International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org **Original Research Paper**

Uncover and Identify Accounting Frauds in Publicly Traded Firms Using Machine Learning Techniques

Siddharth Nanda[1]*, Dr. Vinod Moreshwar Vaze[2]

Submitted: 31/12/2023 **Revised**: 03/02/2024 **Accepted**: 11/02/2024

Abstract: Financial fraud has increased dramatically along with the rise of advanced technologies and worldwide connection. There are several types of financial fraud, each with its unique characteristics. This paper focuses on detecting accounting fraud in publicly traded firms. This study proposed a framework for financial fraud prediction and detection using machine learning (ML). This study utilized single ML models like Logistic Regression (LR), Naïve Byes (NB), Extreme Gradient Boosting (XG-BOOST), and ensemble techniques to identify fraud. Each classifier was assessed for accuracy, recall, precision, and testing and training time. The proposed ensemble classifier, which includes NB, LR, and XGBOOST, outperformed the single models by achieving accuracy, precision, and recall of 99.46%, 99.6%, and 99.82%, respectively. The findings suggest that the proposed ensemble model can forecast financial fraud more precisely and efficiently than other classifiers.

Keywords: *Financial fraud, Fraud detection, fraud detection system, Machine learning, ensemble model*

1. Introduction

There has been an enormous increase in fraud in recent years, with negative consequences for both financial firms and their clients. The identification of accounting fraud in publicly listed organizations is one of the most exciting and demanding applications of ML in computational finance [1]. There are many different sectors in the financial sector where ML is being used to enhance operations, such as fraud detection, payment processing, and regulation [2]. Insiders (including managers and controlling owners) committing accounting fraud at publicly traded companies is a global issue [3]. Suppose frauds are not discovered and stopped promptly. In that case, they have the potential to inflict enormous damage to the stakeholders of the companies that are directly involved in the frauds (for example, Enron and WorldCom) [4]. As a result of competing with fraudulent enterprises for limited financial resources and customer spending, genuine businesses may be affected indirectly by fraudulent ones [5]. In addition, the financial market's ambiguity regarding the occurrence of frauds could hinder the proper functioning of financial markets and economic development owing to information asymmetry among company insiders and outside investors [6]. Fraud in the accounting profession is very difficult to uncover. And even if it does, by the time the problem is identified, the damage is usually already done to a considerable extent [7]. Therefore, regulators, auditors, and investors would all benefit significantly from having access to methodologies that are both efficient and effective for detecting corporate accounting fraud [8].

Financial statement fraud is the rarest but most expensive fraud, with the Association of Certified Fraud Examiners (2020) estimating annual fraud losses for firms at 5% of overall revenue, or $4.5 trillion [9]. Investors, regulators, and auditors would all benefit from the timely discovery of accounting fraud to save related expenses [8]. Two major concerns are to be considered. First, large sample sizes are required because of the significant class imbalance between identified fraud and non-fraud instances and "partial observability of fraud” [10]. The second issue is that knowledge asymmetry about organizational behavior is not completely reflected in aggregate financial data [11]. When it comes to publicly listed companies, accounting fraud may have far-reaching economic implications, endanger the financial security of investors, and affect market confidence. This study aims to address these issues and develop an efficient method for the detection of financial fraud by using the ensemble learning approach, which is one of the most potent ML approaches, instead of the more standard method of logistic regression. For evaluating the efficacy of fraud prediction algorithms, this study proposes a novel performance assessment measure inspired by ranking problems but better suited to fraud prediction tasks. The study begins with the same theoretically motivated raw accounting data and demonstrates that the proposed fraud detection model significantly outperforms various ML classifiers, including LR, NB, XGBOOST, etc.

This study employs a live implementation of an ML method to uncover accounting fraud in publicly traded businesses. This study is structured as follows: Section 2 provides a literature review of relevant prior work; Section 3 provides a brief description of the dataset used in this study; Section 4 defines the technique used in this study, which is ML, along with the proposed methodology. In Section 5, the result and comparative analysis have been done based on various performance metrics. In section 6 conclusion and future have been explained [13].

1.1. Financial Fraud

Bank robbery does have far-reaching consequences for the investing industry in daily life. If a person engages in misleading

1 Research Scholar, Department. of CSE, Shri JJT University, Jhunjhunu, Rajasthan, India

2 Guide, Department. of CSE, Shri JJT University, Jhunjhunu, Rajasthan, India

** Corresponding Author Email: siddnanda89@gmail.com*

behavior with the malicious motive of gaining some kind of illegal benefit, then it is called fraudulent behavior. The goal of every fraudulent act is to get some benefit or value at the expense of another party. Changing tax and insurance records and faking sales are the most common types of fraud in the real estate market. Even though such actions are uncommon, they are often carried out by people, organizations, and even businesses. The confidence in an industry, people's ability to save money, and the cost of living can all be impacted by these kinds of fraud [14]. Financial institutions utilize numerous anti-fraud strategies. Criminals who commit fraud are resourceful and continually devise novel strategies to defeat security measures. Crimes against the economy continue to be committed despite the efforts of banks, law enforcement, and the government. Criminals operating in today's world may be a highly inventive, exceptionally intelligent, and lightning-fast bunch. In this study, several techniques for detecting fraud are compared. It's also used to explore large datasets for hidden patterns. In addition, because of the exponential increase in fraud that annually affects the financial industry, fresh methods of detecting fraud are constantly being created and used in other industries [15]. Figure 1 shows the overall categorization scheme of financial fraud.

The various types of fraud can be explained as follows:

1. Security and exchange commodities fraud: Connell University Law School (CULS) classifies market manipulation, securities account theft, and wire fraud as types of security fraud [16]. Market manipulation, high-yield investment fraud, commodity fraud, foreign currency fraud, after-hours trading, broker theft, etc., are just some of the many services it offers.
2. Bank fraud: It is carried out with the intent to deceive a financial institution or to fraudulently gain funds, assets, credits, securities, or other property belonging to such an entity. Criminal offenses include mortgage fraud, money laundering, etc.
3. Insurance fraud: These are the ones that arise in the middle of the insurance procedure. Healthcare providers, patients, agents, brokers, and staff members could be affected at any point in the application, billing, rating, claims, and eligibility processes.
4. Other related financial fraud: These include the categories mentioned above, such as corporate fraud and mass marketing fraud [17].

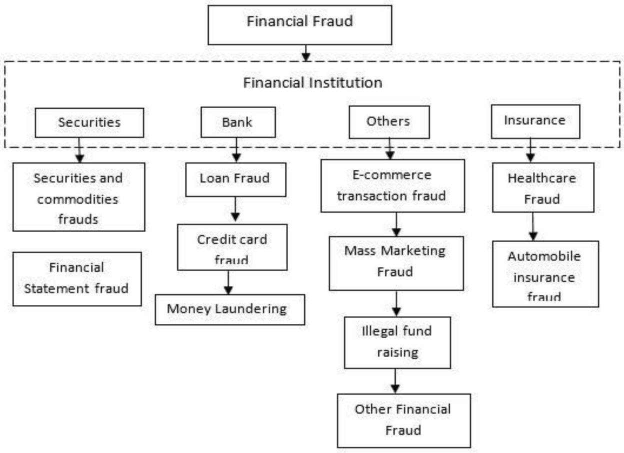

Fig. 1: Financial fraud classification [18].

1.2. Significance of Machine Learning in Fraud Detection

Traditional methods of fraud detection take up a lot of time. Hence, there should be some AI models for finding and blocking financial fraud [19]. Many computational intelligence-based methods may be found among these intelligence-gathering methods. Both supervised and unsupervised ML techniques are used in the fraud detection system. On the other hand, the supervised technique depends on the transaction based on fraudulent and legal and then classifications freshly happened transactions based on the learned model, in contrast to the unsupervised model of identifying fraud, which focuses on transactions that fall in outliers. Backpropagation of mistakes is an example of an algorithm used to identify fraud that employs both forward and backward passes [20].

ML improves data management and data processing. Recent research teaches robots autonomy. Robots must learn and make smart judgments. Mathematicians & computer scientists tried several solutions [21]. Supervised learning enables computers to anticipate and classify tagged datasets. This helps robots adapt old data to new data. Supervised learning algorithms find data patterns, linkages, and correlations using mathematical models for precise forecasts and judgments. Unlabeled datasets teach computers. Algorithms find unlabeled patterns. Unsupervised learning lets robots find hidden patterns, synthesize data, and spot abnormalities. It helps experimental data analysis and categorization [22]. Reinforcement learning teaches robots. Robots learn reinforcement. Reward and punishment drive them. Mathematical algorithms and models help robots make decisions and reach objectives [23].

Deep learning is trendy. Multi-layered artificial neural networks reveal complicated patterns and representations from difficult data. Computer vision, Natural Language Processing (NLP), and robotics are its strengths. ML teaches computers supervised, unsupervised, reinforcement, and deep learning. These methods let robots learn, improve, and make data-handling judgments. ML approaches help intelligent computers organize and use data in numerous fields [24]. Figure 2 shows the ML types.

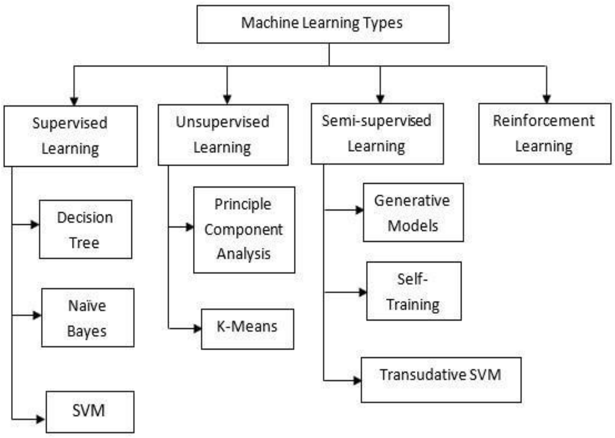

Fig. 2: Various types of ML techniques [24]

1.2.1 Supervised Learning (SL)

ML is frequently used to train a function that converts input to output from example, input-output pairs [25]. It learns a function by comparing it to labelled training data and a collection of examples. S.L. is implemented in a task-driven setting, where certain outputs are sought from a given set of inputs [26]. Common supervised tasks include "classification," which involves grouping data, and "regression," which involves transforming data. An application of S.L. is text categorization, which involves determining the likely category or opinion of a piece of text, such as a tweet or a review of a product [27].

2. Related Works

This section provides an overview of the literature on the topic of employing an ML technique to identify accounting fraud in publicly traded businesses.

Zhao et al., (2022) [28] suggested a novel ML-based approach to combat financial fraud among publicly traded corporations. Five individual classification models and three ensemble models were developed to forecast the financial fraud records of listed businesses. After evaluating five different ML techniques, a single best model among them and an overall best ensemble model is selected. According to the findings, the ideal single model achieved an accuracy of 97% to 99%, while the ensemble models (both Logistic regression and XGBOOST) achieved an accuracy of above 99%. This demonstrates the superior performance of the optimum ensemble model and its ability to identify and predict corporate fraud.

Chen et al., (2022) [29] offered a framework for anticipating financial fraud using fundamental financial data. The technique relies heavily on the principal component analysis Random Forest (PCA-RF) approach. In this study, authors apply an ensemble learning strategy to the prediction of financial fraud for publicly traded firms, therefore introducing the PCA-RF technique. The investigation shows that the PCA-RF model is better at predicting domestic financial fraud in China compared to R.F. and neural network approaches.

Fukas et al., (2022) [30] suggested employing Generative Adversarial Networks (GANs) to construct synthetic fraud cases using a dataset of publicly traded companies that the U.S. Securities and Exchange Commission has fined for accounting malfeasance. The goal of this method is to train a logit, Support Vector Machine (SVM), or XG-Boost classifier on a more evenly distributed dataset to improve the classifier's prediction accuracy. Findings show that existing ML models like XG-Boost can beat legacy fraud detection methods on the same data; however, training an ML method on simulated fraud instances does not provide better results.

Pranto et al., (2022) [31] offered a blockchain and smart contract-based method to develop a powerful ML system for detecting fraudulent online purchases. The performance of the blockchain network is evaluated by subjecting it to varying degrees of difficulty and diverse data loads. After eight iterations, the model's F-beta score was 98.22%, and its testing accuracy was 98.93%. The results demonstrate that blockchain mining time is affected by both the amount of data and the difficulty level.

Ileber et al., (2022) [32] suggested the Genetic Algorithm (GA) for feature selection for use in an ML-based credit card fraud detection engine. The suggested detection engine employs the ML classifiers Decision Tree (DT), RF, LR, Artificial Neural Network (ANN), and NB once the optimum features have been selected. The effectiveness of the suggested credit card fraud detection engine is verified by testing it on a dataset produced by European cardholders. Recommended the Genetic Algorithm (GA) for feature selection in a machine learning (ML) based credit card fraud recognition system. Once the best features have been chosen, the recommended detection system uses ML classifiers such as DT, RF, LR, ANN, and NB. The proposed credit card fraud detection engine is validated by testing it on a dataset built from European cardholders, demonstrating its efficacy. The outcome confirmed that the proposed method is superior to the status standard.

Hassanniakalager et al., (2022) [33] developed a novel ML-based fraud detection model in the field of accounting. They term the LR-enhanced ensemble learning model Logit-Boost. In forecasting fraud that occurs beyond the current accounting period, the model performs better than the others. Importantly, the method authors use fewer predictors than those in earlier ML studies, which always worries about multicollinearity and possible overfitting caused by ML techniques.

Sánchez-Aguayo et al., (2022) [34] suggested a method for identifying instances of possible fraud by studying how individuals interact with data. To create an alert from potentially fraudulent material, this method combines a predetermined topic model with a supervised classifier. They compare the performance of several topic modeling methods and supervised and Deep Learning (DL) classifiers and conclude that Linear Discriminant Analysis (LDA), R.F., and Convolutional Neural Networks (CNN) perform the best. The study suggests that the method can be implemented since multiple examples of such models achieve an average Area Under Curve (AUC) greater than 0.8.

Saheed et al., (2022) [35] proposed a new model for credit card fraud detection (CCFD) using PCA for financial security and supervised ML techniques to categorize transactions as fraudulent or not. The effectiveness of the proposed method in detecting fraudulent transactions is demonstrated when compared to previous. The model is developed to be robust by using PCA's prowess in identifying the best predictive characteristics. Data sets from Germany and Taiwan were used for the experimental analysis. Based on the testing results, the KNN is the highest-performing model for the German data set, with 96.29% accuracy, 100% recall, and 96.29% precision. With an accuracy of 81.75%, recall of 34.89%, and precision of 66.61%, the ridge classifier was the top-performing model on Taiwan Credit data.

Liu et al., (2021) [36] investigated a reliable and understandable approach for detecting financial scams. When compared to Smoothly Clipped Absolute Deviation (SCAD), Minimax Concave Penalty (MCP), Stepwise, and Lasso algorithms using the model confidence algorithm (MCB), They discovered that Adaptive Lasso was the most stable. When compared to classic models like SVM and L.R., integrated models like XG-Boost, Light-GBM, and R.F. demonstrated superior performance in detecting financial fraud.

Hamal et al., (2021) [37] analyzed the financial accounts of 341 Turkish SMEs between 2013 and 2017 to discover how well ML classifiers can detect financial accounting fraud. Seven distinct classifiers (SVM, NN, k-nearest neighbor, RF, L.R., and bagging) are implemented and evaluated concerning their performance measures. In addition, they compare classifiers without any feature selection or sampling methods. According to the findings, the best-performing model is the R.F. without feature selection-oversampling model.

Mqadi et al., (2021) [38] applied to oversample using Synthetic Minority Oversampling Technique (SMOTe), which is a data-point method to an unbalanced credit card dataset. The classifications were carried out using state-of-the-art classical ML methods such as SVM, LR, DT, and RF classifiers, and accuracy was measured with precision, recall, and F1-score along with the average precision. The findings demonstrate that the model has difficulty identifying fraudulent transactions when the data is substantially asymmetrical. The accuracy of positive class predictions was greatly enhanced by employing the SMOTe-based Oversampling method. The findings suggest that among the available algorithms, RF is the most effective. RF, DT, LR, and SVM are the finest algorithms in that order.

Trivedi et al., (2020) [39] introduced an ML-based feedback system for detecting credit card fraud. Its cost-effectiveness and detection rate improvements can be attributed to the classifier's feedback technique. The purpose of this study was to compare the efficiency of several approaches using imprecise credit card fraud data collection. These methods included RF, tree classifiers, ANNs, SVM, NB, LR, and XG-BOOST classifier algorithms. Precision, recall, F1-score, accuracy, and FPR % have traditionally been used as the only performance assessment metrics for comparing classifiers' efficacy. According to the results, random forest methods have a 95.988% rate of accuracy.

Nguyen et al., (2020) [40] conducted a comprehensive study on deep learning approaches to address the credit card fraud detection problem, along with its comparisons to other ML algorithms and the results from tests conducted on three discrete financial datasets. The proposed DL approaches outperform traditional ML models, as shown by experimental results, highlighting their potential usage in real-world credit card fraud detection systems. A 50-block Long Short-Term Memory (LSTM) outperformed the other algorithms with an F1-Score of 84.85% in a head-to-head comparison.

Thennakoon et al., (2019) [41] focused on the detection of four major instances of fraud in actual financial dealings. Multiple ML models, such as LR, NB, LR, and SVM, are applied to each fraud scenario before the most effective model is chosen. The study utilized a suitable performance measure that provides thorough guidance to select an ideal algorithm about the kind of fraud. The report also discusses the crucial problem of real-time credit card fraud detection. The study uses the predictive analytics provided by the installed ML models and an API module to ascertain whether a specific transaction is fraudulent. The study used data obtained from a financial institution under the terms of a nondisclosure agreement, and the four fraud patterns were best caught by LR, NB, k-nearest neighbor (KNN), and SVM by achieving accuracies of 74%, 83%, 72%, and 91% respectively.

Raghavan et al., (2019) [42] compared several popular machines learning techniques, including KNN, RF, and SVM, with popular DL techniques, including autoencoders, CNNs etc. The study demonstrated that SVMs, maybe in conjunction with CNNs for improved performance, are among the best approaches for detecting fraud when applied to huge datasets. Ensemble methods like SVM, Random Forest, and KNNs might yield useful improvements, especially for smaller datasets. Among the several deep learning techniques, CNNs are typically the most effective.

3. Techniques Used

In this part, the author discussed the technique used in this study for detecting accounting fraud. In this paper, the ML technique, including Supervised learning (L.R., R.F., and SVM), is proposed.

3.1. Logistic Regression

The ML method of L.R. is called S.L. The sigmoid function (logarithmic) provides the foundation since it takes a real number and returns a value between 0 and 1. An N-sample training dataset is given to the algorithm during the training phase. Each instance has certain X properties and a Y label that specifies the way it should be categorized. As a result of the training process, the system generates a model that can be used to label data that was not involved in the training set [44][45].

3.2. Naïve Byes

The Naive Bayes approach is a kind of Bayesian statistics in which the most likely outcome is used for making predictions. The probability of the unknown value is estimated using the known value. This classifier uses its previous knowledge for prediction. Naive Bayes relies heavily on categorical data and conditional probabilities.

$$p(fe_k) = \frac{p(c_j) * p(c_j)}{p(fe_k)}$$

$$p(c_j) = \prod_{j=1}^{m} p(c_j)$$

Maximum number of features is denoted by m in Eq. (1) and Eq. (2) the probability of producing feature value fe_k given in class c_j is denoted by $p(c_j)$; and the probabilities of occurrence of feature value and class are denoted by $p\ (fe_k)$ and $p(c_j)$, respectively. In order to do binary classification using the Bayesian principle, this classifier was implemented [46].

3.3. XG-Boost

XG-Boost is a scalable and computationally effective version of gradient-enhanced decision trees, which are used to incrementally construct additive models. By incorporating additive models based on shortcomings discovered in the preceding processes, the overall error is gradually lowered. The resulting outperform base learners outperform the individual classifiers in terms of predictive ability. This is accomplished by ensuring that each of the individual learners contributes equally to the final composite model while simultaneously increasing its accuracy and decreasing its tree depth [47].

A random sampling strategy was added to it to make it more resistant to noise and overfitting. To prevent overfitting, XG-Boost uses an increasingly regularized model than previous implementations. Here is the minimized XG-Boost objective function [48]

To increase model performance, a greedy decision tree model is added, denoted by f_t (x_i), and penalize model complexity with the regularization term Ω (f_t), where y_i is the desired outcome for the i-th instance, and yˆ_i^((t)) is its predicted value at the t-th iteration. In addition to its effectiveness in other sectors, such as insurance fraud detection, XGBoost is now one of the highest-performing classifiers overall [49].

4. Proposed Methodology

See figure 3 for a flowchart of the suggested method, which is centered on identifying accounting fraud in publicly listed corporations using an ML technique with live implementation.

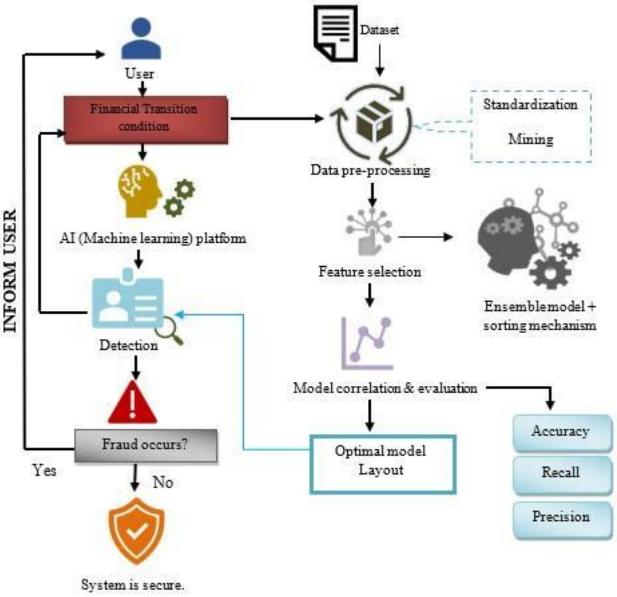

Fig. 3: Block Diagram of Proposed Fraud Accounting Detecting Model

5. Results and Discussion

5.1. Tools Used

In this study, ML models are based on Python 3.10. The machine consists of an i5 processor, 8GB RAM, and google collab to write and run code.

5.2. Performance Analysis

5.2.1. Training and Testing Dataset

The entire dataset has been split up into two parts: the training data set, which has around 14,613 record items, of which 13458 are authentic records, and the test data set, which has 1151 fake records. The testing dataset has around 3,447 records in it, including 2,413 records that are authentic and 1034 fake records, as shown in Table 1.

Table 1 Training and testing dataset

Dataset	**No. of authentic records**	**No. of fraud records**
Training data set	13458	1155
Testing data set	2,413	1034

5.2.2. Confusion Matrix

It can be seen from Table 2 given below that the proposed ensemble model successfully recognized fraud events with the highest number of true positives using Logistic Regression, Naive Bayes, and XG-BOOST. There were very few false positives and negatives. By achieving a better balance between true positives, false positives, and false negatives, the ensemble model outperforms the other classifiers. Figure 4 illustrates the confusion matrix of Logistic Regression, Naive Bayes, and XG-BOOST AND ensemble model.

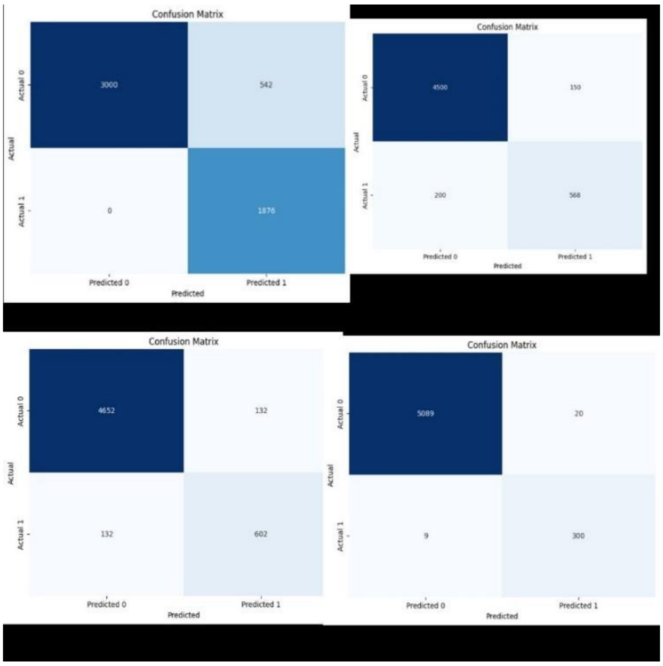

Fig. 4: Confusion matrix for: (a) LR (b) NB (c) X-GBOOST (d) proposed ensemble model

Table 2 Confusion matrix of ML models

Algorithm	**TP**	**FP**	**FN**	**TN**
LR	3000	542	0	1876
NB	4500	150	200	568
XG-BOOST	4652	132	132	602
Proposed Ensemble model (LR+NB+XGBOOST)	5089	20	9	300

5.2.3. Performance Evaluation

Table 3 below compares the accuracy of several fraud-prediction classifiers. The training period using Logistic Regression was larger, but its 90.3% accuracy and strong recall made up for it. With improved precision and recall but more time spent in training, Naive Bayes was able to reach an accuracy of 93.54%. In addition to its speed improvements, XG-BOOST's 96.97% accuracy was the greatest of any method tested. The Ensemble model outperformed the others, although it required more time to train and test due to its high accuracy of 99.46%, precision of 99.60%, and recall of 99.82%. Overall, it can be observed that the Ensemble model outperforms the other models.

Table 3 Performance metrics of ML models

Technique	**Accuracy**	**Precision**	**Recall**	**Testing time**	**Training time**
LR	90.3	84.69	99.99	0.01s	0.49s
NB	93.54	96.77	95.74	0.01s	1.34s
XG-BOOST	96.97	97.24	97.24	0.02s	0.39s
Proposed Ensemble model	99.46	99.60	99.82	0.03s	1.15s

5.2.4. Comparative Analysis

The results obtained by several studies for the purpose of financial fraud detection using ML classifiers are compared as given in Table 4. Kaur et al., [50] used three models; among them, Naive Bayes (NB) had 86.36% accuracy, 95.88% precision, and 86.79% recall, missing some fraudulent transactions. The KNN model outperformed with 97.14% accuracy, 98.87% precision, and 97.95% recall. MLP had 84.22% accuracy, decent precision (89.89%), and recall (64.29%), but the study conducted by Zhao et al., [28] used Logistic Regression (LR) + XG-Boost model and obtained 98.52% accuracy, 99.02% precision, and 99.49% recall. The suggested ensemble model (NB + LR + XG-Boost) detected financial fraud with the greatest accuracy of 99.46%, precision of 99.6%, and recall of 99.82%, as shown in Figure 5.

Table 4 Comparative Analysis.

Author	Technique	Accuracy (%)	Precision (%)	Recall (%)	Testing time (s)	Training time (s)
Kaur et al., [50]	NB	86.36	95.88	86.79	0.52	2.38
	KNN	97.14	98.87	97.95	1.13	0.12
	MLP	84.22	89.89	64.29	0.01	17.84
Zhao et al., (2022) [28]	LR+XGBOOST	98.52	99.02	99.49	0.01	1.54
Proposed method	NB+LR+ XGBOOST	99.46	99.60	99.82	0.03	1.15

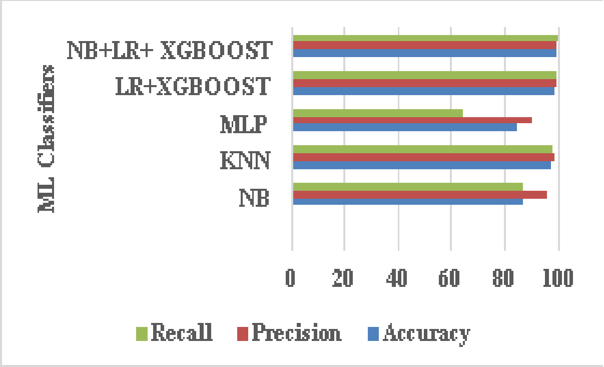

Fig. 5: Comparative analysis of proposed with existing models.

6. Conclusion and Future Scope

An ML approach was presented in the study to help detect and prevent financial fraud. In this study, the author employed a single ML model for fraud detection, such as LR, NB, XGBOOST, and ensemble methods. Finally, the study recorded the amount of time spent on testing and training each classifier, along with several performance criteria such as accuracy, recall, and precision. When compared to a single model, the suggested ensemble classifier clearly excels. The proposed ensemble model (NB+LR+XGBOOST) model performed best of all by achieving accuracy, precision, and recall percentages of 99.46%, 99.6%, and 99.82%, respectively. This indicates the ensemble model is superior to others in its ability to forecast whether a company has committed financial fraud.

According to the findings, the parameters used by the ML algorithms in this study are appropriate for detecting financial fraud. This indicates that the most financially troubled businesses may be pinpointed with this method. In addition, it is crucial for businesses operating in the financial industry to clear up any confusion their employees may have about financial reports or initiatives. As a result of the booming economy, more and more businesses are going public every year. Further, using ML to determine if publicly traded firms have financial fraud issues can significantly ease the burden on employees.

Acknowledgements

The author would like to express his sincere gratitude to his department for their invaluable contributions to this research paper. Their expertise and insights have been instrumental in shaping the direction of this study. The author extends our heartfelt thanks to all those who have contributed to the successful completion of this research.

Author contributions

Siddharth Nanda[1]: Conceptualization, Methodology, Software, Field study Visualization, Investigation **Dr. Vinod Moreshwar Vaze[2]:** Data curation, Writing-Original draft preparation, Software, Validation., Field study, Writing-Reviewing and Editing.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] B. Li and S. C. H. Hoi, “Online portfolio selection: A survey,” ACM Comput. Surv., vol. 46, no. 3, pp. 1-36, 2014 [doi:10.1145/2512962].

[2] S. Emerson et al., “Trends and applications of machine learning in quantitative finance” in 8th international conference on economics and finance research (ICEFR 2019), 2019.

[3] P. Ravisankar et al., “Detection of financial statement fraud and feature selection using data mining techniques,” Decis. Support Syst., vol. 50, no. 2, pp. 491-500, 2011 [doi:10.1016/j.dss.2010.11.006].

[4] A. Abbasi et al., “Metafraud: A meta-learning framework for detecting financial fraud,” MIS Q., vol. 36, no. 4, pp. 1293-1327, 2012 [doi:10.2307/41703508].

[5] Y. Bao et al., “Detecting accounting frauds in publicly traded US firms: New perspective and new method”. Available at: https://i/ssrn, Com/abstract 2670703, 2018.

[6] B. Li et al., “Detecting accounting frauds in publicly traded US firms: A machine learning approach” in

Asian Conference on Machine Learning, 2016, pp. 173-188. PMLR.

[7] A. Dyck et al., “Who blows the whistle on corporate fraud?,” J. Fin., vol. 65, no. 6, pp. 2213-2253, 2010 [doi:10.1111/j.1540-6261.2010.01614.x].

[8] Y. Bao et al., “Detecting accounting fraud in publicly traded US firms using a machine learning approach,” J. Acc. Res., vol. 58, no. 1, pp. 199-235, 2020 [doi:10.1111/1475-679X.12292].

[9] B. Dorris, Report to the Nations, 2018 Global Study on Occupational Fraud and Abuse. New York: Association of Certified Fraud Examiners, 2018.

[10] M. D. Beneish, “Detecting GAAP violation: Implications for assessing earnings management among firms with extreme financial performance,” J. Acc. Public Policy, vol. 16, no. 3, pp. 271-309, 1997 [doi:10.1016/S0278-4254(97)00023-9].

[11] T. Y. Wang et al., “Corporate fraud and business conditions: Evidence from IPOs,” J. Fin., vol. 65, no. 6, pp. 2255-2292, 2010 [doi:10.1111/j.1540-6261.2010.01615.x].

[12] D. W. Campbell and Ruidi Shang, “Tone at the bottom: Measuring corporate misconduct risk from the text of employee reviews,” Manag. Sci., vol. 68, no. 9, pp. 7034-7053, 2022 [doi:10.1287/mnsc.2021.4211].

[13] M. Schneider and R. Brühl, “Disentangling the black box around CEO and financial information-based accounting fraud detection: Machine learning-based evidence from publicly listed US firms,” J. Bus. Econ., pp. 1-9, 2023.

[14] J. O. Awoyemi et al., “Credit card fraud detection using machine learning techniques: A comparative analysis” in international conference on computing networking and informatics (ICCNI). IEEE, 2017, pp. 1-9 [doi:10.1109/ICCNI.2017.8123782].

[15] A. Reurink, “Financial fraud: A literature review,” Contemp. Top. Fin. Collect. Lit. Surv., pp. 79-115, 2019.

[16] X. Zhu et al., “Intelligent financial fraud detection practices in post-pandemic era,” Innovation (Camb), vol. 2, no. 4, 100176, 2021 [doi:10.1016/j.xinn.2021.100176].

[17] A. Sudjianto et al., “Statistical methods for fighting financial crimes,” Technometrics, vol. 52, no. 1, pp. 5-19, 2010 [doi:10.1198/TECH.2010.07032].

[18] [18] A. E. Omolara et al., “State-of-the-art in big data application techniques to financial crime: A survey,” Int. J. Comput. Sci. Netw. Sec., vol. 18, no. 7, pp. 6-16, 2018.

[19] A. E. Omolara et al., “State-of-the-art in big data application techniques to financial crime: A survey,” Int. J. Comput. Sci. Netw. Sec., vol. 18, no. 7, pp. 6-16, 2018.

[20] J. T. S. Quah and M. Sriganesh, “Real-time credit card fraud detection using computational intelligence,” Expert Syst. Appl., vol. 35, no. 4, pp. 1721-1732, 2008 [doi:10.1016/j.eswa.2007.08.093].

[21] W. Richert, Building Machine Learning Systems with Python. Packt Publishing Ltd, 2013.

[22] L. P. Coelho and W. Richert, Building Machine Learning Systems with Python. Packt Publishing Ltd, 2015.

[23] M. Welling, A First Encounter with Machine Learning. Irvine, CA: University of California, 2011, p. 12.

[24] M. Bowles, Machine Learning in Python: Essential Techniques for Predictive Analysis. John Wiley & Sons, 2015.

[25] J. Han, M. Kamber in J. Pei, Data Mining: Concepts and Techniques: Concepts and Techniques, vol. 3. izd.", 2011.

[26] I. H. Sarker et al., “Cybersecurity data science: An overview from machine learning perspective,” J. Big Data, vol. 7, pp. 1-29, 2020.

[27] Y. Yusof et al., “Utilizing unsupervised weightless neural network as autonomous states classifier in reinforcement learning algorithm” in, 2017 IEEE 13th International Colloquium on Signal Processing & Its Applications (CSPA). IEEE. IEEE, 2017, pp. 264-269 [doi:10.1109/CSPA.2017.8064963].

[28] Z. Zhao and Tongyuan Bai, “Financial fraud detection and prediction in listed companies using SMOTE and machine learning algorithms,” Entropy (Basel), vol. 24, no. 8, p. 1157, 2022 [doi:10.3390/e24081157].

[29] D. Chen, “Predicting accounting fraud in publicly traded Chinese firms via a PCA-RF method” in, Advances in Computer Science Research International Conference on Computer Science, Information Engineering and Digital Economy (CSIEDE 2022). Atlantis Press, pp. 739-748, 2022 [doi:10.2991/978-94-6463-108-1_82].

[30] P. Fukas et al., Augmenting Data with Generative Adversarial Networks to Improve Machine Learning-Based Fraud Detection, 2022.

[31] T. H. Pranto, Kazi Tamzid Akhter Md Hasib, Tahsinur Rahman, Akm Bahalul Haque, AKM Najmul Islam, and Rashedur M. Rahman. "Blockchain and Machine Learning for Fraud Detection: A Privacy-Preserving and Adaptive Incentive Based Approach." IEEE Access 10 (2022): 87115-87134.

[32] E. Ileberi et al., “A machine learning based credit card fraud detection using the GA algorithm for feature selection,” J. Big Data, vol. 9, no. 1, pp. 1-17, 2022.

[33] A. Hassanniakalager et al., “A machine learning approach to detect accounting frauds” Available at SSRN 4117764 (2022), SSRN Journal [doi:10.2139/ssrn.4117764].

[34] M. Sánchez-Aguayo et al., “Predictive fraud analysis applying the fraud triangle theory through data mining techniques,” Appl. Sci., vol. 12, no. 7, p. 3382, 2022 [doi:10.3390/app12073382].

[35] Saheed et al., “Big data analytics for credit card fraud detection using supervised machine learning models” in Big Data Analytics in the Insurance Market. Emerald Publishing Limited, 2022, pp. 31-56.

[36] Z. Liu et al., Detecting Financial Statement Fraud with Interpretable Machine Learning, 2021.

[37] S. Hamal and O. Senvar, “Comparing performances and effectiveness of machine learning classifiers in detecting financial accounting fraud for Turkish SMEs,” Int. J. Comput. Intell. Syst., vol. 14, no. 1, pp. 769-782, 2021 [doi:10.2991/ijcis.d.210203.007].

[38] N. Mqadi et al., “A SMOTe based oversampling data-

point approach to solving the credit card data imbalance problem in financial fraud detection," Int. J. Comput. Digit. Syst., vol. 10, no. 1, pp. 277-286, 2021 [doi:10.12785/ijcds/100128].

[39] N. K. Trivedi et al., "An efficient credit card fraud detection model based on machine learning methods," Int. J. Adv. Sci. Technol., vol. 29, no. 5, pp. 3414-3424, 2020.

[40] T. T. Nguyen et al., 'Deep learning methods for credit card fraud detection.' arXiv Preprint ArXiv:2012.03754, 2020.

[41] A. Thennakoon et al., "Real-time credit card fraud detection using machine learning" in Data Sci. Eng. (Confluence) 9th International Conference on Cloud Computing. IEEE, 2019, pp. 488-493 [doi:10.1109/CONFLUENCE.2019.8776942].

[42] P. Raghavan and N. E. El Gayar, "Fraud detection using machine learning and deep learning" in international conference on computational intelligence and knowledge economy (ICCIKE). IEEE, 2019, pp. 334-339 [doi:10.1109/ICCIKE47802.2019.9004231].

[43] Available at: https://www.tipdm.org:10010/#/competition/1354705811842195456/question.

[44] A. Argentiero et al., "The applications of artificial intelligence in cardiovascular magnetic resonance—A comprehensive review," J. Clin. Med., vol. 11, no. 10, p. 2866, 2022 [doi:10.3390/jcm11102866].

[45] A. Robles-Velasco et al., "Prediction of pipe failures in water supply networks using logistic regression and support vector classification," Reliab. Eng. Syst. Saf., vol. 196, p. 106754, 2020 [doi:10.1016/j.ress.2019.106754].

[46] A. Mehbodniya et al., "Financial fraud detection in healthcare using machine learning and deep learning techniques," Sec. Commun. Netw., vol. 2021, pp. 1-8, 2021 [doi:10.1155/2021/9293877].

[47] P. Hajek et al., "Fraud detection in mobile payment systems using an XGBoost-based framework," Inf. Syst. Front., vol. 25, no. 5, pp. 1985-2003, 2023.

[48] T. Chen and C. Guestrin, 'XGBoost: A scalable tree boosting system. arXiv 2016.' arXiv Preprint ArXiv:1603.02754, vol. 11, 2016.

[49] N. Dhieb et al., "Extreme gradient boosting machine learning algorithm for safe auto insurance operations" in IEEE international conference on vehicular electronics and safety (ICVES). IEEE, 2019, pp. 1-5 [doi:10.1109/ICVES.2019.8906396].

[50] H. Kaur et al., "A systematic review on imbalanced data challenges in machine learning: Applications and solutions," ACM Comput. Surv., vol. 52, no. 4, pp. 1-36, 2020 [doi:10.1145/3343440].2nd ed., vol. 3, J. Peters, Ed. New York, NY, USA: McGraw-Hill, 1964, pp. 15–64. 2020 [doi:10.1145/3343440].