predictive performance of the Random Forest (RF) and Gradient-Boosted Decision Tree (GBDT) models applied to the Type I and Type II samples, respectively.

| | | Accuracy | Precision | Recall | F1-Score | MCC | AUC |
|---|---|---|---|---|---|---|---|
| ***Panel A: Type I Sample*** | | | | | | | |
| **SOE** | RF | 0.8097 | 0.7158 | 0.2372 | 0.3563 | 0.3348 | 0.6052 |
| | GBDT | 0.8186 | 0.6398 | 0.4191 | 0.5065 | 0.4147 | 0.6759 |
| **Non-SOE** | RF | 0.7358 | 0.7081 | 0.5056 | 0.5900 | 0.4154 | 0.6900 |
| | GBDT | 0.7506 | 0.6969 | 0.5957 | 0.6423 | 0.4560 | 0.7198 |
| ***Panel B: Type II Sample*** | | | | | | | |
| **SOE** | RF | 0.8276 | 0.7233 | 0.2265 | 0.3449 | 0.3379 | 0.6024 |
| | GBDT | 0.8410 | 0.6704 | 0.4065 | 0.5061 | 0.4366 | 0.6782 |
| **Non-SOE** | RF | 0.7612 | 0.7123 | 0.5187 | 0.6003 | 0.4471 | 0.7040 |
| | GBDT | 0.7747 | 0.7018 | 0.6051 | 0.6499 | 0.4880 | 0.7347 |

### Table 9: Comparison of the Predictive Performance by Information Disclosure Quality

This table compares the predictive performance of listed companies with different information disclosure situations (i.e., the evaluation result of corporate information disclosure is excellent (the highest quality), good, qualified, and unqualified or missing (the lowest quality)) for the likelihood of any violations occurring in the next year. The evaluation metrics featured in the table include the Accuracy, Precision, Recall, F1-Score, MCC (Matthew's Correlation Coefficient), and AUC (Area under the ROC) score. Panels A and B report the predictive performance of the Random Forest (RF) and Gradient-Boosted Decision Tree (GBDT) models applied to the Type I and Type II samples, respectively.

| | | Accuracy | Precision | Recall | F1-Score | MCC | AUC |
|---|---|---|---|---|---|---|---|
| ***Panel A: Type I Sample*** | | | | | | | |
| **Excellent** | RF | 0.8202 | 0.6019 | 0.1191 | 0.1986 | 0.2068 | 0.5501 |
| | GBDT | 0.8088 | 0.4718 | 0.2057 | 0.2865 | 0.2172 | 0.5766 |
| **Good** | RF | 0.7369 | 0.6772 | 0.3390 | 0.4518 | 0.3344 | 0.6315 |
| | GBDT | 0.7504 | 0.6537 | 0.4671 | 0.5449 | 0.3896 | 0.6754 |
| **Qualified** | RF | 0.6849 | 0.6848 | 0.7622 | 0.7214 | 0.3643 | 0.6792 |
| | GBDT | 0.6795 | 0.6816 | 0.7526 | 0.7154 | 0.3531 | 0.6741 |
| **Unqualified or Missing** | RF | 0.7980 | 0.7260 | 0.4354 | 0.5443 | 0.4476 | 0.6862 |
| | GBDT | 0.8065 | 0.6888 | 0.5507 | 0.6121 | 0.4907 | 0.7277 |
| ***Panel B: Type II Sample*** | | | | | | | |
| **Excellent** | RF | 0.8354 | 0.7469 | 0.1079 | 0.1885 | 0.2402 | 0.5500 |
| | GBDT | 0.8328 | 0.5621 | 0.2513 | 0.3474 | 0.2957 | 0.6046 |
| **Good** | RF | 0.7576 | 0.7174 | 0.3400 | 0.4614 | 0.3681 | 0.6406 |
| | GBDT | 0.7690 | 0.6696 | 0.4807 | 0.5597 | 0.4191 | 0.6882 |
| **Qualified** | RF | 0.6959 | 0.6711 | 0.7097 | 0.6898 | 0.3930 | 0.6966 |
| | GBDT | 0.7083 | 0.6785 | 0.7366 | 0.7064 | 0.4190 | 0.7096 |
| **Unqualified or Missing** | RF | 0.8325 | 0.7283 | 0.4187 | 0.5317 | 0.4635 | 0.6864 |
| | GBDT | 0.8386 | 0.6862 | 0.5325 | 0.5997 | 0.5069 | 0.7305 |

### Table 10: Comparison of the Predictive Performance by Past Violations

This table compares the predictive performance of listed companies that have engaged in different numbers of violations within a year (i.e., 0, 1, and 2 or more, respectively) for the likelihood of any violations occurring in the next year. The evaluation metrics featured in the table include the Accuracy, Precision, Recall, F1-Score, MCC (Matthew's Correlation Coefficient), and AUC

(Area under the ROC) score. Panels A and B report the predictive performance of the Random Forest (RF) and Gradient-Boosted Decision Tree (GBDT) models applied to the Type I and Type II samples, respectively.

| | | Accuracy | Precision | Recall | F1-Score | MCC | AUC |
|---|---|---|---|---|---|---|---|
| ***Panel A: Type I Sample*** | | | | | | | |
| **Number = 0** | RF | 0.8302 | 0.8043 | 0.0954 | 0.1706 | 0.2392 | 0.5451 |
| | GBDT | 0.8489 | 0.7223 | 0.2829 | 0.4066 | 0.3876 | 0.6293 |
| **Number = 1** | RF | 0.6158 | 0.6153 | 0.7932 | 0.6930 | 0.2120 | 0.5970 |
| | GBDT | 0.6200 | 0.6344 | 0.7225 | 0.6756 | 0.2256 | 0.6098 |
| **Number ≥ 2** | RF | 0.7292 | 0.7570 | 0.8953 | 0.8203 | 0.3037 | 0.6271 |
| | GBDT | 0.7220 | 0.7565 | 0.8808 | 0.8140 | 0.2900 | 0.6245 |
| ***Panel B: Type II Sample*** | | | | | | | |
| **Number = 0** | RF | 0.8547 | 0.8560 | 0.0898 | 0.1625 | 0.2484 | 0.5435 |
| | GBDT | 0.8706 | 0.7509 | 0.2631 | 0.3896 | 0.3938 | 0.6234 |
| **Number = 1** | RF | 0.6114 | 0.6188 | 0.8009 | 0.6982 | 0.1947 | 0.5872 |
| | GBDT | 0.6189 | 0.6375 | 0.7393 | 0.6846 | 0.2145 | 0.6031 |
| **Number ≥ 2** | RF | 0.7269 | 0.7569 | 0.8963 | 0.8207 | 0.2835 | 0.6171 |
| | GBDT | 0.7221 | 0.7572 | 0.8853 | 0.8163 | 0.2755 | 0.6164 |

## References


Achakzai, M.A.K., Peng, J., 2023. Detecting financial statement fraud using dynamic ensemble machine learning. *International Review of Financial Analysis* 89, 102827.

Adams, R.B., Funk, P., 2012. Beyond the glass ceiling: Does gender matter? *Management Science* 58, 219–235.

Agrawal, A., Chadha, S., 2005. Corporate governance and accounting scandals. *The Journal of Law and Economics* 48, 371–406.

Alexander, C.R., Cohen, M.A., 1999. Why do corporations become criminals? Ownership, hidden actions, and crime as an agency cost. *Journal of Corporate Finance* 5, 1–34.

Alok, S., Ayyagari, M., 2020. Politics, state ownership, and corporate investments. *The Review of Financial Studies* 33, 3031–3087.

Ananchotikul, N., Kouwenberg, R., Phunnarungsi, V., 2010. Do firms decouple corporate governance policy and practice? *European Financial Management* 16, 712–737.

Anand, V., Tina Dacin, M., Murphy, P.R., 2015. The continued need for diversity in fraud research. *Journal of Business Ethics* 131, 751–755.

Armstrong, C.S., Balakrishnan, K., Cohen, D., 2012. Corporate governance and the information environment: Evidence from state antitakeover laws. *Journal of Accounting and Economics* 53, 185–204.

Bantel, K.A., Jackson, S.E., 1989. Top management and innovations in banking: Does the composition of the top team make a difference? *Strategic Management Journal* 10, 107–124.

Bao, Y., Ke, B., Li, B., Yu, Y.J., Zhang, J., 2020. Detecting accounting fraud in publicly traded us firms using a machine learning approach. *Journal of Accounting Research* 58, 199–235.

Bear, S., Rahman, N., Post, C., 2010. The impact of board diversity and gender composition on corporate social responsibility and firm reputation. *Journal of Business Ethics* 97, 207–221.

Beasley, M.S., 1996. An empirical analysis of the relation between the board of director composition and financial statement fraud. *The Accounting Review*, 443–465.

Becker, G.S., 1968. Crime and punishment: An economic approach, in: *The Economic Dimensions of Crime*. Springer, pp. 13–68.

Beneish, M.D., 1999. Incentives and penalties related to earnings overstatements that violate GAAP. *The Accounting Review* 74, 425–457.

Benmelech, E., Frydman, C., 2015. Military CEOs. *Journal of Financial Economics* 117, 43–59.

Bertomeu, J., Cheynel, E., Floyd, E., Pan, W., 2021. Using machine learning to detect misstatements. *Review of Accounting Studies* 26, 468–519.

Bertrand, M., Mullainathan, S., 2003. Enjoying the quiet life? Corporate governance and managerial preferences. *Journal of Political Economy* 111, 1043–1075.

Biggerstaff, L., Cicero, D.C., Puckett, A., 2015. Suspect CEOs, unethical culture, and corporate misbehavior. *Journal of Financial Economics* 117, 98–121.

Bologna, J., Lindquist, R.J., Wells, J.T., 1993. *The Accountant's Handbook of Fraud and Commercial Crime*. Wiley New York, NY.

Brochet, F., Srinivasan, S., 2014. Accountability of independent directors: Evidence from firms subject to securities litigation. *Journal of Financial Economics* 111, 430–449.

Brown, N.C., Crowley, R.M., Elliott, W.B., 2020. What are you saying? Using topic to detect financial misreporting. *Journal of Accounting Research* 58, 237–291.

Burns, N., Kedia, S., 2006. The impact of performance-based compensation on misreporting. *Journal of Financial Economics* 79, 35–67.

Burns, N., Kedia, S., Lipson, M., 2010. Institutional ownership and monitoring: Evidence from financial misreporting. *Journal of Corporate Finance* 16, 443–455.

Byun, H.S., Kim, W., Lee, E.J., Park, K.S., 2019. When and why do takeovers lead to fraud? *Financial Management* 48, 45–76.

Cai, Z., Wu, S., 2007. The empirical study on behaviour of committing fraud in listed companies: Evidence from the characteristics of board of directors. *Nankai Business Review* 10, 62–68.

Carcello, J.V., Hermanson, D.R., Neal, T.L., Riley Jr, R.A., 2002. Board characteristics and audit fees. *Contemporary Accounting Research* 19, 365–384.

Cecchini, M., Aytug, H., Koehler, G.J., Pathak, P., 2010. Detecting management fraud in public companies. *Management Science* 56, 1146–1160.

Chen, C.J., Li, Z., Su, X., Sun, Z., 2011. Rent-seeking incentives, corporate political connections, and the control structure of private firms: Chinese evidence. *Journal of Corporate Finance* 17, 229–243.

Chen, D., Chen, Y., Li, O.Z., Ni, C., 2018. Foreign residency rights and corporate fraud. *Journal of Corporate Finance* 51, 142–163.

Chen, G., Firth, M., Gao, D.N., Rui, O.M., 2006. Ownership structure, corporate governance, and fraud: Evidence from China. *Journal of Corporate Finance* 12, 424–448.

Chen, J., Cumming, D., Hou, W., Lee, E., 2013. Executive integrity, audit opinion, and fraud in Chinese listed firms. *Emerging Markets Review* 15, 72–91.

Chen, J., Cumming, D., Hou, W., Lee, E., 2016. Does the external monitoring effect of financial analysts deter corporate fraud in China? *Journal of Business Ethics* 134, 727–742.

Chen, Z., 2020. Research on financial irregularities of listed companies based on decision tree and random forest algorithm. *Contemporary Finance*, 23–25.

Clifford, C.P., Ellis, J.A., Gerken, W.C., 2018. Hedge fund boards and the market for independent directors. *Journal of Financial and Quantitative Analysis* 53, 2067–2101.

Cline, B.N., Walkling, R.A., Yore, A.S., 2018. The consequences of managerial indiscretions: Sex, lies, and firm value. *Journal of Financial Economics* 127, 389–415.

Coles, J.L., Daniel, N.D., Naveen, L., 2014. Co-opted boards. *The Review of Financial Studies* 27, 1751–1796.

Conyon, M.J., He, L., 2016. Executive compensation and corporate fraud in China. *Journal of Business Ethics* 134, 669–691.

Cowen, A.P., Marcel, J.J., 2011. Damaged goods: Board decisions to dismiss reputationally compromised directors. *Academy of Management Journal* 54, 509–527.

Cumming, D., Dannhauser, R., Johan, S., 2015a. Financial market misconduct and agency conflicts: A synthesis and future directions. *Journal of Corporate Finance* 34, 150–168.

Cumming, D., Groh, A.P., Johan, S., 2018a. Same rules, different enforcement: Market abuse in Europe. *Journal of International Financial Markets, Institutions and Money* 54, 130–151.

Cumming, D., Johan, S., Peter, R., 2018b. Developments in financial institutions, governance, agency costs, and misconduct. *Journal of International Financial Markets, Institutions and Money* 54, 1–14.

Cumming, D., Leung, T.Y., Rui, O., 2015b. Gender diversity and securities fraud. *Academy of Management Journal* 58, 1572–1593.

Dalton, D.R., Daily, C.M., Johnson, J.L., Ellstrand, A.E., 1999. Number of directors and financial performance: A meta-analysis. *Academy of Management Journal* 42, 674–686.

Davidson, R., Dey, A., Smith, A., 2015. Executives' "off-the-job" behavior, corporate culture, and financial reporting risk. *Journal of Financial Economics* 117, 5–28.

De Roux, D., Perez, B., Moreno, A., Villamil, M.d.P., Figueroa, C., 2018. Tax fraud detection for under-reporting declarations using an unsupervised machine learning approach, in: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 215–222.

Dechow, P.M., Ge, W., Larson, C.R., Sloan, R.G., 2011. Predicting material accounting misstatements. *Contemporary Accounting Research* 28, 17–82.

Del Guercio, D., Seery, L., Woidtke, T., 2008. Do boards pay attention when institutional investor activists "just vote no"? *Journal of Financial Economics* 90, 84–103.

Djankov, S., La Porta, R., Lopez-de Silanes, F., Shleifer, A., 2008. The law and economics of self-dealing. *Journal of Financial Economics* 88, 430–465.

Doidge, C., Karolyi, G.A., Stulz, R.M., 2007. Why do countries matter so much for corporate governance? *Journal of Financial Economics* 86, 1–39.

Dong, W., Liao, S., Zhang, Z., 2018. Leveraging financial social media data for corporate fraud detection. *Journal of Management Information Systems* 35, 461–487.

Dyck, A., Morse, A., Zingales, L., 2010. Who blows the whistle on corporate fraud? *The Journal of Finance* 65, 2213–2253.

Eagly, A.H., Crowley, M., 1986. Gender and helping behavior: A meta-analytic review of the social psychological literature. *Psychological Bulletin* 100, 283.

Erhardt, N.L., Werbel, J.D., Shrader, C.B., 2003. Board of director diversity and firm financial performance. *Corporate Governance: An International Review* 11, 102–111.

Fama, E.F., Jensen, M.C., 1983. Separation of ownership and control. *The Journal of Law and Economics* 26, 301–325.

Fang, L., Peress, J., 2009. Media coverage and the cross-section of stock returns. *The Journal of Finance* 64, 2023–2052.

Fanning, K.M., Cogger, K.O., 1998. Neural network detection of management fraud using published financial data. *Intelligent Systems in Accounting, Finance & Management* 7, 21–41.

Feng, X., Johansson, A.C., 2018. Living through the great Chinese famine: Early-life experiences and managerial decisions. *Journal of Corporate Finance* 48, 638–657.

Firth, M., Rui, O.M., Wu, W., 2011. Cooking the books: Recipes and costs of falsified financial statements in china. *Journal of Corporate Finance* 17, 371–390.

Fisman, R., Wang, Y., 2015. The mortality cost of political connections. *The Review of Economic Studies* 82, 1346–1382.

Fitzpatrick, T., Mues, C., 2016. An empirical comparison of classification algorithms for mortgage default prediction: Evidence from a distressed mortgage market. *European Journal of Operational Research* 249, 427–439.

Fombrun, C., Van Riel, C., 1997. The reputational landscape. *Corporate Reputation Review*, 1–16.

Francis, J.R., 2004. What do we know about audit quality? *The British Accounting Review* 36, 345–368.

Free, C., Murphy, P.R., 2015. The ties that bind: The decision to co-offend in fraud. *Contemporary Accounting Research* 32, 18–54.

Galeshchuk, S., Mukherjee, S., 2017. Deep networks for predicting direction of change in foreign exchange rates. *Intelligent Systems in Accounting, Finance and Management* 24, 100–110.

Gam, Y.K., Gupta, P., Im, J., Shin, H., 2021. Evasive shareholder meetings and corporate fraud. *Journal of Corporate Finance* 66, 101807.

Gande, A., Lewis, C.M., 2009. Shareholder-initiated class action lawsuits: Shareholder wealth effects and industry spillovers. *Journal of Financial and Quantitative Analysis* 44, 823–850.

Gelman, M., Khan, Z., Shoham, A., Tarba, S.Y., 2021. Does local competition and firm market power affect investment adviser misconduct? *Journal of Corporate Finance* 66, 101810.

Giannetti, M., Wang, T.Y., 2016. Corporate scandals and household stock market participation. *The Journal of Finance* 71, 2591–2636.

Gilligan, C., 1977. In a different voice: Women's conceptions of self and of morality. *Harvard Educational Review* 47, 481–517.

Goodell, J.W., Li, M., Liu, D., 2021. Price informativeness and state-owned enterprises: Considering their heterogeneity. *International Review of Financial Analysis* 76, 101783.

Gotsi, M., Wilson, A.M., 2001. Corporate reputation: Seeking a definition. *Corporate Communications: An International Journal* 6(1), 24–30.

Gow, I.D., Kaplan, S.N., Larcker, D.F., Zakolyukina, A.A., 2016. CEO personality and firm policies (No. w22435). *National Bureau of Economic Research*.

Greve, H.R., Palmer, D., Pozner, J.E., 2010. Organizations gone wild: The causes, processes, and consequences of organizational misconduct. *Academy of Management Annals* 4, 53–107.

Griffin, J.M., Maturana, G., 2016. Who facilitated misreporting in securitized loans? *The Review of Financial Studies* 29, 384–419.

Gu, N., Feng, Q., 2009. The empirical studying on detecting the fraudulent financial statements based on LVQ neural network. *Value Engineering* 28, 111–113.

Gu, S., Kelly, B., Xiu, D., 2020. Empirical asset pricing via machine learning. *The Review of Financial Studies* 33, 2223–2273.

Hagendorff, J., Le, N., Nguyen, D.D., 2021. The walls have ears: Local information environments and corporate fraud. *Journal of Money, Credit and Banking* 54(8), 2377–2410.

Hajek, P., Henriques, R., 2017. Mining corporate annual reports for intelligent detection of financial statement fraud–A comparative study of machine learning methods. *Knowledge-Based Systems* 128, 139–152.

Hambrick, D.C., Mason, P.A., 1984. Upper echelons: The organization as a reflection of its top managers. *Academy of Management Review* 9, 193–206.

Harford, J., Kecsk´es, A., Mansi, S., 2018. Do long-term investors improve corporate decision making? *Journal of Corporate Finance* 50, 424–452.

Harris, J., Bromiley, P., 2007. Incentives to cheat: The influence of executive compensation and firm performance on financial misrepresentation. *Organization Science* 18, 350–367.

Hartzell, J.C., Starks, L.T., 2003. Institutional investors and executive compensation. *The Journal of Finance* 58, 2351–2374.

Haß, L.H., M¨uller, M.A., Vergauwe, S., 2015. Tournament incentives and corporate fraud. *Journal of Corporate Finance* 34, 251–267.

Hastie, T., Tibshirani, R., Friedman, J., 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: springer.

He, L., Conyon, M.J., Chen, J., 2020. Signalling via earnings downgrades: How reputational concerns shape analyst responses to corporate fraud? *Corporate Governance: An International Review* 28, 240–263.

He, W., Luo, J.h., 2018. Agency problems in firms with an even number of directors: Evidence from China. *Journal of Banking & Finance* 93, 139–150.

Heese, J., P´erez-Cavazos, G., Peter, C.D., 2022. When the local newspaper leaves town: The effects of local newspaper closures on corporate misconduct. *Journal of Financial Economics* 145, 445–463.

Hill, C.W., Jones, T.M., 1992. Stakeholder-agency theory. *Journal of Management Studies* 29, 131–154.

Hoberg, G., Lewis, C., 2017. Do fraudulent firms produce abnormal disclosure? *Journal of Corporate Finance* 43, 58–85.

Hong, H., Lim, T., Stein, J.C., 2000. Bad news travels slowly: Size, analyst coverage, and the profitability of momentum strategies. *The Journal of Finance* 55, 265–295.

Hou, W., Moore, G., 2010. Player and referee roles held jointly: The effect of state ownership on China's regulatory enforcement against fraud. *Journal of Business Ethics* 95, 317–335.

Huang, J., Kisgen, D.J., 2013. Gender and corporate finance: Are male executives overconfident relative to female executives? *Journal of Financial Economics* 108, 822–839.

Huang, Z., Liu, J., Lin, C., 2020. A comparative study on the cutting-edge methods in identifying financial statement fraud of listed companies based on machine learning. *Journal of Systems Science and Mathematical Sciences* 40, 1882–1900.

Ismail, G.M., Huseynov, F., Jain, P.K., McInish, T.H., 2021. Brand equity, earnings management, and financial reporting irregularities. *The Review of Corporate Finance Studies* 10, 402–435.

Jensen, M.C., 1993. The modern industrial revolution, exit, and the failure of internal control systems. *The Journal of Finance* 48, 831–880.

Jensen, M.C., Meckling, W.H., 2019. Theory of the firm: Managerial behavior, agency costs and ownership structure, in: *Corporate Governance*, 77–132. Gower.

Jiang, G., Lee, C.M., Yue, H., 2010. Tunneling through intercorporate loans: The China experience. *Journal of Financial Economics* 98, 1–20.

Jiang, L., Levine, R., Lin, C., 2016. Competition and bank opacity. *The Review of Financial Studies* 29, 1911–1942.

Johnson, S., Boone, P., Breach, A., Friedman, E., 2000. Corporate governance in the Asian financial crisis. *Journal of Financial Economics* 58, 141–186.

Karpoff, J.M., 2021. The future of financial fraud. *Journal of Corporate Finance* 66, 101694.

Karpoff, J.M., Koester, A., Lee, D.S., Martin, G.S., 2017. Proxies and databases in financial misconduct research. *The Accounting Review* 92, 129–163.

Karpoff, J.M., Lee, D.S., Martin, G.S., 2008. The cost to firms of cooking the books. *Journal of Financial and Quantitative Analysis* 43, 581–611.

Kedia, S., Philippon, T., 2009. The economics of fraudulent accounting. *The Review of Financial Studies* 22, 2169–2199.

Khanna, V., Kim, E.H., Lu, Y., 2015. CEO connectedness and corporate fraud. *The Journal of Finance* 70, 1203–1252.

Kim, E.H., Lu, Y., 2011. CEO ownership, external governance, and risk-taking. *Journal of Financial Economics* 102, 272–292.

Kizilaslan, A., Manakyan Mathers, A., 2014. Strategic credit line usage and performance. *Journal of Financial Research* 37, 243–265.

Kuang, Y.F., Lee, G., 2017. Corporate fraud and external social connectedness of independent directors. *Journal of Corporate Finance* 45, 401–427.

Lee, J., Lee, K.J., Nagarajan, N.J., 2014. Birds of a feather: Value implications of political alignment between top management and directors. *Journal of Financial Economics* 112, 232–250.

Li, F., 2010. The information content of forward-looking statements in corporate filings—A naïve Bayesian machine learning approach. *Journal of Accounting Research* 48, 1049–1102.

Li, J., Yu, L., Mei, X., Feng, X., 2022. Do social media constrain or promote company violations? *Accounting & Finance* 62, 31–70.

Li, K., Mai, F., Shen, R., Yan, X., 2021a. Measuring corporate culture using machine learning. *The Review of Financial Studies* 34, 3265–3315.

Li, X., Pukthuanthong, K., Walker, M.G., Walker, T.J., 2016. The determinants of IPO-related shareholder litigation: The role of CEO equity incentives and corporate governance. *Journal of Financial Markets* 31, 81–126.

Li, Y., Shi, H., Zhou, Y., 2021b. The influence of the media on government decisions: Evidence from IPOs in China. *Journal of Corporate Finance* 70, 102056.

Lin, C., Morck, R., Yeung, B.Y., Zhao, X., 2016. Anti-corruption reforms and shareholder valuations: Event study evidence from China. *Available at SSRN 2729087*.

Liu, C., 2018. Are women greener? Corporate gender diversity and environmental violations. *Journal of Corporate Finance* 52, 118–142.

Liu, X., 2016. Corruption culture and corporate misconduct. *Journal of Financial Economics* 122, 307–327.

Lokanan, M., Liu, S., 2021. Predicting fraud victimization using classical machine learning. *Entropy* 23, 300.

Lu, Y., Zhu, Y., Hu, X., 2012. Institutional shareholding and corporate fraud: Evidence from China. *Nankai Business Review* 15, 13–23.

Mackenzie, C., Rees, W., Rodionova, T., 2013. Do responsible investment indices improve corporate social responsibility? ftse4good's impact on environmental management. *Corporate Governance: An International Review* 21, 495–512.

Malmendier, U., Tate, G., Yan, J., 2011. Overconfidence and early-life experiences: The effect of managerial traits on corporate financial policies. *The Journal of Finance* 66, 1687–1733.

Markarian, G., Santalo, J., 2014. Product market competition, information and earnings management. *Journal of Business Finance & Accounting* 41, 572–599.

Miller, G.S., 2006. The press as a watchdog for accounting fraud. *Journal of Accounting Research* 44, 1001–1033.

Miller, G.S., Skinner, D.J., 2015. The evolving disclosure landscape: How changes in technology, the media, and capital markets are affecting disclosure. *Journal of Accounting Research* 53, 221–239.

Munkhdalai, L., Munkhdalai, T., Namsrai, O.E., Lee, J.Y., Ryu, K.H., 2019. An empirical comparison of machine-learning methods on bank client credit assessments. *Sustainability* 11, 699.

Ning, C., McGuinness, P.B., Chao, X., 2021. Does securities enforcement improve disclosure quality? An examination of Chinese listed companies' restatement activities. *Journal of Corporate Finance* 67, 101877.

Pagano, M., Immordino, G., 2012. Corporate fraud, governance, and auditing. *The Review of Corporate Finance Studies* 1, 109–133.

Pai, P.F., Hsu, M.F., Wang, M.C., 2011. A support vector machine-based model for detecting top management fraud. *Knowledge-Based Systems* 24, 314–321.

Peng, L., Roell, A., 2014. Managerial incentives and stock price manipulation. *The Journal of Finance* 69, 487–526.

Perols, J., 2011. Financial statement fraud detection: An analysis of statistical and machine learning algorithms. *Auditing: A Journal of Practice & Theory* 30, 19–50.

Piotroski, J.D., Wong, T., Zhang, T., 2015. Political incentives to suppress negative information: Evidence from Chinese listed firms. *Journal of Accounting Research* 53, 405–459.

Poirier, D.J., 1980. Partial observability in bivariate probit models. *Journal of Econometrics* 12, 209–217.

Post, C., Byron, K., 2015. Women on boards and firm financial performance: A meta-analysis. *Academy of Management Journal* 58, 1546–1571.

Purda, L., Skillicorn, D., 2015. Accounting variables, deception, and a bag of words: Assessing the tools of fraud detection. *Contemporary Accounting Research* 32, 1193–1223.

Ren, L., Zhong, X., Wan, L., 2022. Missing analyst forecasts and corporate fraud: Evidence from China. *Journal of Business Ethics* 181, 171–194.

Rosenstein, S., Wyatt, J.G., 1990. Outside directors, board independence, and shareholder wealth. *Journal of Financial Economics* 26, 175–191.

Sheedy, E., Zhang, L., Tam, K.C.H., 2019. Incentives and culture in risk compliance. *Journal of Banking & Finance* 107, 105611.

Shi, W., Aguilera, R., Wang, K., 2020. State ownership and securities fraud: A political governance perspective. *Corporate Governance: An International Review* 28, 157–176.

Shi, W., Connelly, B.L., Hoskisson, R.E., 2017. External corporate governance and financial fraud: Cognitive evaluation theory insights on agency theory prescriptions. *Strategic Management Journal* 38, 1268–1286.

Shleifer, A., 2004. Does competition destroy ethical behavior? *American Economic Review* 94, 414–418.

Shleifer, A., Vishny, R.W., 1986. Large shareholders and corporate control. *Journal of Political Economy* 94, 461–488.

Song, X.P., Hu, Z.H., Du, J.G., Sheng, Z.H., 2014. Application of machine learning methods to risk assessment of financial statement fraud: Evidence from China. *Journal of Forecasting* 33, 611–626.

Su, F., Feng, X., Tang, S., 2021. Do site visits mitigate corporate fraudulence? Evidence from China. *International Review of Financial Analysis* 78, 101940.

Sun, F., Dutta, S., Zhu, P., Ren, W., 2021a. Female insiders' ethics and trading profitability. *International Review of Financial Analysis* 74, 101710.

Sun, Y., Sun, X., Wu, W., 2021b. Who detects corporate fraud under the thriving of the new media? Evidence from Chinese-listed firms. *Accounting & Finance* 61, 1313–1343.

Ullah, S., Ahmad, S., Akbar, S., Kodwani, D., Frecknall-Hughes, J., 2021. Governance disclosure quality and market valuation of firms in UK and Germany. *International Journal of Finance & Economics* 26, 5031–5055.

Vafeas, N., 1999. Board meeting frequency and firm performance. *Journal of Financial Economics* 53, 113–142.

Wang, T., Hsu, C., 2013. Board composition and operational risk events of financial institutions. *Journal of Banking & Finance* 37, 2042–2051.

Wang, T.Y., 2013. Corporate securities fraud: Insights from a new empirical framework. *The Journal of Law, Economics, & Organization* 29, 535–568.

Wang, T.Y., Winton, A., 2021. Industry informational interactions and corporate fraud. *Journal of Corporate Finance* 69, 102024.

Wang, T.Y., Winton, A., Yu, X., 2010. Corporate fraud and business conditions: Evidence from IPOs. *The Journal of Finance* 65, 2255–2292.

Wei, Y., Chen, J., Wirth, C., 2017. Detecting fraud in Chinese listed company balance sheets. *Pacific Accounting Review* 29, 356–379.

Whiting, D.G., Hansen, J.V., McDonald, J.B., Albrecht, C., Albrecht, W.S., 2012. Machine learning methods for detecting patterns of management fraud. *Computational Intelligence* 28, 505–527.

Wu, W., Johan, S.A., Rui, O.M., 2016. Institutional investors, political connections, and the incidence of regulatory enforcement against corporate fraud. *Journal of Business Ethics* 134, 709–726.

Jiang, F. and Kim, K.A., 2015. Corporate governance in China: A modern perspective. *Journal of Corporate Finance* 32, 190–216.

Xiong, J., Ouyang, C., Tong, J.Y., Zhang, F.F., 2021. Fraud commitment in a smaller world: Evidence from a natural experiment. *Journal of Corporate Finance* 70, 102090.

Xu, X., Xiong, F., An, Z., 2022. Using machine learning to predict corporate fraud: Evidence based on the gone framework. *Journal of Business Ethics*, 1–22.

Yang, D., Jiao, H., Buckland, R., 2017. The determinants of financial fraud in Chinese firms: Does corporate governance as an institutional innovation matter? *Technological Forecasting and Social Change* 125, 309–320.

Yao, S., Zhao, W., Sensoy, A., Cheng, F., Goodell, J.W., 2021. The dark side of marital leadership: Evidence from China. *International Review of Financial Analysis* 77, 101844.

Yiu, D.W., Wan, W.P., Xu, Y., 2019. Alternative governance and corporate financial fraud in transition economies: Evidence from China. *Journal of Management* 45, 2685–2720.

Young, M.N., Peng, M.W., Ahlstrom, D., Bruton, G.D., Jiang, Y., 2008. Corporate governance in emerging economies: A review of the principal–principal perspective. *Journal of Management Studies* 45, 196–220.

Young, S.M., Peng, E.Y., 2013. An analysis of accounting frauds and the timing of analyst coverage decisions and recommendation revisions: Evidence from the US. *Journal of Business Finance & Accounting* 40, 399–437.

Yu, F., Yu, X., 2011. Corporate lobbying and fraud detection. J*ournal of Financial and Quantitative Analysis* 46, 1865–1891.

Zaman, R., Atawnah, N., Baghdadi, G.A., Liu, J., 2021. Fiduciary duty or loyalty? Evidence from co-opted boards and corporate misconduct. *Journal of Corporate Finance* 70, 102066.

Zhou, Z.g., Hussein, M., Deng, Q., 2021. Chinext IPOs' initial returns before and after the 2013 stock market reform: What can we learn? *Emerging Markets Review* 48, 100817.