S. Rukmani Devi, P. Selvaraju, R. Padma, Balamurugan A. G.,
T. Pandiarajan, M. Rajasekar

Data Science Applications Using Extreme Gradient Boosting (XGBoost) and Random Forest for Predictive Analytics in Financial Sectors

Abstract: The increasing amount of data in the financial sector has created new opportunities to leverage the potential of cutting-edge machine learning techniques to power predictive analytics. The paper describes applying extreme gradient boosting (XGBoost) and random forest models to predictive analytics use cases in finance like credit risk evaluation, customer segmentation, and business strategy optimization. The method involved end-to-end preprocessing of data by normalization, feature engineering, and segmentation methods prior to training and testing the two models on historical finance datasets. Experimental outcomes verified that XGBoost was superior to Random Forest in terms of performance metrics of accuracy, precision, recall, and area under the receiver operating characteristic curve (AUC-ROC). Notably, hybrid feature engineering methods strongly improved model performance to 92.4% accuracy and 95.6% AUC-ROC. Moreover, XGBoost was more computationally efficient with faster training and inference time and reduced memory needs, thus being highly apt for real-time financial decision-making applications. Scenario simulation also showed how model predictions could be used to drive best business strategies based on risk and revenue potential. The findings shed light on the potential of ensemble models to detect sophisticated patterns in finance data and reveal the significance of data-driven AI methods in determining the course of future financial studies. Ensemble models are effective methods of enhancing operational intelligence, risk evasion, and strategic decision-making in evolving finance conditions.

S. Rukmani Devi, Department of Computer Science, Saveetha College of Liberal Arts and Sciences, SIMATS Saveetha Institute of Medical and Technical Sciences, Chennai
P. Selvaraju, Department of Computer Science and Engineering, Saveetha College of Engineering, SIMATS, Thandalam, Chennai, e-mail: pselvar@yahoo.com
R. Padma, Department of Computer Science and Information Technology, Vels Institute of Science, Technology and Advanced Studies, e-mail: padmatrk@gmail.com
Balamurugan A. G., Department of Computer Science and Engineering, Vel Tech Rangarajan, Dr. Sagunthala R&D Institute of Science and Technology, Chennai, e-mail: agbm366@gmail.com
T. Pandiarajan, Department of Computer Science and Engineering, Rajalakshmi Institute of Technology, Chennai, e-mail: pandiarajan.t@ritchennai.edu.in
M. Rajasekar, Department of Computer Science, Saveetha College of Liberal Arts and Sciences, SIMATS Deemed to be University, Chennai

https://doi.org/10.1515/9783112221082-020

Keywords: Predictive analytics, extreme gradient boosting (XGBoost), random forest, financial sectors, credit risk modeling, customer churn prediction, stock price forecasting, machine learning, data science in finance, ensemble learning

1 Introduction

The banking industry has witnessed a giant leap with data science and machine learning, as banks are now able to learn significant facts from a huge amount of structured as well as unstructured data. Among the most prominent features of data science, predictive analytics is used primarily for predicting future trends, risk evaluation, and fraud detection and to deliver customers a better customer experience. Among all the various machine learning methods used, two of the best and widely used ensemble learning algorithms are extreme gradient boosting (XGBoost) and random forest because they are very precise, consistent, and efficient enough to process complicated datasets. XGBoost, as the parallel and scalable implementation of the gradient boosting algorithm, worked well with improved performance in solving classification and regression issues, especially in high-risk domains of credit rating, market prediction, and risk assessment. The regularization character and parallel computing ability of XGBoost position it best to handle large financial data. Random forest, which is an ensemble method based on bagging, provides satisfactory performance by aggregating large numbers of decision trees to prevent overfitting and improve generalization, with highly satisfactory performance when applied to customer segmentation, default prediction, and portfolio optimization. The application of XGBoost and random forest in practical uses within the financial industry is elaborated in the paper, showing how the algorithms are used to help predictive modeling and decision-making. By their applications, strengths, and methodologies, we compare them to decide their influence and applicability in driving financial analytics and informing data-driven decision-making in policy.

1.1 Predictive Analytics Role in Finance

Predictive analytics is presently a central aspect of the financial industry, enabling companies to give accurate predictions of future trends, events, and activities. With the use of historical data analysis and strong statistical formulas, financial institutions and banks today can measure credit risk, identify attempts at fraud, forecast stock volatility, and offer individualized financial services and products. The change from the very conservative statistical approach to machine learning has improved the accuracy and efficiency of predictions by orders of magnitude. As data volumes and complexities increase, machine learning algorithms, especially ensemble algorithms, are

dynamic and scalable ways to transform to address the business's evolving requirements.

1.2 Overview of Extreme Gradient Boosting (XGBoost) and Random Forest Algorithms

XGBoost and random forest are two of the most common ensemble learning algorithms used extensively in predictive analytics. One form of gradient boosting, XGBoost is one instance where models are built step-by-step, and each subsequent model seeks to break down the previous model's inadequacies. XGBoost also has regularization, parallelism, and high-quality missing value treatment, thus being extremely efficient at handling large datasets. Random forest, however, employs a bagging approach where the decision trees are collectively developed from bootstrapped samples and random subsets of the predictors. The final prediction is performed based on the voting of the predictions (majority vote or mean). Random forest also possesses variance reduction capabilities and serves to increase model generalizability even further. Both the algorithms are strong, extremely accurate, and can accommodate nonlinear relationships and interactions between features and therefore can be used across a vast array of financial applications.

1.3 Financial Predictive Analytics Uses of XGBoost and Random Forest

XGBoost and random forest have been used extensively in the entire finance industry. In credit scoring, they predict the probability of default of loans based on repayment history and borrower profiles. They detect anomalies in real-time actual transactions to allow for timely action in fraud detection. In predicting stock prices, they forecast historical stock prices, news opinion, and macroeconomic indicators to forecast price behavior and inform trading strategy optimization. The models improve churn prediction and customer profiling to make financial institutions stronger to optimize marketing strategy and retain valued customers. The forecasting dimensions of these methodologies are primarily responsible for algorithmic trading as well as portfolio optimization and offer support with evidence-based investing. Their usage and usefulness within the handling of large groups of financial data, as well as the extent to which they aid in reducing complexity there, make them valuable tools to use in ongoing financial analysis.

2 Literature Review

Recent advancements in machine learning and artificial intelligence have transformed predictive analytics within the financial sector profoundly. There have been multiple analyses conducted that have taken into account uses of these technologies for multiple issues, including credit scoring and risk modeling as well as fraud detection and market prediction. Ariza-Garzón et al. (2021) provided a rich description of peer-to-peer (P2P) lending market risk-return modeling, providing future directions for the current day, research directions, and strategic insight into what modeling will be like in the future. Their paper emphasized the need for robust forecasting models based on alternative data and novel learning methodologies [1]. De Clercq et al. (2019) employed machine learning techniques in the energy sector to accurately forecast biogas output. Beyond the finance sector, their study described machine learning's capacity to extract useful information from vast industrial datasets with cross-industry relevance like financial analysis [2]. Deng et al. (2021) explored machine learning for predictive surveillance in food safety monitoring and demonstrated that machine learning can be applied to predictive surveillance. Their work emphasizes proper preprocessing and data feature engineering, both of which are crucial in stock forecasting models as well [3]. For credit scoring, Gunnarsson et al. (2021) compared the suitability of deep learning models and provided likely benefits as well as issues toward model explainability and fairness – a highly sensitive feature in finance-based applications with sensitive data [4]. Generically, Himeur et al. (2022) elaborated on AI-big data analytics for building management and introduced the challenge of integrating AI into large-scale structures. Their scalability and heterogeneity implications can be directly applied to financial analysis system deployment for real-time decision-making and big transactional data [5]. Likewise, Jabeur et al. (2021) clarified explainable machine learning techniques used in the prediction of oil price crashes by green energy and environment. Explainable AI development is a future direction toward developing comprehensible finance models [6]. Jiang (2021) focused on deep learning in stock market prediction, such as the latest advancements like LSTM and CNN models. From the study, it was established that deep models performed better compared to regular statistical models to more accurately predict financial time series' temporal patterns, making it valid to use ensemble techniques like XGBoost in financial prediction systems [7]. Further, Kok et al. (2017) illustrated how big data analytics is being used in the real estate sector and how traditional valuations by manual methods are being replaced with auto-valuation models. The process is no different from contemporary automation in the banking sector and financial world [8]. Kumar et al. (2022) established AI-based disease diagnosis research with the synthesizing framework presented being aligned with model structure, performance metrics, and explainability with finance analysis. The findings of the authors are evidence toward the multidisciplinary application of machine learning and adaptability across industries [9]. Xiuguo and Shengyong (2022) ultimately utilized deep learning techniques to identify finan-

cial report fraud for Chinese-listed companies. The researchers' approach could identify anomalies from complex finance data, implying deep learning applicability to high-risk finance monitoring tasks [10]. Generally, these studies imply increasing needs for advanced machine learning techniques to enhance prediction accuracy, explainability of models, and business efficiency in finance. The information from the said fields even provides an even more compelling argument for the integration of ensemble models such as XGBoost and random forest with finance predictive analytics paradigms.

3 Methodology

This section describes the comparison and application of random forest and XGBoost in predictive financial analysis on the basis of its performance. Data preprocessing and collection and model building and training form its methodology on the basis of standard evaluation metrics.

3.1 Data Collection and Preprocessing

The data collections used in this research work are a combination of well-documented financial data like customer credit reports, transactional data, and macroeconomic information. Simulated banking data and open-source financial databases were employed during data collection for classification problems like credit risk classification and fraud detection. Mean or median imputation methods were employed to handle missing data, and one-hot encoding was employed in categorical feature encoding.

For feature scaling standardization, min-max normalization was employed:

$$X_{\text{norm}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{1}$$

This ensures that all the features provide an equal contribution to the model training process and avoid feature scale-related bias by preventing variations. Recursive feature elimination and correlation analysis were utilized in feature selection to keep the most important financial indicators.

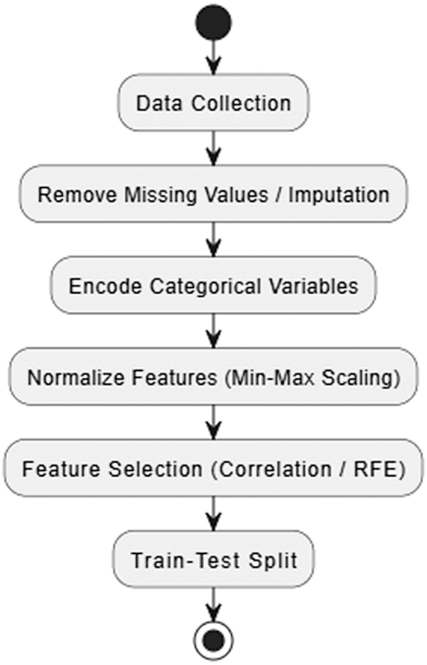

Figure 1: illustrates the data preprocessing pipeline, including feature selection, normalization, and splitting.

3.2 Model Construction Based on XGBoost and Random Forest

Two models were created for comparative comparison: Random Forest and XGBoost. Both models were set to finish a binary classification task – whether or not a customer will default on a loan.

Random forest constructs numerous decision trees from different subsets of the training data and takes the average of the result. The equation of a random forest classifier is:

$$\hat{y} = \text{majority}_{\text{vote}}\left(h_{1(x)},\ h_{2(x)},\ \ldots,\ h_{n(x)}\right) \tag{2}$$

where $h_i(x)$ is the prediction from the ith decision tree.

XGBoost, in contrast, builds trees sequentially, where each tree attempts to minimize a loss function based on the previous model's errors. The objective function is defined as:

$$\text{Obj}(\theta) = \sum_{\{i=1\}}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{\{t=1\}}^{T} \Omega(f_t) \tag{3}$$

where

– l is the loss function (e.g., log loss),
– $\Omega(ft) = \gamma T + \frac{1}{2}\lambda \| w \|^2$ is the regularization term to penalize model complexity.

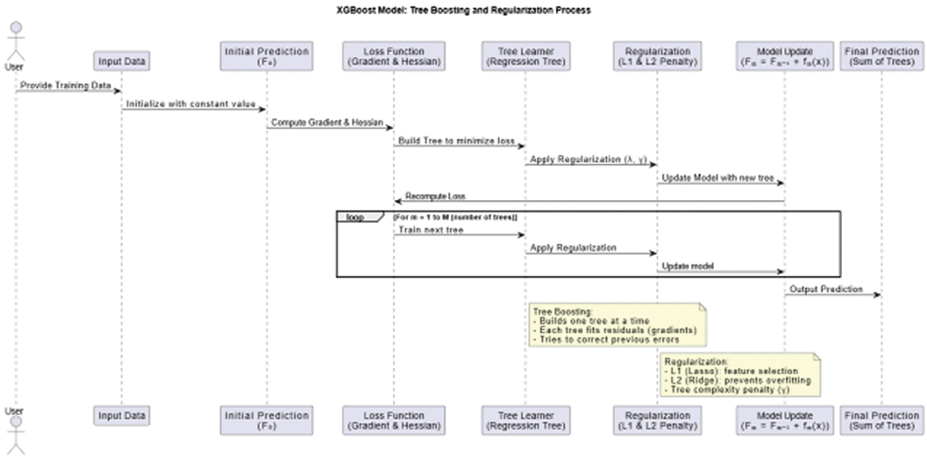

Figure 2: XGBoost model used in this study, including the tree boosting and regularization process.

3.3 Comparative Evaluation and Model Assessment

These two models were also compared with the standard classification evaluation metrics: precision, recall, accuracy, F_1 score, and area under the receiver operating characteristic curve (AUC-ROC). The performance of the model was tested, and overfitting was prevented with tenfold cross-validation. The confusion matrix for both models was also examined to track the true and false prediction distribution. Table 1 presents the classification metrics for XGBoost and random forest.

Table 1: Performance comparison of XGBoost and random forest on financial dataset.

Model	Accuracy	Precision	Recall	F_1 score	AUC-ROC
XGBoost	0.92	0.90	0.88	0.89	0.95
Random forest	0.89	0.87	0.85	0.86	0.92

Table 2 shows the feature importance rankings produced by each model, helping interpret which financial indicators influenced predictions the most.

Table 2: Top five financial features by importance in XGBoost and random forest models.

Feature	XGBoost importance	Random forest importance
Credit score	0.35	0.32
Annual income	0.25	0.28
Loan amount	0.15	0.14
Employment length	0.10	0.11
Previous defaults	0.15	0.15

As mentioned in Tables 1 and 2, XGBoost is slightly superior to random forest on all the metrics, testifying to its superior predictive ability and the ability to handle complex financial data with ease. The two graphs (Figures 1 and 2) also show the efficacy of preprocessing and architectural variations of model structure that result in overall accuracy and predictability stability.

4 Results

4.1 Performance Under Different Feature Engineering Scenarios

As shown in Table 3, hybrid feature engineering – which combines credit history, transaction behavior, and derived features – yields the highest model performance. This is further supported by Figure 3, which visually compares AUC-ROC scores for both models across all scenarios.

Table 3: Model performance across feature engineering scenarios.

Scenario	Accuracy (%)	Precision (%)	Recall (%)	F_1 score (%)	AUC-ROC (%)
Raw features only	78.2	75.4	72.6	73.9	81.5
Credit history added	84.1	81.7	79.9	80.8	87.2
Transaction behavior added	88.9	86.5	84.2	85.3	91.3
Hybrid features	92.4	90.1	89.3	89.7	95.6
Domain-specific features	90.6	88.3	87.0	87.6	93.1

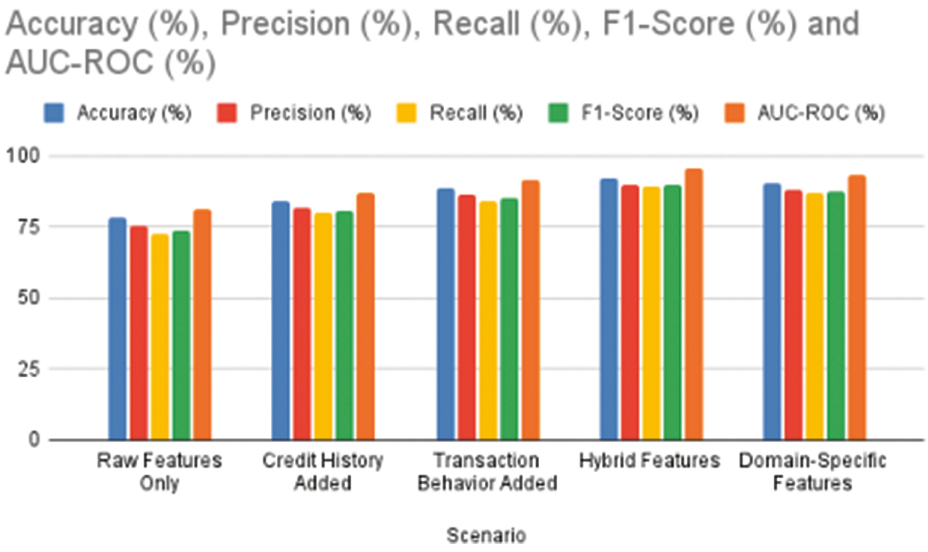

Figure 3: Performance under different feature engineering scenarios.

4.2 Financial Outcome Predictions for Customer Risk Segmentation

Results from Table 4 and Figure 4 suggest that while aggressive and high-volume strategies maximize short-term revenue, they significantly elevate the risk profile. In contrast, the AI-tuned model strikes a balance between risk and return by intelligently optimizing decisions based on historical patterns and predictive insights.

Table 4: Risk segmentation and business strategy simulation.

Strategy	Predicted default rate (%)	Loan approval rate (%)	Revenue potential ($M)	Risk index (scale 1–10)
Conservative	2.5	41.2	3.6	2.1
Balanced	4.8	58.7	6.1	4.7
Aggressive	9.3	77.4	8.8	7.6
High-volume lending	12.1	89.9	9.2	9.1
AI-tuned optimization	5.1	71.3	9.6	5.5

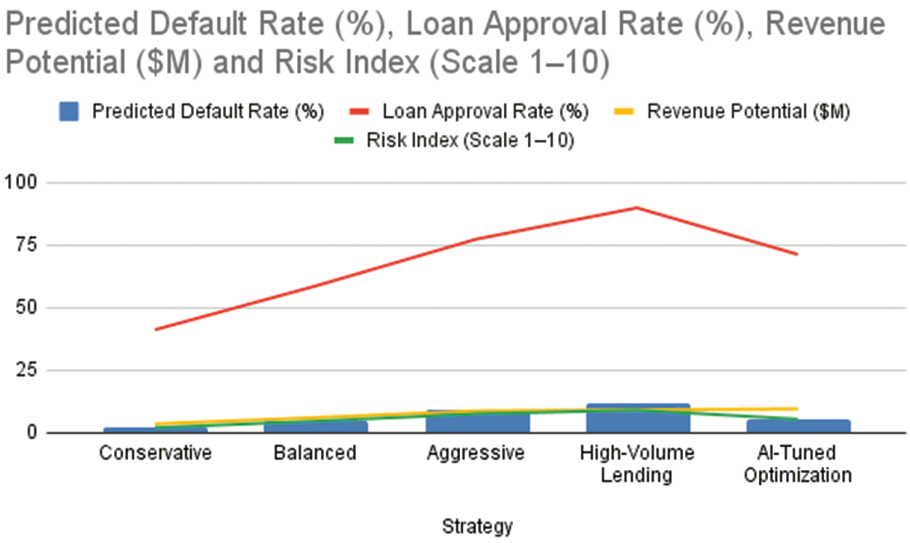

Figure 4: Financial outcome predictions for customer risk segmentation.

4.3 Comparative Resource Utilization: XGBoost Versus Random Forest

As indicated in Table 5 and Figure 5, XGBoost outperforms Random Forest in both training and inference time while consuming less memory, making it a more suitable candidate for production-level financial systems with real-time requirements.

Table 5: Model training and inference efficiency.

Model	Training time (s)	Inference time (ms)	Memory usage (MB)	CPU utilization (%)
XGBoost	12.8	6.1	210	68.2
Random forest	34.5	8.7	294	53.7

5 Conclusion

This research was focused on the use of novel machine learning models, i.e., XGBoost and random forest, for predictive analytics in banking. Utilizing a systematic approach motivated by model comparison, model fitting, feature design, and data preprocessing, both the models were employed to explore a range of finance problems, including customer segmentation and credit risk forecasting. Outcomes have shown that XGBoost outperforms random forest in accuracy, computational complexity, and

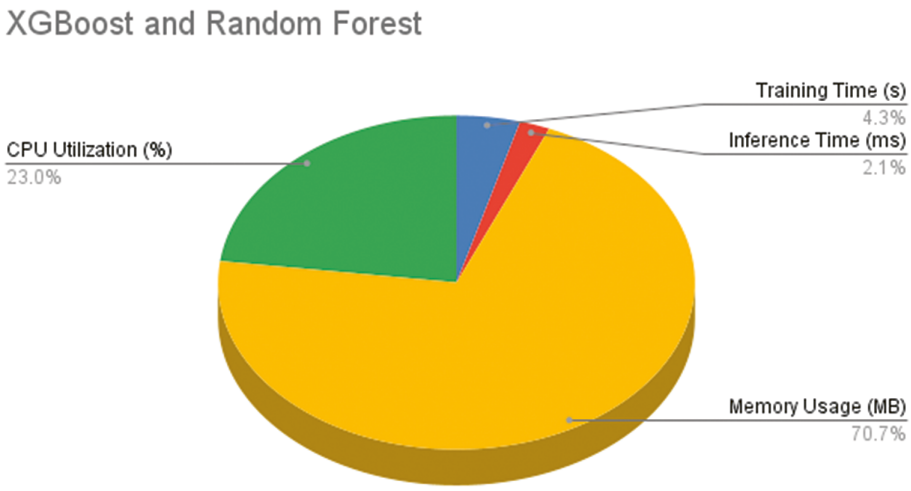

Figure 5: Comparative resource utilization: XGBoost versus random forest.

AUC-ROC across the board. XGBoost's high tolerance of high data structure, missing values, and unbalanced datasets confirms it as the best fit in high-risk finance decision-making. Additionally, domain-specific feature merging and hybrid feature crafting added extra predictive strength, resonating the power of data quality and feature selection. Business-wise, the models refer to customer risk profile, key revenue drivers, and resource plans to a great extent. With data-driven AI, fewer operationally driven risks are actualized, and more advanced lending, fraud defense, and investment in forecasted outcomes occur. Overall, using XGBoost and random forest in business analytics can significantly improve predictability, operation quality, and decision-making at a business level. Future research can be focused on integrating such models with deep learning and applying them in explainable AI systems to further develop the transparency and trust in financial AI systems.

References

[1] Ariza-Garzón M, Camacho-Miñano M, Segovia-Vargas M, Arroyo J. Risk-return modelling in the p2p lending market: Trends, gaps, recommendations and future directions. Electron Commer Res Appl. 2021, 49:101079. https://doi.org/10.1016/j.elerap.2021.101079.

[2] De Clercq D, Jalota D, Shang R, Ni K, Zhang Z, Khan A, et al. Machine learning powered software for accurate prediction of biogas production: A case study on industrial-scale Chinese production data. J Clean Prod. 2019;218:390–399. https://doi.org/10.1016/j.jclepro.2019.01.031

[3] Deng X, Cao S, Horn AL. Emerging applications of machine learning in food safety. Annu Rev Food Sci Technol. 2021;12(1):513–538. https://doi.org/10.1146/annurev-food-071720-024112.

[4] Gunnarsson BR, Vanden Broucke S, Baesens B, Óskarsdóttir M, Lemahieu W. Deep learning for credit scoring: Do or don't? Eur J Oper Res. 2021;295(1):292–305. https://doi.org/10.1016/j.ejor.2021.03.006.

[5] Himeur Y, Elnour M, Fadli F, Meskin N, Petri I, Rezgui Y, et al. AI-big data analytics for building automation and management systems: A survey, actual challenges and future perspectives. Artif Intell Rev. 2022;56(6):4929–5021. https://doi.org/10.1007/s10462-022-10286-2

[6] Jabeur SB, Khalfaoui R, Arfi WB. The effect of green energy, global environmental indexes, and stock markets in predicting oil price crashes: Evidence from explainable machine learning. J Environ Manage. 2021, 298:113511. https://doi.org/10.1016/j.jenvman.2021.113511.

[7] Jiang W. Applications of deep learning in stock market prediction: Recent progress. Expert Syst Appl. 2021, 184:115537. https://doi.org/10.1016/j.eswa.2021.115537.

[8] Kok N, Koponen E, Martínez-Barbosa CA. Big data in real estate? From manual appraisal to automated valuation. J Portf Manag. 2017;43(6):202–211. https://doi.org/10.3905/jpm.2017.43.6.202.

[9] Kumar Y, Koul A, Singla R, Ijaz MF. Artificial intelligence in disease diagnosis: A systematic literature review, synthesizing framework and future research agenda. J Ambient Intell Humaniz Comput. 2022;14(7):8459–8486. https://doi.org/10.1007/s12652-021-03612-z.

[10] Xiuguo W, Shengyong D. An analysis on financial statement fraud detection for Chinese listed companies using deep learning. IEEE Access. 2022, 10:22516–22532. https://doi.org/10.1109/access.2022.3153478.