International Review of Financial Analysis 89 (2023) 102827

Contents lists available at ScienceDirect

International Review of Financial Analysis

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

Detecting financial statement fraud using dynamic ensemble machine learning

Muhammad Atif Khan Achakzai *, Juan Peng

Antai College of Economics and Management, Shanghai Jiao Tong University, 1954 Huashan Road, Shanghai 200030, China

ARTICLE INFO

JEL classification:
C52
C53
C61
D83
M41

Keywords:
Fraud
Detection
Dynamic ensemble selection
Machine learning

ABSTRACT

Our study uses Machine learning to develop an advanced fraud detection model that can detect fraudulent firms. We build our model using raw financial and non-financial variables following prior literature. In addition, we introduce the Dynamic Ensemble Selection algorithm to the fraud detection literature, which combines individual classifiers dynamically to make a final prediction. Using several performance evaluation metrics, we find that our model can outperform several machine learning models used in recent studies.

1. Introduction

Financial statement fraud is a pervasive problem that harms businesses and their stakeholders globally. Financial statement fraud which is the act of publishing false or misleading information in financial statements has significant adverse effects on investor confidence and market order. The impact of fraud has reached a staggering figure, with estimates putting the cost of financial fraud at US$5.38 trillion (Gee & Button, 2021). Fraud generally also deteriorates trust in company disclosures, increasing uncertainty about the financial statements and leading to a higher capital cost (Graham, Li, & Qiu, 2008).

In the last decade, technological improvements coupled with access to a wide variety of information have improved corporate fraud detection. Earlier research on corporate fraud detection focused primarily on using logistic regression. The M-Score and F-Score models developed by Beneish (1999) and Dechow, Ge, Larson, and Sloan (2011) relied on Logistic Regression as the main algorithm used in their fraud detection studies. However, with the advancement in ML (Machine Learning) algorithms and the availability of high computing power in recent years, more studies have emerged that deploy ML models. The latest studies by Bao, Ke, Bin Li, Julia, and Zhang (2020) and Bertomeu, Cheynel, Floyd, and Pan (2021) showed that ML algorithms could be effectively deployed to detect financial fraud and outperform the traditionally used fraud detection models used in previous studies. However, we note that most of the previous ML literature on corporate fraud detection primarily relied on using individual ML classifiers to construct their fraud detection models and did not take advantage of ensemble learning.[1] We do note that Perols (2011) used a Stacking Classifier[2] to detect financial fraud and compared its performance with individual classifiers. However, the study did not consider the base classifiers' diversity when combining the models leading to sub-optimal results for the Stacked Classifier. Wolpert (1996) proposed the no free-lunch theorem and argued that no single ML classifier could work best across all scenarios. As a result, combining several accurate and diverse classifiers can lead to constructing an improved fraud detection model. In our study, we implement the DES (Dynamic Ensemble Selection) technique to construct our ensemble model, which allows us to enhance the fraud detection ability. Previous studies (García, Zhang, Altalhi, Alshomrani, & Herrera, 2018; Sergio, de Lima, & Ludermir, 2016) have shown that DES models can improve a model's accuracy, robustness, and

* Corresponding author.
E-mail addresses: atif007@sjtu.edu.cn (M.A.K. Achakzai), jpeng@sjtu.edu.cn (J. Peng).

[1] Ensemble learning involves combining different ML models to improve the overall performance and generalization of the model.

[2] Stacking is a technique that employs ensembles by merging the predictions of various base classifiers on the same dataset. The outputs of these base classifiers are then integrated using a meta-classifier. The logistic regression is commonly used as the the meta-classifier and we also use it in our Stacked Classifier model.

https://doi.org/10.1016/j.irfa.2023.102827
Received 19 March 2023; Received in revised form 4 May 2023; Accepted 31 July 2023
Available online 2 August 2023

generalization capabilities.

Moreover, previous literature on corporate fraud detection has primarily relied on financial ratios as the independent variables. Nonetheless, in recent years, an increased number of independent variables have been employed to enhance the performance of these models. Bao et al. (2020) used raw-financial variables extracted from the firm's financial statements to build their model and compared their performance against models that used financial ratios. Their findings suggested that raw-financial variables can lead to an improved fraud detection model compared to financial ratios. Wei, Chen, and Wirth (2017) also used raw financial statement variables to detect financial statement fraud and found that several balance sheet values were linked to fraud. Similarly, non-financial variables have also received less attention in the fraud detection literature that employs ML models. A study conducted by Bertomeu et al. (2021), which was used to detect misstatements, revealed that when financial and non-financial variables were combined to construct the model, the non-financial variables' importance was 63.8%. In contrast, the financial variable's importance accounted for 36.2% of the model.

Similarly, previous studies have shown a strong correlation between financial fraud and non-financial variables. Wu, Johan, and Rui (2016) found that firms with a higher proportion of institutional ownership can lower the incidence of fraud. Chen, Firth, Gao, and Rui (2006) found that firms with high-quality auditors reduce their propensity to commit fraud. As a result, the findings from previous literature allow us to construct a fraud detection model comprising raw-financial and non-financial variables that can be used to construct an improved fraud detection model.

Our study focuses on China for constructing our fraud detection model. Since the turn of the century, China's economy has grown significantly, which has coincided with the increased prevalence of corporate fraud in the country. Developing markets like China suffer from a weak regulatory regime and lack of transparency related to corporate governance and financial reporting, leading to opportunities for earnings management (Chen et al., 2006). In recent years, the CSRC (China Securities Regulatory Commission) and regulatory authorities (Shenzhen and Shanghai Stock Exchange) have issued multiple bulletins related to corporate fraud. In addition, some firms have been issued multiple bulletins published against them in a year, which we term as serial offenders. As a result, the detection of corporate fraud in China remains a significant issue.

Moreover, research on financial statement fraud and administrative penalties in Chinese listed companies reveals significant economic consequences. Prior research has reported considerable adverse effects on equity markets for fraudulent firms in China, including a decline in stock prices, larger bid-ask spreads, and a rise in the cost of capital (Firth, Rui, & Wu, 2011). Chen, Zhu, and Wang (2011) also find that the punishment for corporate fraud can impact a firm's financing contracts with its bank due to increased credit and information risk. In Chinese firms, post-fraud borrowing behavior reveals reduced bank loans and higher interest rates than in non-fraudulent firms. Corporate fraud also indirectly disrupts the relationship between performance and bank loans. These findings suggest that fraud negatively affects a firm's ability to secure debt financing, shedding light on the economic repercussions of fraud. In their study, Liebman and Milhaupt (2008) explored the repercussions of public denunciation from stock exchanges on fraudulent companies. Their findings revealed that such condemnation significantly damages the stock prices of these firms and hinders their future business endeavors, including obtaining bank loans. Niu, Li, Fan, and Zhang (2019) investigated how corporate fraud impacts households, discovering that those more exposed to such fraud are less inclined to invest in stocks and private insurance but more prone to invest in residential real estate. The variation in effects on investment decisions based on the type of fraud suggests that corporate fraud can negatively influence the entire financial system. Karpoff, Scott Lee, and Martin (2008) emphasized that accounting fraud, when uncovered, adversely affects a company's valuation. Companies accused of manipulating earnings by regulators face a negative response from the stock market, with average losses tripling the inflated market value gained through fraudulent activities.

In addition, Fu, Deng, and Tang (2023) examine Chinese public companies initiating cross-border M&As and investigate the impact of fraud revelations on such transactions. They discovered that acquirers with a fraud history have a lower likelihood of completing cross-border M&As, and if they do, the transactions take longer than those by other acquirers, which suggests that corporate fraud carries implicit costs, such as reputational loss, which can hinder the progress of cross-border M&As. The detrimental effects of corporate fraud in Chinese listed companies extend beyond the firms themselves, impacting equity markets, household investment decisions, debt financing, and cross-border M&As. These findings highlight the negative externalities and the high costs of financial fraud, leading to increased attention from regulators and practitioners and making fraud detection a popular topic in academic research.

In addition, this study focuses on detecting corporate fraud in China for two main reasons. Firstly, although various studies have been used to detect financial fraud using ML techniques, these have been mostly restricted to developed markets. Therefore, it is unclear whether emerging markets with weaker regulatory regimes can also use these techniques to detect financial fraud. Moreover, as China is the world's second-largest economy, it is also an important developing market. The study aims to detect financial fraud in Chinese firms that can be used in other emerging economies with weaker institutional regimes and high economic growth. Secondly, in China, a comprehensive list of data allows the study to use financial and non-financial information to construct a fraud detection model and explore its effectiveness. This study takes advantage of the newly introduced ML algorithms to improve the detection ability of corporate fraud. By using advanced techniques and wide-ranging information, the study aims to enhance corporate fraud detection and provide an early warning system for regulators and investors.

We are able to make several contributions through our study to the current literature. Firstly, we add to the growing literature on using machine learning algorithms to detect financial fraud (Bao et al., 2020; Bertomeu et al., 2021; Cecchini, Aytug, Koehler, & Pathak, 2010; Dong, Liao, & Zhang, 2018; Perols, 2011). Whereas most of the previous studies using ML and detecting corporate fraud have been related to developed markets, only a few have focused on developing markets. Our study adds to this scant literature and introduces the DES algorithm to the fraud detection literature. Our results show that the DES algorithm outperforms models used in previous studies when detecting financial fraud and is also effective at detecting serial offenders, which provides a valuable technique to regulators and practitioners. Moreover, since previous studies in developed markets have primarily relied on using single ML classifiers, with the exception of Perols (2011) that used a static ensemble of Stacked classifiers, we believe the DES model can also be used to detect financial fraud in these markets to get an improved fraud detection model. In addition, the model can also assist investors and regulators in emerging economies in better-grasping ML applications to detect corporate fraud.

Secondly, we also contribute to the financial fraud literature in China. Earlier studies had primarily relied on using logistic regression to detect financial fraud by identifying several factors which led to it. Many of these factors were related to non-financial variables, which included corporate governance (Jia, Ding, Li, & Zhenyu, 2009), forecasts made by analysts (Ren, Zhong, & Wan, 2022), executive compensation (Conyon & He, 2016) and executive characteristics (Luo, Peng, & Zhang, 2020). In addition, previous literature using financial variables to detect financial fraud primarily relied on financial ratios. Wei et al. (2017) used raw financial variables to detect corporate fraud. They found that these variables can help to detect corporate fraud. However, no study has used a model comprising raw financial and non-financial variables to detect

financial fraud. Therefore, our study is the first to use such a model and enhance the predictive performance of fraud detection.

2. A review of machine learning and corporate fraud

Most of the prior literature on fraud detection primarily relied on econometric models. Financial statement fraud was treated as a binary variable, and the focus of the studies was to establish causality. Moreover, many previous studies used logistic regressions and financial ratios to explain the determinants of accounting fraud. Before ML became more common in prediction-related studies, the M-Score and F-Score models developed by Beneish (1999) and Dechow et al. (2011) were considered the most effective models for detecting fraudulent firms. However, with the advancement in ML algorithms and improved computing power, there has been a shift towards using ML algorithms to detect financial fraud in recent years.

The application of ML algorithms to identify fraudulent firms has experienced a surge in the past few years. An SVM (Support Vector Machine) algorithm using financial ratios was used by Cecchini et al. (2010) to detect fraudulent firms, and their model identified 80% of fraud cases used in their study. Perols (2011) also undertook a fraud-related study using six machine learning algorithms to detect financial fraud; their findings suggested that the Logistic Regression and SVM model were the best at identifying fraud cases. In addition, Purda and Skillicorn (2015) inspected the financial reports' MD&A (Management Discussion and Analysis) section and examined their textual features using Random Forests and SVM. They found that a firm's MD&A report contains information that can be used to detect financial fraud. A more recent study by Bao et al. (2020) used raw financial variables in contrast to financial ratios used in previous studies and a RusBoost algorithm. Their study found that using such a model can outperform the best-identified ML algorithms used in previous studies, i.e., SVM and Logistic Regression algorithms. In addition, Brown, Crowley, and Brooke Elliot (2020) used textual analysis to study the thematic content in the financial reports using a Latent Dirichlet Allocation algorithm for their fraud detection model. They found that when the thematic content in the financial reports improves, it leads to improved detection of financial misreporting. Bertomeu et al.'s latest study in 2021 employed a comprehensive array of financial and non-financial factors to identify accounting misstatements by utilizing GBRT (Gradient Boosted Regression Tree) algorithms. Their findings demonstrated that GBRT surpassed the performance of Random Forests, RusBoost, and Logistic Regression in detecting these misstatements.

In addition, it is also important to acknowledge several challenges related to using ML and detecting corporate fraud. The first challenge lies in identifying variables to be used in the fraud detection model, which affects the prediction accuracy of the models. Recent studies have used various input variables ranging from financial ratios, raw financial variables, textual analysis variables, and corporate finance-related and audit-related variables. Bao et al. (2020) was the first ML fraud detection study that deviated from the traditional use of financial ratios and incorporated financial variables in their raw form from financial statements. They formed a list of financial variables by referring to Cecchini et al. (2010) and Dechow et al. (2011); their results suggest that using raw financial variables can lead to improved performance of the fraud detection model when compared to financial ratios. In addition, Bertomeu et al. (2021) employed a comprehensive list of financial variables following Dechow et al. (2011) and non-financial variables following Larcker, Richardson, and Tuna (2007) and DeFond, Raghunandan, and Subramanyam (2002). Their study found that adding non-financial variables to financial variables can lead to improved performance of the ML models, where the non-financial variables account for higher importance in the model. Following previous literature allows us to form a comprehensive list of variables that can be used to construct a good-performing fraud detection model.

The second challenge in financial fraud detection is its application in real-world situations. Since fraud is not a typical situation, there is a situation of imbalance between the two types of firms (fraud versus non-fraud). This has led to some researchers using matched samples between fraudulent and non-fraud firms, which raises questions about the real-world application of these models. Cecchini et al. (2010) used a matched sample of firms in their fraud detection model (6426 non-fraudulent firms versus 205 fraudulent firms). Similarly, a one-to-one ratio of fraudulent and non-fraud firms (64 firms) was used by Dong et al. (2018). However, recent studies questioned the use of matching samples, arguing that using such a sample in a real-world situation will lead to biased results and make the models' use impractical. As a result, the latest studies by Bao et al. (2020) and Bertomeu et al. (2021) avoid using matched samples. We follow the latest literature by constructing a practical model and avoid using matched samples since it would lead to overestimating the algorithm's performance.

The third challenge lies in the identification of the ML algorithm for the detection of corporate fraud. Previous literature has identified several ML models that detect financial fraud. Recent research has shown that the top-performing models for detecting fraudulent firms include GBRT (Bertomeu et al., 2021), Random Forests (Purda & Skillicorn, 2015), SVM (Perols, 2011), RusBoost (Bao et al., 2020) and Logistic Regression (Dechow et al., 2011). Furthermore, we stress the significance of adjusting the ML algorithm's hyper-parameters. Bertomeu (2020) noted that default models are ill-suited for handling imbalanced datasets. Walker's (2020) fraud detection model used an XGBoost algorithm without fine-tuning the hyper-parameters and yielded unsatisfactory outcomes. Consequently, it is crucial to fine-tune the ML algorithm's hyper-parameters to obtain optimal results.

The last significant challenge lies in the generalizability of ML detection algorithms for financial fraud. Much of the existing literature concentrates on the US market, a developed economy characterized by more sophisticated capital markets and strict regulatory frameworks. In contrast, the application of ML in emerging markets for detecting financial fraud remains relatively unexplored. Consequently, it is crucial to ascertain whether ML models used in developed markets are also applicable to developing countries with weaker regulatory regimes. Similarly, choosing performance evaluation metrics appropriate for imbalanced datasets is vital. Utilizing commonly used metrics suited for balanced datasets (such as accuracy) may yield misleading results. Therefore, it is essential to select evaluation metrics that consider the problem's nature.

3. An overview of dynamic ensemble selection (DES)

In recent years, applying DES models to real-world issues has seen notable success in different domains, including credit scoring, face recognition, music genre classification, time series forecasting, and biomedicine (Cruz, Sabourin, & Cavalcanti, 2018). Recent studies by Lessmann, Baesens, Seow, and Thomas (2015), Xiao, Xiao, and Wang (2016), and Hou, Wang, Zhang, Wang, and Li (2020) applied the DES models to credit scoring and risk assessment, which is an imbalanced problem similar to fraud detection and found that these models outperformed individual classifiers.

In contrast to conventional ML models like Logistic Regression, which generates a single estimator, DES combines predictions of a group of base classifiers dynamically at runtime to make a final prediction. This approach differs from static ensemble learning like Random Forests and Stacked Classifiers, where the ensemble is fixed at training time, and all the models in the ensemble are used to make predictions. Conversely, the base models used in the DES algorithm may change depending on the input data or other factors, which can improve the generalizability and robustness of the model.

In addition, DES classifiers can outperform individual and static ensemble classifiers in many ways, including capturing local competence, exploiting base classifiers' diversity and complementarity, reducing overfitting, dynamically adapting to data changes, and

effectively handling noisy data. Firstly, the DES classifier dynamically selects the most appropriate classifiers for each instance, considering their competence in specific input space regions, which can lead to improved performance compared to static ensemble classifiers, which do not account for local competence. Static ensemble classifiers, like Stacked Classifiers or Bagging, use the same set of classifiers for all instances without considering their local competence. DES classifiers, on the other hand, adapt their ensemble for each instance, leading to improved performance (Cruz et al., 2018). Secondly, DES classifiers exploit the diversity and complementarity of the base classifiers in the ensemble. Using multiple base classifiers, they capture different aspects of the data and make better predictions. This can improve performance over single base classifiers with limited capacity to capture complex patterns (Kuncheva & Whitaker, 2003).

Thirdly, DES classifiers have a better generalization capability than single-base classifiers because they reduce the risk of overfitting. Combining the predictions of multiple base classifiers makes DES ensembles less likely to overfit the training data than individual classifiers, resulting in better performance on unseen data (Ko, Sabourin, & Jr. Britto Alceu Souza., 2008). Fourthly, DES classifiers can dynamically adapt to changes in the data distribution or concept drift, a common issue in many real-world classification problems. This adaptability allows DES classifiers to maintain high performance even in the presence of changing data, where static ensemble classifiers might struggle to adapt (Gomes et al., 2017). Finally, DES classifiers can be more robust to noise in the data since they can identify and rely on the classifiers that perform well on noisy instances. This can lead to better performance than static ensemble classifiers, which use the same classifiers for all instances without considering their performance on noisy data (Dos Santos, Eulanda, Sabourin, & Maupin, 2008).

Our study uses the DESRRC (Dynamic Ensemble Selection-Randomized Reference Classifier) that Woloszynski and Kurzynski (2011) developed. A review by Cruz et al. (2018) compared several DES models and found that the DESRRC is one of the best-performing models when used across multiple real-world datasets. We first introduce the DES algorithm and then describe how the DESRRC works. The DES algorithm starts by selecting a group of base classifiers trained on a subset of the available data. During the classification phase, the DES algorithm selects a subset of the base classifiers for each instance to be classified based on their past performance on similar instances. The algorithm then combines the nominated base classifiers' outputs to obtain the final classification output. The key idea behind DES is to use the most accurate classifiers for each instance while avoiding using classifiers that are likely to make errors on that particular instance. This allows DES to adapt to changing data distributions and achieve better classification accuracy than a single classifier. A graphical representation of a DES model is shown in Fig. 1.

The DESRRC is an ML-based algorithm which combines the concept of DES and RRC (Randomized Reference Classifier) to determine if the base classifier significantly outperforms a random classifier. During the classification process, this algorithm employs a randomization method to choose a subset of base classifiers for each instance requiring classification. This approach enables the classifier to examine various combinations of classifiers in order to identify the most accurate one. After that, the DESRRC selects the most competent base classifier from the subset using a competence level estimate based on the classification performance of each classifier and combines them. Mathematically, DESRRC selects a subset of base classifiers, C_s, for a given input x based on the output of the reference classifiers, $R_i(x)$, where $i = 1,2,3, \ldots,m$, and m is the number of reference classifiers. The reference classifiers are trained on subsets of the data using random sampling. The subset of base classifiers C_s is chosen as the set of classifiers whose outputs are closest to the average output of the reference classifiers, as shown in the following equation:

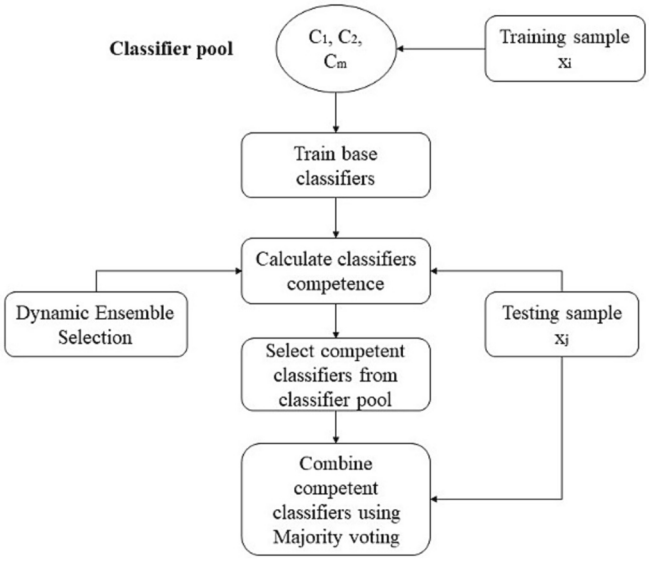

Fig. 1. DES model flowchart.

$$C_s = C_i \mid R_i(x) \geq \left(\frac{1}{k}\right) * Sum\left(R_j(x)\right)$$

Where k is a user-defined parameter that controls the size of the subset. Once the subset of classifiers is selected, DESRRC applies a dynamic ensemble selection mechanism to choose the best one(s) for the given input. The final prediction is made by combining the outputs of the selected classifiers using majority voting. Mathematically, the DESRRC algorithm can be expressed as follows:

$$f(x) = h(C_s(x))$$

Where f(x) is the final prediction for input x, h is a function that combines the outputs of the selected classifiers, and $C_s(x)$ is the subset of classifiers chosen for input x. (See Woloszynski and Kurzynski (2011) for more details on DESRRC).

Previous research has also shown that ensemble models based on the diversity of classifiers, such as the DES model or static models like Stacked Classifiers, can outperform single-base estimators in various classification problems. However, it is important to note that diversity is an essential trait to obtain better performance from an ensemble model compared to an individual classifier (the classifiers should not make the same errors). Therefore, constructing a competent ensemble model requires training accurate and diverse base classifiers on a training set and creating a pool of candidate classifiers (Sesmero, Iglesias, Magán, Ledezma, & Sanchis, 2021). We draw on previous fraud detection literature to select the base classifiers for comparison and construct the DES and Stacked Classifier model. Our study employs eight different base classifiers, including (1) Logistic Regression, (2) Balanced Bagging, (3) XGBoost, (4) Ridge Classifier, (5) RusBoost, (6) Random Forests, (7) GBRT, and (8) SVM.

To select diverse base classifiers for constructing the DESRRC and Stacked Classifier model, we refer to Sesmero et al. (2021), who used several diversity measures to construct an ensemble of classifiers for different datasets. Their study found that the DF (Double-Fault) metric performed the best at selecting diverse base classifiers that can be used to construct an effective ensemble model. The DF measure is a way to measure how good a classifier is in comparison to other classifiers in an ensemble. The idea behind this measure is that a good classifier should be able to correctly predict the class of an instance that other classifiers in the ensemble cannot. It is called a “double fault” - when two or more classifiers in the ensemble make an incorrect prediction on the same instance.

The DF measure counts the instances that a given classifier misclassifies and is also misclassified by at least one other classifier in the ensemble, which is then divided by the sum of instances in the dataset. A lower value of the DF measure indicates that the classifier is better at predicting the correct class, i.e., it makes fewer "double faults" and is more competent than the other classifiers in the ensemble. The DF measure is the proportion of misclassified examples by both classifiers. A lower value of the DF measure means that the base classifiers are less likely to make the same error, thus increasing the ensemble's diversity.

4. Data and research design

4.1. Sample selection

Our fraud data is obtained from CSMAR (China Stock Market and Accounting Research), which includes fraud-sanctioned cases by CSRC and other regulatory authorities, including the Shenzhen and Shanghai stock exchanges. The database contains information on every fraud case's revelation and occurrence dates. A dummy variable is created that equals one if the regulators sanctioned the firm for fraud in a specific year, and it equals zero when there has been no sanction against a firm. Moreover, we also observe that some companies have more than one fraud sanction against them in a calendar year. We term firms with more than one fraud case in a single year as serial offenders and use the dummy variable one where there has been more than one fraud case and zero when no fraud case or only one fraud case has occurred.

All of the study's financial and non-financial variables data are collected from CSMAR, excluding the MD&A tone, which is collected from CNRDS (Chinese Research Data Services Platform), and the Internal Control Index collected from the DIB database. For the raw financial variables used in our study, we follow previous literature by Wei et al. (2017) and Bao et al. (2020) to obtain 26 variables. Whereas for the non-financial variables, we refer to previous studies by Mutlu, Van Essen, Peng, Saleh, and Duran (2018), Wu et al. (2016), Lin, Chiu, Huang, and Yen (2015), and Bertomeu et al. (2021) and obtain 24 non-financial variables. In addition, firms with B-share listings and missing values are dropped from our sample. Our final sample comprises 50 independent variables.[3]

After merging all the variables used in our study, we obtained 22,329 firm years from 2007 to 2020. We observe that 2453 firms committed at least one fraud in a single year in our study. Fraud-related data can be observed in Table 1, which shows the fraud occurrences throughout our sample period. We observed a sharp increase in the number of frauds starting in 2015. This increase can be explained due to two factors. Firstly, the Chinese IPO market was halted in October 2012 by the CSRC due to concerns over weak market conditions, financial fraud, and insufficient transparency. The IPO market reopened in January 2014 after regulatory reforms were implemented to address these concerns and improve the market's overall health. The reopening of the IPO market was accompanied by new rules and regulations to improve transparency and protect investors (Zhou, Hussein, & Deng, 2021). Secondly, in October 2013, the regional offices (38 offices in each region of China) of the CSRC were delegated the authority to impose sanctions on fraudulent firms, which led to an increase in the number of frauds (Xu, 2022). Due to these factors, we can observe a marked increase in fraud starting in 2015.

The proportion of fraud for the study period is approximately 11%, which aligns with previous studies on corporate fraud in China. A study by Luo et al. (2020) that studied financial-reporting fraud and CFO gender showed that in their sample from 2004 to 2014, the fraud average was 11%. Similarly, a study by Ren et al. (2022) that examined corporate fraud and analyst forecasts from 2008 to 2019 observed 14.59% fraudulent occurrences.

Table 1
Fraud distributions.

Year	No. of listed firms	No. of fraud firms	Percentage of fraud firms	No. of serial offenders
2007	747	32	4.28%	3
2008	867	40	4.61%	4
2009	960	36	3.75%	5
2010	1059	34	3.21%	5
2011	1424	55	3.87%	10
2012	1722	80	4.65%	13
2013	1795	109	6.07%	17
2014	1707	128	7.50%	21
2015	1814	228	12.57%	41
2016	1896	319	16.82%	63
2017	1998	350	17.52%	109
2018	2213	418	18.89%	126
2019	2218	374	16.86%	112
2020	1909	250	13.10%	58
Total	**22,329**	**2453**		**587**

The percentage of fraud firms represents the proportion of fraud firms out of the total listed firms annually. Whereas serial offenders are firms that have committed more than one fraud case annually.

It is worth noting that the proportion of fraud firms in China is significantly higher than in the US, where the fraud firms are 1% (Bao et al., 2020) and restatement firms are 6.6% (Bertomeu et al., 2021) due to several reasons. Firstly, the CSMAR database includes all types of fraud. In contrast, the US studies focus on serious material frauds disclosed through AAERs (Accounting and Auditing Enforcement Releases). Therefore, to ensure that our study can be replicated and comparable to the US setting, we also analyze serial offenders: firms with more than one fraud sanction in a calendar year. As shown in Table 1, our fraud sample for serial offenders reduces to 587 firms, which is approximately 2.9% of the firm years.

4.2. Serial frauds & train-test split

We also recognize the situation where fraud may have spanned over multiple periods, as Bao et al. (2020) noted in their study in the US setting. In our sample, we also observe that fraud cases identified by the regulators have spanned over several consecutive reporting periods. As a result, if the fraud cases span both the training and testing years, then using such fraud cases in our model would most likely overstate the performance of our models. Therefore, to ensure that the improvement in the performance of our models is not inflated due to the training and test sample period comprising of the same fraud firms. All firms in the training sample are recorded as non-fraud (zero), where the same fraud has spanned both periods.

In addition, to ensure that our model is practical and can be used in a real-world situation. We perform our train-test split chronologically following Bertomeu et al. (2021). Our training period spans 2007–2015, the validation period from 2016 to 2017, and the test period from 2018 to 2020. We utilize the training period to train our models, while the validation period determines the optimal hyperparameters for each model. After obtaining the optimal hyperparameters for each model, we combine the training and validation dataset (2007–2017) to train the final model and test them on the out-of-sample test period from 2018 to 2020. Fig. 2 provides us with a graphical representation of the data split.[4]

[3] All of the financial and non-financial variables used in the study are listed in Appendix 1.

[4] The sample of firms is reduced to 20,462 firm-years for the serial offenders, since we remove the firms with a single case of fraud when building the model. The training data consists of 11,470 firm-years, validation data 3398 firm-years and test data 5594 firm-years.

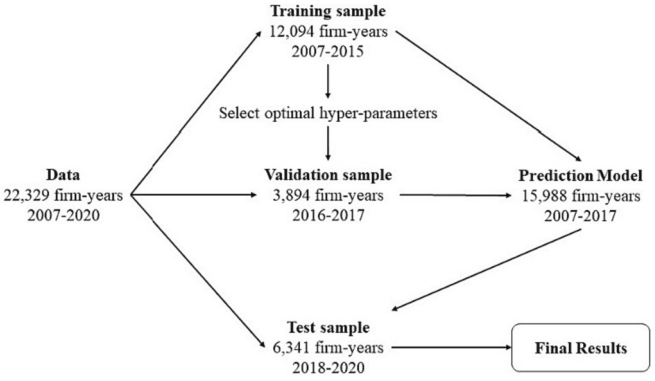

Fig. 2. Data split for all firms.
The train data is set to 2007–2015, and validation data to 2016–2017 to fine-tune the hyperparameters. Once we obtain the hyper-parameters for each model, the train and validation datasets are combined (2007–2017) and retrained to test on the out-of-sample test data 2018–2020.

4.3. Performance evaluation metrics

Since fraud detection can be viewed as a binary classification situation (fraudulent versus non-fraudulent), we use evaluation metrics suitable for classification problems. Moreover, since our dataset is imbalanced with more non-fraud firms compared to fraud firms, choosing suitable performance metrics is vital for the results to be meaningful. To ensure the robustness and unbiasedness of the results, we employ multiple metrics to evaluate the performance of our models.

Our first performance metric is the ROC (receiver operating characteristic) curve, demonstrating the models' ability to differentiate between fraud and non-fraud firms. The ROC curve is plotted using various classification thresholds and plotting a graph of the True Positive Rate $TPR = \left[\frac{TP}{TP+FN}\right]$ against the False Positive Rate $FPR = \left[\frac{FP}{TN+FP}\right]$.[5] A model's performance is better if the curve is closer to the top-left part of the graph. The ROC curve can also be reduced to a single value: the AUC (Area under the ROC curve). The probability that a fraud firm selected at random has a higher predicted fraud probability than a non-fraud firm selected at random is quantified by the AUC score (Fawcett, 2006). The AUC score ranges from 0 to 1, where a random guess has a value of 0.5. Moreover, for the models' results to be meaningful, they should have a score higher than 0.5. An advantage of the AUC is that it is insensitive to changes in the balance of the dataset and class distribution. Therefore, the AUC score can be used for both balanced and imbalanced datasets.

Our second metric is the PR (Precision-Recall) curve, which is used to evaluate the performance of a classifier in situations where the minority class (fraud) is rare (Japkowicz, 2013). The PR curve is used to plot the trade-off between $Precision = \frac{TP}{TP+FP}$ and $Recall = \frac{TP}{TP+FN}$ at different thresholds. A model with a PR curve to the top-right part of the graph is considered a better-performing model. Similarly, the PR curve can be reduced to a single value: AP (Average Precision). The AP is calculated using the area under the PR curve by averaging the precision values at various classification thresholds. The AP score ranges from 0 to 1, with a higher score indicating better performance. The AP is sensitive to class distribution and dataset balance changes, making it more appropriate when the minority class (fraud) is rare. The AP is calculated as follows $AP = \sum_n (R_n - R_{n-1})P_n$.[6] The AP is beneficial in tasks that rank positive instances (fraud firms) higher than negative cases (non-fraud firms). The main advantage of the AP measure is that it helps identify the ML classifiers' effectiveness in ranking fraudulent firms.

Our third metric, the MCC (Matthew's correlation coefficient), is another valid measure for evaluating the relationship between actual and predicted values, mainly when dealing with unbalanced datasets (Chicco & Jurman, 2020). This measure calculates the Pearson product-moment correlation coefficient using a contingency matrix. The MCC's value ranges from −1 to +1, with high scores indicating an accurate prediction of positive and negative data instances. It is the only binary classification rate that considers the correct prediction of both instances (fraud and non-fraud). The formula for the MCC metric is: $MCC = \frac{TP.TN - FP.FN}{\sqrt{(TP+FP).(TP+FN).(TN+FP).(TN+FN)}}$. The main advantage of MCC is that it is less sensitive to class imbalance since it considers both true and false positives and negatives, irrespective of the relative frequencies of each class in the dataset, and is unbiased.

Additionally, to account for the resource and time constraints placed on the regulators, it would be impractical to investigate every fraud case. To account for these constraints, we can consider fraud a ranking problem where firms with a higher probability of fraud should be investigated first. As a result, we also use the NDCG@k (Normalized Discounted Cumulative Gain), Recall@k, and Precision@k metrics proposed by Bao et al. (2020). The performance of the fraud detection model can be assessed by choosing the top-k firms with the greatest likelihood of engaging in fraudulent activities.

In order to evaluate how well the models can be used in ranking instances, the NDCG@k is a frequently used measure in recommendation and web search engine algorithms (Järvelin & Kekäläinen, 2002). The NDCG is a metric used to evaluate the quality of a ranking system. It takes into account the relevance of the items ranked and their position on the list. NDCG@k is a variant of this measure that considers only the top-k results. Mathematically NDCG@k can be written as $NDCG@k = \frac{DCG@k}{iDCG@k}$, where $DCG@k = \sum_{i=1}^{k} \frac{2^{rel_i}-1}{log_2(i+1)}$, rel_i is the relevance score of the item at position i in the ranked list, and k is the number of items to be considered. The DCG@k calculates the sum of the relevance scores of the top-k items in the ranked list, with a discount factor that increases logarithmically as the rank position increases. Similarly, the iDCG@k refers to the calculation of the ideal DCG at a particular ranking position, k. It represents the highest achievable DCG@k score, which can be attained if the items are perfectly ranked in decreasing order of their relevance scores. Finally, the NDCG@k normalizes DCG@k by dividing it with the iDCG@k, which gives a score between 0 and 1. A higher score indicates a better ranking. Given that the average incidence of fraud in our test data (2018–2020) for the overall fraudulent firms (serial offenders) model is 16.4% (5.3%), we set k to 16.4% for the overall fraud detection model and 5.3% for the model used to detect serial offenders.[7] We supplement the NDCG@k with the Recall@k and Precision@k to evaluate the model's ability to correctly identify fraudulent firms among the top k firms.

5. Results and discussion

Table 2 presents the DF diversity measure between the different base classifiers using the validation dataset. As explained in the earlier section, a lower value of the DF measure between two base classifiers means that they are less likely to make the same error, leading to increased diversity in the ensemble and possibly leading to better performance. The table shows that the Ridge Classifier and Random Forests have the lowest DF score of 0.138 compared to the other base classifiers. Therefore, we use them as the base classifiers in the Stacked Classifier and our DESRRC model.

We report the results of the various ML models when we use the all firms sample in Table 3. Our first performance metric, the AUC,

[5] Here FN, FP, TP and TN represent false negatives, false positives, true positives and true negatives, respectively.

[6] where *Pn* is precision and *Rn* is recall at '*n*' threshold.

[7] We also set k = 11% for all fraudulent firms and 2.9% for serial offenders as a robustness check, which is the average incidence of fraud during our sample period.

Table 2
Diversity measure (Double Fault).

Models	Logistic Regression	SVM	Random Forests	Ridge Classifier	RusBoost	XGBoost	GBRT	Balanced Bagging
Logistic Regression	1	0.154	0.186	0.183	0.152	0.163	0.188	0.191
SVM	0.154	1	0.169	0.180	0.209	0.173	0.151	0.158
Random Forests	0.186	0.169	1	0.138	0.175	0.166	0.187	0.207
Ridge Classifier	0.183	0.180	0.138	1	0.173	0.169	0.177	0.179
RusBoost	0.152	0.209	0.175	0.173	1	0.170	0.151	0.168
XGBoost	0.163	0.173	0.166	0.169	0.170	1	0.168	0.164
GBRT	0.188	0.151	0.187	0.177	0.151	0.168	1	0.189
Balanced Bagging	0.191	0.158	0.207	0.179	0.168	0.164	0.189	1

The table shows the double-fault measure between the base classifiers. A higher value indicates that both classifiers incorrectly predicted the outcome. In contrast, a lower value indicates that the base classifiers are less likely to make the same error, thus increasing the ensemble's diversity.

Table 3
Model performance of all firms (2018–2020).

Models	AUC	MCC	AP	NDCG@ 16.4%	Precision@ 16.4%	Recall@ 16.4%	NDCG@ 11%	Precision@ 11%	Recall@ 11%
Logistic Regression	0.694	0.270	0.318	0.392	0.410	0.354	0.420	0.433	0.433
SVM	0.713	0.288	0.317	0.331	0.310	0.310	0.371	0.351	0.351
Random Forests	0.725	0.276	0.344	0.405	0.421	0.393	0.439	0.430	0.430
Ridge Classifier	0.730	0.308	0.380	0.431	0.441	0.392	0.484	0.483	0.483
RusBoost	0.719	0.271	0.363	0.417	0.396	0.396	0.467	0.450	0.450
XGBoost	0.704	0.255	0.345	0.396	0.380	0.380	0.442	0.431	0.431
GBRT	0.684	0.179	0.330	0.387	0.498	0.115	0.430	0.498	0.172
Balanced Bagging	0.690	0.231	0.307	0.365	0.397	0.288	0.406	0.405	0.405
Stacked Classifier	0.734	0.316	0.388	0.441	0.471	0.352	0.498	0.483	0.483
DESRRC	0.763	0.354	0.443	0.485	0.475	0.438	0.556	0.536	0.536

The table presents the performance of each ML model applied to the sample of all firms in our study, spanning the test period from 2018 to 2020. The evaluation metrics featured in the table include the AUC (Area under the ROC) score, MCC (Matthew's Correlation Coefficient), AP (Average Precision) score, NDCG@k (Normalized Discounted Cumulative Gain at k), Recall@k and Precision@k. A higher value for any of these metrics signifies a more effective model.

measures the model's ability to differentiate between non-fraud and fraudulent firms. The DESRRC has the highest AUC at 0.763, whereas the second-best model is the Stacked Classifier with an AUC of 0.734. These results show that when the DESRRC model selects a fraud firm randomly, it has a higher predicted fraud probability than a non-fraud firm. In Fig. 3, we refer to the ROC curve and observe that the ROC curve of the DESRRC is the highest and above all the other curves, demonstrating that the model can distinguish between the fraud and non-fraud firms across all classification thresholds. We also observe that all the classifiers used in the study have a higher ROC curve than a random guess (represented by a white line in Fig. 3).

Our second evaluation metric, the MCC, measures the model's ability for the binary classifications (fraud and non-fraud) by considering the TP, FP, TN, and FN. Here again, the DESRRC outperforms the other models with an MCC of 0.354, and the second-best model is the Stacked Classifier at 0.316. The higher MCC score of the DESRRC suggests that the model has a better quality of binary classification (fraud and non-fraud cases), with a higher number of TP and TN and lower FP and FN.

Our third metric, the AP score, measures the quality of the different model's ability to correctly classify fraudulent firms using Precision and Recall at various thresholds. The best-performing models are the DESRRC and Stacked Classifier, with AP of 0.443 and 0.388. The higher AP score of 0.443 suggests that the DESRRC model performs better in accurately classifying the fraudulent firms, with higher Precision and Recall at various thresholds. Fig. 4 shows the PR curve, and we observe that the DESRRC model is towards the top right part of the graph and above all the other models' curves. This shows that the DESRRC is better at correctly predicting fraudulent firms across the different thresholds in the PR curve.

Next, we evaluate and compare the effectiveness of the different models in identifying firms with a high probability of committing fraud, specifically focusing on the top 16.4% of observations. The NDCG@k is the first metric for our study's top 16.4% of firms, and it measures the model's effectiveness in ranking the fraudulent firms. Here again, we find that the DESRRC and Stacked Classifier are the best-performing models, with a score of 0.485 and 0.441, respectively. The higher NDCG@k score of 0.485 suggests that the DESRRC model is more effective at ranking the fraudulent firms in the top k positions. Moreover, for the Precision@k metric, we find that the GBRT performs the best at 0.498, followed by DESRRC at 0.475. This implies that the model can accurately detect actual fraud firms when it ranks them within the top 16.4% of the total number of fraud firms identified. However, when we look at Recall@k, we find that the DESRRC is the best model for correctly identifying the highest number of actual fraudulent firms at 0.438, followed by RusBoost at 0.396.

We further report the performance evaluation metrics for the top 11% of firms as a robustness check. We obtain similar results for all three measures—NDCG@k, Recall@k, and Precision@k at 11%. Our findings remain consistent, with the DESRRC model surpassing the performance of other models. The results from Table 3 demonstrate that the DESRRC model outperforms the other tested models when selecting the top k firms. In addition, we also observe that the Stacked Classifier also performs across most of the performance evaluation metrics. These results suggests that the DESRRC model could be well-suited for regulators, as it can effectively identify and detect firms engaged in fraudulent activities.

5.1. Serial offenders

When using the serial offenders' sample in our study, firms with more than one sanction against them in a calendar year, we report the findings in Table 4. Due to the significantly smaller number of serial offenders compared to the total number of fraudulent firms, it becomes increasingly difficult for the models to identify fraudulent firms accurately. Nevertheless, our results align with previous findings. When considering the overall model performance evaluation metrics, such as AUC, MCC, and AP, the DESRRC model achieves the highest score and exhibits the best performance. Furthermore, by examining Figs. 5 and 6, we can view the ROC and PR curves. Comparing these two figures demonstrates that the DESRRC model consistently outperforms other models across all thresholds.

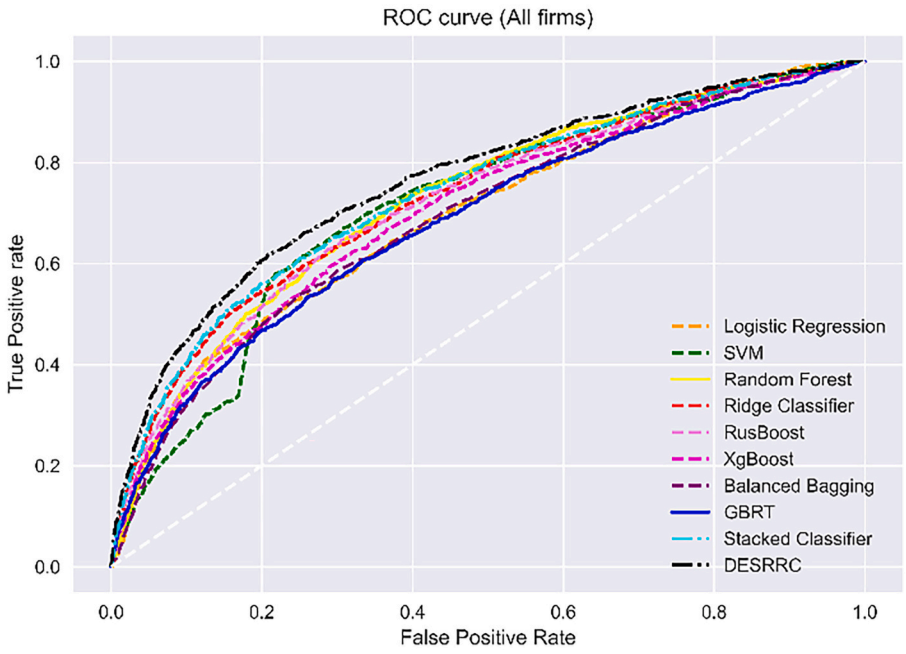

Fig. 3. ROC curve (All firms) for the test period 2018–2020.
The ROC (Receiver Operating Characteristic) curve illustrates a classifier's performance by plotting the TPR (True Positive Rate) against the FPR (False Positive Rate) at different classification thresholds. A model with better performance will have a curve closer to the top-left section of the graph. The white line depicted in the graph symbolizes the performance of a random guess model.

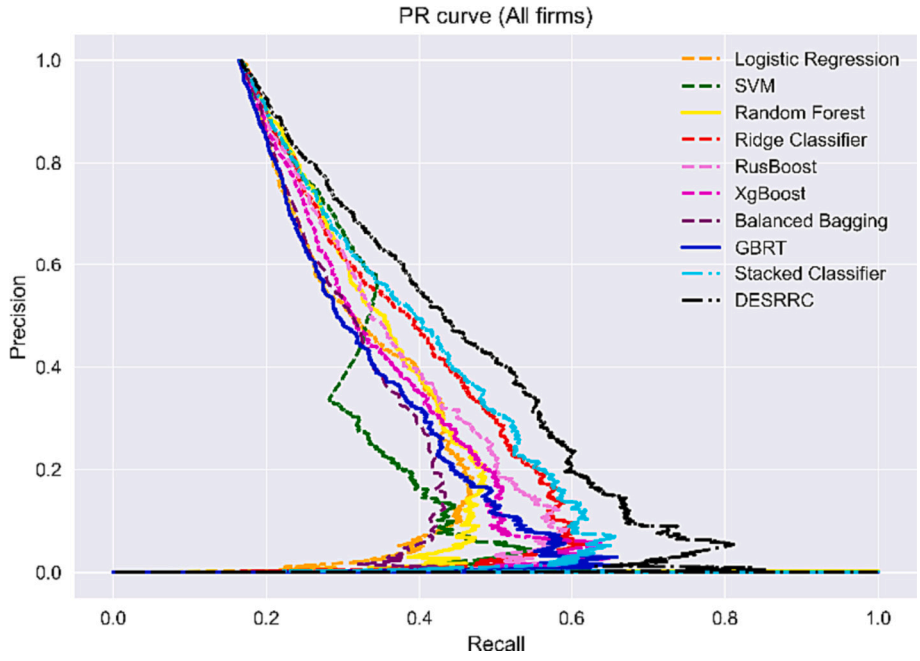

Fig. 4. PR curve (All firms) for the test period 2018–2020.
The PR (Precision-Recall) curve represents a classifier's performance by plotting the Precision (Positive Predicted Value) against the Recall (True Positive Rate) at various classification thresholds. The closer the curve is to the top-right part of the graph, the better the model's performance.

Analogous outcomes are achieved when employing the performance evaluation metrics for the top k firms. We present the results for k at 5.3% and 2.9% to ensure the robustness of our findings and observe that the DESRRC consistently outperforms all other models. We also note a reduction in the Stacked Classifier's performance when detecting serial offenders. This indicates that, unlike dynamic ensemble models, static ensemble models lack consistency, which could be due to their static combination of base classifiers and potential struggles with noisy data. Moreover, these results showcase the DESRRC model's consistency over individual classifiers and static ensemble models. Our findings imply that the DESRRC model is also more adept at accurately identifying serial offenders.

5.2. *Alternative periods*

To further verify the robustness of our models, we re-evaluate them

Table 4
Model performance of serial offenders (2018–2020).

Models	AUC	MCC	AP	NDCG@ 5.3%	Precision@ 5.3%	Recall@ 2.9%	NDCG@ 2.9%	Precision@ 2.9%	Recall@ 2.9%
Logistic Regression	0.724	0.269	0.176	0.282	0.304	0.304	0.277	0.317	0.317
SVM	0.796	0.318	0.312	0.401	0.365	0.341	0.457	0.391	0.391
Random Forests	0.831	0.309	0.297	0.402	0.364	0.324	0.464	0.398	0.398
Ridge Classifier	0.842	0.337	0.264	0.337	0.389	0.348	0.365	0.398	0.398
RusBoost	0.695	0.134	0.133	0.212	0.196	0.196	0.227	0.205	0.205
XGBoost	0.807	0.298	0.253	0.361	0.339	0.331	0.391	0.366	0.366
GBRT	0.827	0.279	0.260	0.337	0.353	0.277	0.365	0.348	0.348
Balanced Bagging	0.779	0.172	0.225	0.312	0.330	0.118	0.321	0.330	0.217
Stacked Classifier	0.843	0.314	0.297	0.391	0.375	0.321	0.442	0.410	0.410
DESRRC	0.867	0.378	0.377	0.463	0.412	0.412	0.527	0.466	0.466

The table presents the performance of each ML model applied to our study's sample of serial offenders, spanning the test period from 2018 to 2020. The evaluation metrics featured in the table include the AUC (Area under the ROC) score, MCC (Matthew's Correlation Coefficient), AP (Average Precision) score, NDCG@k (Normalized Discounted Cumulative Gain at k), Recall@k and Precision@k. A higher value for any of these metrics signifies a more effective model.

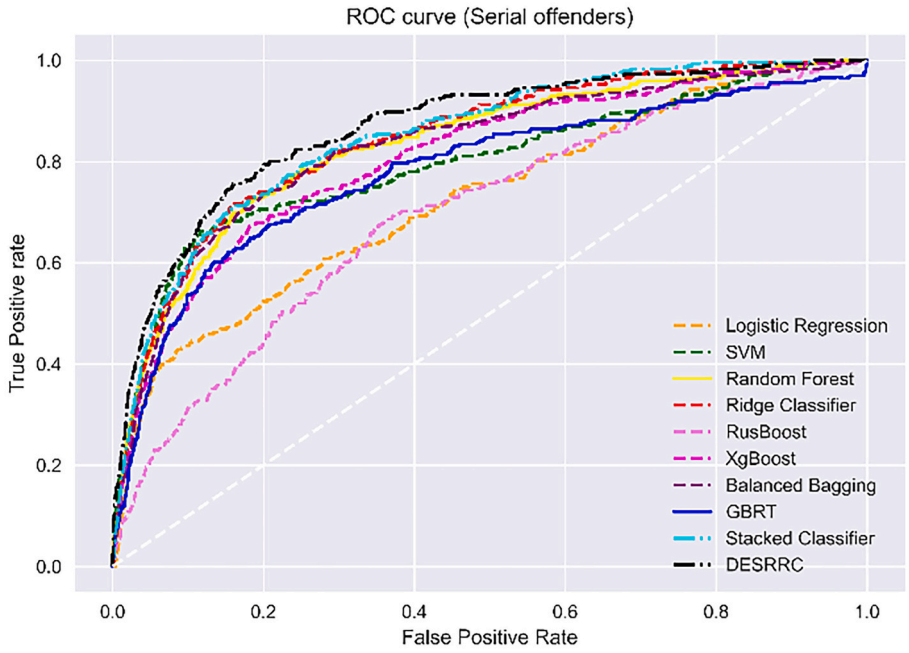

Fig. 5. ROC curve (Serial offenders) for the test period 2018–2020.
The ROC curve illustrates a classifier's performance by plotting the TPR against the FPR at different classification thresholds. A model with better performance will have a curve closer to the top-left section of the graph. The white line depicted in the graph symbolizes the performance of a random guess model.

using an alternative test period of 2016–2017. With the training period now shortened to 2007–2015,[8] we anticipate a decrease in the performance of the models. Table 5 displays the results for all firms during the 2016–2017 test period. Our findings align with the primary outcomes of our study, as the DESRRC model continues to surpass all other models across various evaluation metrics. We also observe that the performance of the Stacked Classifier model is again less consistent when using the alternative period. In contrast, the DESRRC model's performance remains stable across the performance metrics, illustrating the reliability of our model.

Comparable results are documented for the serial offenders' subsample in the 2016–2017 time period, as shown in Table 6. Our findings remain consistent, with the DESRRC model persistently outperforming all other models. These outcomes indicate that the DESRRC model exhibits robustness across our study's diverse performance evaluation metrics and time frames.

6. Conclusion

Financial statement fraud detection is a challenging activity. The consequences of fraudulent activities have severe implications, which include a higher cost of capital, distrust in capital markets, and significant costs to economies. As a result, timely fraud detection has become an important topic of interest to investors, practitioners, and academics. The main objective of this study was to develop an out-of-sample fraud detection model that used a comprehensive list of variables identified in recent studies. Our study used raw-financial and non-financial variables to build our model, which is the first model to do so. We also used the latest ML algorithms to build our fraud detection model and found that the DESRRC algorithm can lead to improvements in fraud detection. Furthermore, to show the robustness and unbiasedness of our results, we used several performance evaluation metrics and different time periods on all the fraud firms and serial offenders. Our results remained consistent throughout the study, and the DESRRC model performed better across the various performance metrics. Therefore, it can be used as an effective fraud detection model. Moreover, we also believe that the

[8] Similar to our main model, we set the train data at 2007–2013, validation dataset at 2014–2015. Once we fine-tune the hyper-parameters of each model, we then combine the data from 2007 to 2015 to retrain our models and use it to test the out-of-sample period of 2016–2017.

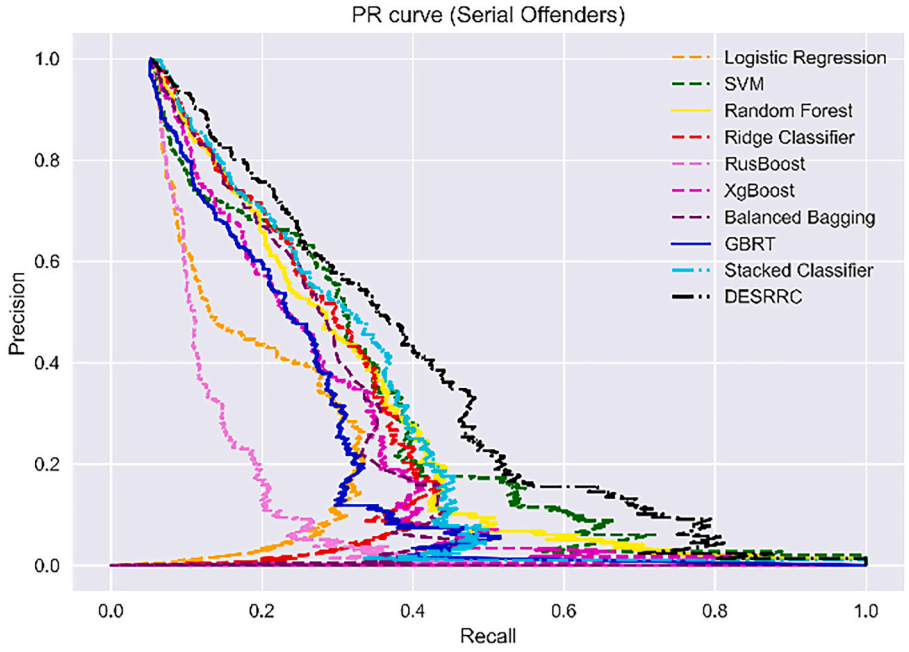

Fig. 6. PR curve (Serial offenders) for the test period 2018–2020.
The PR curve represents a classifier's performance by plotting the Precision against the Recall at various classification thresholds. The closer the curve is to the top-right part of the graph, the better the model's performance.

Table 5
Model performance of all firms (2016–2017).

Models	AUC	MCC	AP	NDCG@ 16.4%	Precision@ 16.4%	Recall@ 16.4%	NDCG@ 11%	Precision@ 11%	Recall@ 11%
Logistic Regression	0.616	0.176	0.262	0.328	0.333	0.293	0.344	0.343	0.343
SVM	0.610	0.133	0.244	0.304	0.269	0.239	0.295	0.343	0.343
Random Forests	0.659	0.212	0.318	0.379	0.381	0.297	0.408	0.390	0.390
Ridge Classifier	0.674	0.207	0.336	0.386	0.360	0.345	0.442	0.421	0.421
RusBoost	0.651	0.208	0.317	0.372	0.344	0.344	0.421	0.395	0.395
XGBoost	0.605	0.197	0.267	0.330	0.368	0.280	0.364	0.395	0.395
GBRT	0.631	0.121	0.274	0.323	0.401	0.092	0.377	0.401	0.138
Balanced Bagging	0.634	0.171	0.290	0.351	0.327	0.327	0.427	0.376	0.376
Stacked Classifier	0.646	0.227	0.324	0.375	0.360	0.336	0.438	0.418	0.418
DESRRC	0.703	0.266	0.388	0.438	0.410	0.377	0.494	0.453	0.453

The table presents the performance of each ML model applied to the sample of all firms in our study, spanning the test period from 2016 to 2017. The evaluation metrics featured in the table include the AUC (Area under the ROC) score, MCC (Matthew's Correlation Coefficient), AP (Average Precision) score, NDCG@k (Normalized Discounted Cumulative Gain at k), Recall@k and Precision@k. A higher value for any of these metrics signifies a more effective model.

Table 6
Model performance of serial offenders (2016–2017).

Models	AUC	MCC	AP	NDCG@ 5.3%	Precision@ 5.3%	Recall@ 5.3%	NDCG@ 2.9%	Precision@ 2.9%	Recall@ 2.9%
Logistic Regression	0.753	0.203	0.197	0.279	0.338	0.169	0.297	0.338	0.314
SVM	0.650	0.138	0.175	0.198	0.194	0.154	0.223	0.205	0.205
Random Forests	0.746	0.188	0.169	0.268	0.278	0.188	0.299	0.314	0.314
Ridge Classifier	0.744	0.239	0.201	0.310	0.306	0.263	0.370	0.395	0.395
RusBoost	0.680	0.149	0.166	0.257	0.225	0.225	0.344	0.326	0.326
XGBoost	0.750	0.221	0.202	0.313	0.337	0.200	0.399	0.349	0.349
GBRT	0.678	0.083	0.116	0.163	0.200	0.163	0.140	0.200	0.116
Balanced Bagging	0.747	0.173	0.155	0.235	0.219	0.219	0.277	0.267	0.267
Stacked Classifier	0.735	0.257	0.182	0.309	0.341	0.263	0.353	0.395	0.395
DESRRC	0.797	0.286	0.253	0.357	0.367	0.294	0.444	0.453	0.453

The table presents the performance of each ML model applied to our study's sample of serial offenders, spanning the test period from 2016 to 2017. The evaluation metrics featured in the table include the AUC (Area under the ROC) score, MCC (Matthew's Correlation Coefficient), AP (Average Precision) score, NDCG@k (Normalized Discounted Cumulative Gain at k), Recall@k and Precision@k. A higher value for any of these metrics signifies a more effective model.

DES methodology adopted in this paper can be extended to studies in accounting and finance, where enhancing the predictive performance of the models is vital.

Our study adds to the fraud detection literature by constructing an effective model. By using the sample of all fraudulent firms in the Chinese market, we are able to develop a fraud detection model which can effectively detect fraudulent firms in China. In addition, we try to match the fraud distribution in the US setting by focusing on serial offenders

and find that the DESRRC continues to perform well. We believe that our model can also be used in other emerging and developed markets to detect financial statement fraud. However, it would need to be tested empirically in these markets. Therefore, we suggest that future research be undertaken in other markets to see how the model performs. In addition, we also add to a growing list of literature that focuses on fraud detection in emerging markets like China, with high economic growth and a weaker regulatory regime. We believe our findings can help regulators and investors in these markets better understand financial fraud and the use of ML to detect it.

Appendix A. List of independent variables

No.	Financial statement variables	No.	Non-Financial variables
1	Common Shares Outstanding	1	MD&A Tone 1
2	Sale of Common and Preferred Stock	2	MD&A Tone 2
3	Total Income Taxes	3	Internal Control Index
4	Depreciation and Amortization	4	Chairman Tenure
5	Cost of Goods Sold	5	Top 10 auditors
6	Net Income	6	Foreign Investors
7	Sales	7	Institutional Shareholding
8	Prepaid Expenses	8	Age of the firm
9	Employee Benefits Payable	9	State-owned Enterprise
10	Income Taxes Payable	10	Herfindahl Index Top 10 Shareholders
11	Other Receivables	11	Top 2–10 Shareholding
12	Property, Plant and Equipment	12	Shareholding Ratio of the Largest Shareholder
13	Total Current Liabilities	13	Total Number of Committees Established
14	Total Common Equity	14	Number of Shareholder's Meetings
15	Retained Earnings	15	Number of Board Meetings
16	Intangible Assets	16	Manager Sum Salary Ratio
17	Account Payables	17	Top 3 Manager's Remuneration Ratio
18	Total Inventories	18	Top 3 Director's Remuneration Ratio
19	Total Receivables	19	CEO's Shareholding
20	Total Short-Term Investments	20	Chairman's Shareholding
21	Total Liabilities	21	Board Size
22	Investment and Advances	22	Concurrent Position (CEO/Chairman)
23	Total Assets	23	Proportion of Independent Directors
24	Total Current Assets	24	Attendance Rate of Shareholder's Meeting
25	Annual Price Close		
26	Cash and Short-Term Investments		

References

Bao, Y., Ke, B., Bin Li, Y., Julia, Y., & Zhang, J. (2020). Detecting accounting fraud in publicly traded US firms using a machine learning approach. *Journal of Accounting Research, 58*(1), 199–235. https://doi.org/10.1111/1475-679X.12292

Beneish, M. D. (1999). Incentives and penalties related to earnings overstatements that violate GAAP. *The Accounting Review, 74*(4), 425–457. https://doi.org/10.2308/accr.1999.74.4.425

Bertomeu, J. (2020). Machine learning improves accounting: Discussion, implementation and research opportunities. *Review of Accounting Studies, 25*(3), 1135–1155. https://doi.org/10.1007/s11142-020-09554-9

Bertomeu, J., Cheynel, E., Floyd, E., & Pan, W. (2021). Using machine learning to detect misstatements. *Review of Accounting Studies, 26*(2), 468–519. https://doi.org/10.1007/s11142-020-09563-8

Brown, N. C., Crowley, R. M., & Brooke Elliot, W. (2020). What are you saying? Using topic to detect financial misreporting. *Journal of Accounting Research, 58*(1), 237–291. https://doi.org/10.1111/1475-679X.12294

Cecchini, M., Aytug, H., Koehler, G. J., & Pathak, P. (2010). Detecting management fraud in public companies. *Management Science, 56*(7), 1146–1160. https://doi.org/10.1287/mnsc.1100.1174

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*(3), 424–448. https://doi.org/10.1016/j.jcorpfin.2005.09.002

Chen, Y., Zhu, S., & Wang, Y. (2011). Corporate fraud and Bank loans: Evidence from China. *China Journal of Accounting Research, 4*(3), 155–165. https://doi.org/10.1016/j.cjar.2011.07.001

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics, 21*(1), 6. https://doi.org/10.1186/s12864-019-6413-7

Conyon, M. J., & He, L. (2016). Executive compensation and corporate fraud in China. *Journal of Business Ethics, 134*(4), 669–691. https://doi.org/10.1016/j.jcorpfin.2011.04.006

Cruz, R. M. O., Sabourin, R., & Cavalcanti, G. D. C. (2018). Dynamic classifier selection: Recent advances and perspectives. *Information Fusion, 41*, 195–216. https://doi.org/10.1016/j.inffus.2017.09.010

Dechow, P. M., Ge, W., Larson, C. R., & Sloan, R. G. (2011). Predicting material accounting misstatements*. *Contemporary Accounting Research, 28*(1), 17–82. https://doi.org/10.1111/j.1911-3846.2010.01041.x

DeFond, M. L., Raghunandan, K., & Subramanyam, K. R. (2002). Do non–audit service fees impair auditor independence? Evidence from going concern audit opinions. *Journal of Accounting Research, 40*(4), 1247–1274. https://doi.org/10.1111/1475-679X.00088

Dong, W., Liao, S., & Zhang, Z. (2018). Leveraging financial social media data for corporate fraud detection. *Journal of Management Information Systems, 35*(2), 461–487. https://doi.org/10.1080/07421222.2018.1451954

Fawcett, T. (2006). An introduction to ROC analysis. *ROC Analysis in Pattern Recognition, 27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

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*(2), 371–390. https://doi.org/10.1016/j.jcorpfin.2010.09.002

Fu, C., Deng, X., & Tang, H. (2023). Who cares about corporate fraud? Evidence from cross-border mergers and acquisitions of Chinese companies. *Review of Quantitative Finance and Accounting, 60*(2), 747–789. https://doi.org/10.1007/s11156-022-01111-6

García, S., Zhang, Z.-L., Altalhi, A., Alshomrani, S., & Herrera, F. (2018). Dynamic ensemble selection for multi-class imbalanced datasets. *Information Sciences, 445–446*, 22–37. https://doi.org/10.1016/j.ins.2018.03.002

Gee, J., & Button, M. (2021). *The financial cost of fraud 2021*. Tech. Rep. Crowe (January 23, 2023) https://f.datasrvr.com/fr1/521/90994/0031_Financial_Cost_of_Fraud_2021_v5.pdf.

Gomes, H. M., Bifet, A., Read, J., Barddal, J. P., Enembreck, F., Pfharinger, B., … Abdessalem, T. (2017). Adaptive random forests for evolving data stream classification. *Machine Learning, 106*(9), 1469–1495. https://doi.org/10.1007/s10994-017-5642-8

Graham, J. R., Li, S., & Qiu, J. (2008). Corporate misreporting and bank loan contracting. *Journal of Financial Economics, 89*(1), 44–61. https://doi.org/10.1016/j.jfineco.2007.08.005

Hou, W.-h., Wang, X.-k., Zhang, H.-y., Wang, J.-q., & Li, L. (2020). A novel dynamic ensemble selection classifier for an imbalanced data set: An application for credit risk assessment. *Knowledge-Based Systems, 208*, Article 106462. https://doi.org/10.1016/j.knosys.2020.106462

Japkowicz, N. (2013). Assessment metrics for imbalanced learning. *In Imbalanced Learning, 187–206*. https://doi.org/10.1002/9781118646106.ch8

Jia, C., Ding, S., Li, Y., & Zhenyu, W. (2009). Fraud, enforcement action, and the role of corporate governance: Evidence from China. *Journal of Business Ethics, 90*(4), 561–576. https://doi.org/10.1007/s10551-009-0061-9

Karpoff, J. M., Scott Lee, D., & Martin, G. S. (2008). The cost to firms of cooking the books. *Journal of Financial and Quantitative Analysis, 43*(3), 581–611. https://doi.org/10.1017/S0022109000004221

Ko, A. H. R., Sabourin, R., & Jr. Britto Alceu Souza.. (2008). From dynamic classifier selection to dynamic ensemble selection. *Pattern Recognition, 41*(5), 1718–1731. https://doi.org/10.1016/j.patcog.2007.10.015

Kuncheva, L. I., & Whitaker, C. J. (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning, 51*(2), 181–207. https://doi.org/10.1023/A:1022859003006

Larcker, D. F., Richardson, S. A., & Tuna, İ. (2007). Corporate governance, accounting outcomes, and organizational performance. *The Accounting Review, 82*(4), 963–1008. https://doi.org/10.2308/accr.2007.82.4.963

Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research, 247*(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030

Liebman, B. L., & Milhaupt, C. J. (2008). Reputational sanctions in China's securities market. *Columbia Law Review, 108*(4), 929–983. http://www.jstor.org/stable/40041782.

Lin, C.-C., Chiu, A.-A., Huang, S. Y., & Yen, D. C. (2015). Detecting the financial statement fraud: The analysis of the differences between data mining techniques and Experts' judgments. *Knowledge-Based Systems, 89*, 459–470. https://doi.org/10.1016/j.knosys.2015.08.011

Luo, J.-h., Peng, C., & Zhang, X. (2020). The impact of CFO gender on corporate fraud: Evidence from China. *Pacific-Basin Finance Journal, 63*, Article 101404. https://doi.org/10.1016/j.pacfin.2020.101404

Mutlu, C. C., Van Essen, M., Peng, M. W., Saleh, S. F., & Duran, P. (2018). Corporate governance in China: A Meta-analysis. *Journal of Management Studies, 55*(6), 943–979. https://doi.org/10.1111/joms.12331

Niu, G., Li, Y., Fan, G.-Z., & Zhang, D. (2019). Corporate fraud, risk avoidance, and housing Investment in China. *Emerging Markets Review, 39*, 18–33. https://doi.org/10.1016/j.ememar.2019.03.003

Perols, J. (2011). Financial statement fraud detection: An analysis of statistical and machine learning algorithms. *Auditing: A Journal of Practice & Theory, 30*(2), 19–50. https://doi.org/10.2308/ajpt-50009

Purda, L., & Skillicorn, D. (2015). Accounting variables, deception, and a bag of words: Assessing the tools of fraud detection. *Contemporary Accounting Research, 32*(3), 1193–1223. https://doi.org/10.1111/1911-3846.12089

Ren, L., Zhong, X., & Wan, L. (2022). Missing analyst forecasts and corporate fraud: Evidence from China. *Journal of Business Ethics, 181*(1), 171–194. https://doi.org/10.1007/s10551-021-04837-w

Santos, D., Eulanda, M., Sabourin, R., & Maupin, P. (2008). A dynamic overproduce-and-choose strategy for the selection of classifier ensembles. *Pattern Recognition, 41*(10), 2993–3009. https://doi.org/10.1016/j.patcog.2008.03.027

Sergio, A. T., de Lima, T. P. F., & Ludermir, T. B. (2016). Dynamic selection of forecast combiners. *Neurocomputing, 218*, 37–50. https://doi.org/10.1016/j.neucom.2016.08.072

Sesmero, M. P., Iglesias, J. A., Magán, E., Ledezma, A., & Sanchis, A. (2021). Impact of the learners diversity and combination method on the generation of heterogeneous classifier ensembles. *Applied Soft Computing, 111*, Article 107689. https://doi.org/10.1016/j.asoc.2021.107689

Walker, S. (2020). A needle found: Machine learning does not significantly improve corporate fraud detection beyond a simple screen on sales growth. *SSRN*. https://doi.org/10.2139/ssrn.3739480

Wei, Y., Chen, J., & Wirth, C. (2017). Detecting fraud in Chinese listed company balance sheets. *Pacific Accounting Review, 29*(3), 356–379. https://doi.org/10.1108/PAR-04-2016-0044

Woloszynski, T., & Kurzynski, M. (2011). A probabilistic model of classifier competence for dynamic ensemble selection. *Semi-Supervised Learning for Visual Content Analysis and Understanding, 44*(10), 2656–2668. https://doi.org/10.1016/j.patcog.2011.03.020

Wolpert, D. H. (October 1, 1996). The lack of a priori distinctions between learning algorithms. *Neural Computation, 8*(7), 1341–1390. https://doi.org/10.1162/NECO.1996.8.7.1341

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*(4), 709–726. https://doi.org/10.1007/s10551-014-2392-4

Xiao, H., Xiao, Z., & Wang, Y. (2016). Ensemble classification based on supervised clustering for credit scoring. *Applied Soft Computing, 43*, 73–86. https://doi.org/10.1016/j.asoc.2016.02.022

Xu, W. (2022). Public enforcement initiated by the CSRC and its regional offices. In X. Wenming (Ed.), *The enforcement of securities law in China: A law and economics assessment* (pp. 49–80). Singapore: Springer Singapore. https://doi.org/10.1007/978-981-19-0904-7_4.

Zhou, Z., 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*, Article 100817. https://doi.org/10.1016/j.ememar.2021.100817