SCIENTIFIC OASIS

Decision Making: Applications in
Management and Engineering

Journal homepage: www.dmame-journal.org
ISSN: 2560-6018, eISSN: 2620-0104

DECISION MAKING:
APPLICATIONS IN
MANAGEMENT AND
ENGINEERING

Discussion on the Enterprise Financial Risk Management Framework Based on AI Fintech

Yu Liu[1*]

[1] School of Intelligent Finance, Henan Institute of Economics and Trade, Zhengzhou, 450000, China

ARTICLE INFO

Article history:
Received 30 September 2023
Received in revised form 15 November 2023
Accepted 19 November 2023
Available online 5 December 2023

Keywords: Decision tree algorithm; Random forest algorithm; Finance; Risk prediction.

ABSTRACT

Deep learning algorithms lack interpretability and interpretability in the decision-making process. This makes it difficult to understand the judgment basis and decision-making process of financial risks based on algorithms, which may reduce the trust and acceptance of risk decisions by enterprises. To address this issue, this study introduces the improved random forest algorithm based on the decision tree algorithm to discuss its framework. Through analysis of the PR curve in the experiment, it was determined that the AP value of the enhanced random forest algorithm is 0.9919, a significant improvement over the RF algorithm's previous value of 0.9237. It also has a good balance between the precision rate and the recall rate. By introducing the data set to compare and analyze the three algorithms of SVM, CRAT, and the improved random forest algorithm, it is found that the improved random forest algorithm has a higher test value. Through cluster analysis, it is found that the clustering accuracy of the improved random forest algorithm is about 81%. The final analysis of a company's financial data sample showed an accuracy rate of 69.6% on the enterprise access index. The improved random forest algorithm achieves a good balance between accuracy and recall, resulting in the potential for high accuracy in the domain of risk assessment. In addition, compared with other algorithms such as SVM and CRAT, the improved random forest algorithm has higher test values, indicating its excellent performance in financial risk prediction. To sum up, this study firstly verified the feasibility and accuracy of the improved random forest algorithm for financial risk prediction. Additionally, it validated the method's predictive capability using factual enterprise data samples and ultimately established a foundation for the enterprise's financial risk management framework. Construction offers a new theoretical direction.

1. Introduction

In recent years, with the intensification of economic globalization, the world economy has developed rapidly. However, the financial industry has revealed increasingly serious problems and risks, significantly impacting the original industrial model and numerous financial institutions [1]. Most financial technology companies are facing the loss of uncertainty, and the diversity of financial technology is increasing, which also reflects the importance of risk management in financial

* *Corresponding author.*
E-mail addresses: liuyu.lyly@hotmail.com

https://doi.org/10.31181/dmame712024942

technology. In recent decades, there have been several major financial crises in the global financial market. The most recent one was the global financial crisis that broke out in 2008 due to the subprime mortgage crisis in the United States. During this financial crisis, many financial technology companies around the world suffered immeasurable losses, and even many large and medium-sized enterprises were facing bankruptcy [2]. With the development of Internet technology and the emergence of artificial intelligence (AI), it has now become an information age [3]. Due to its efficiency and safety, artificial intelligence is increasingly being utilized in financial risk management and prevention. Currently, advancements in science and technology have significantly enhanced the economic benefits of financial technology firms while also effectively mitigating financial risk for enterprises. However, AI technology has been around for a short time, and financial companies have accumulated less data, which limits the choice of algorithms [4]. And the accuracy and efficiency of the existing algorithms can no longer keep up with the growing speed of the data, so new more efficient, and accurate algorithms must be introduced. Previous studies have found that the random forest algorithm has a good effect on risk prevention, but there are some problems such as high similarity [5]. Although artificial intelligence has great potential, its existence is relatively short-lived, and financial companies have accumulated limited data. The accuracy and efficiency of existing algorithms can no longer keep up with the speed of data growth, so it is necessary to introduce new, more efficient and precise algorithms to address this challenge. This research will introduce the random forest algorithm and add the decision tree algorithm to improve it. Therefore, the algorithm's performance is enhanced, further reducing the financial risks for enterprises.

This study is divided into five parts. The first part explains the necessity and innovation of conducting this study; The second part provides a review of the current research progress on the research topic; The third part studied the random forest algorithm and its improvements in artificial intelligence technology; The fourth part conducted performance testing experiments and result analysis on the improved results; The fifth part summarizes the conclusions of this study.

2. Related works

Through domestic and foreign research, it is found that decision tree algorithm and random forest algorithm have been widely used in all walks of life. When Wang L studied the fit degree in clothing production, he proposed to use the decision tree algorithm to classify body shape and develop a decision tree body shape recognition model to better judge body shape. The results show that the effectiveness of the decision tree algorithm in classifying female body types. And the classification accuracy of various body types is above 95% [6]. When Ariyati et al. studied electronic payment in financial technology, to improve the security and accuracy of electronic payment, a decision tree algorithm was introduced to analyze and process it. The results show that the algorithm provides good classification and increases security and accuracy in digital payments [7]. When Wang et al. studied the conversion of credit funds, to consider risk management, a decision tree algorithm was introduced to evaluate it [6]. The experimental results show that the risk estimation accuracy of the decision tree algorithm reaches 81.2% and 83.6%, respectively, which proves that the model can be used as an effective reference for pre-lending risk assessment. To explain the phenomenon of IPO underpricing, Keuangan proposed a decision tree algorithm and established a model according to the underpricing. The results from the experiments demonstrate that the decision tree algorithm can provide explanations within a specific classification range [8]. Additionally, the decision tree algorithm model can effectively replace the linear regression econometric model.

Saadoon & Abdulamir [9] analyzes the data for their research on big data applications by incorporating a random forest algorithm. The experimental results show that the random forest algorithm has achieved good results in the application of big data. And the accuracy rate is as high as 99.97%, which fully shows the performance of the random forest algorithm in big data processing. Ning et al. proposed a new framework of random forest algorithm combined with feature extraction for leak detection and classification of gas pipelines under severe background noise. The results show that the framework is extremely effective, and the accuracy is also greatly improved compared to the previous one [10]. Ahn proposed a random forest algorithm to establish a data model when studying the diagnosis and classification of Alzheimer’s disease and the identification of biomarkers [11]. The experimental results show that the random forest algorithm exhibits high performance of 92% accuracy, and the standard deviation has no significant deviation, which also shows the effectiveness of the algorithm. Zhang & Cai proposed a random forest algorithm and a digital image analysis framework to analyze the variation trend of permeability with pore structure parameters when studying the relationship between carbonate pore types and complex permeability [12]. The experimental results show that the different discrete rock types are mainly controlled by the product of the shape factor and the square of the tortuosity, and its variation trend curve is accurately drawn. The evidence suggests that the random forest algorithm is highly effective and accurate in predicting carbonate permeability. When Liu conducted research on financial market regulators, they found that there were a lot of financial risks, so they proposed a random forest algorithm to analyze and predict them. The experimental results show that the prediction accuracy of the algorithm is as high as 97%, which shows the accuracy and efficiency of this algorithm [13]. Through previous research, it has been found that the random forest algorithm has been used in various fields and has shown extremely strong performance. It is also widely used in financial risk, which greatly improves the security of financial technology and reduces the possible crisis of financial risk.

In summary, previous studies have shown that decision tree algorithms and random forest algorithms have broad application prospects in various fields, providing effective classification, prediction, and interpretation capabilities to help various industries better cope with complex problems and challenges. However, random forest and decision tree algorithms commonly necessitate a significant quantity of training data and multiple rounds of experimentation and modification to identify the most favorable parameter combination, which consumes substantial time and computational resources. Therefore, improving the application of these two algorithms in the financial risk management framework of fintech enterprises has certain value.

3. Random forest algorithm in AI technology and its improvement

3.1 The characteristics and basic structure of decision tree algorithm

Among data mining algorithms, decision tree algorithms can express more complex nonlinear patterns and feature relationships than other algorithms. And this algorithm is often carried out by automatically generating judgment rules when modeling [14]. The core idea of this algorithm is to solve the classification problem by dividing and conquering and is trained through feature evaluation at each stage, from the top to the bottom. Different data will be allocated to different tree structures to determine whether the recursive termination condition is met, if so, the loop will be terminated, otherwise the loop will continue. Each classification rule in the decision tree model corresponds to a different tree branch [14]. In the financial field, different financial indicators may have different impacts on risk prediction. Through the analysis of historical data and features, decision trees are capable of constructing a tree-like structure in an automatic fashion that classifies and evaluates differing risk factors. This can provide more accurate risk prediction and assessment for fintech enterprises, thereby formulating corresponding risk management strategies.

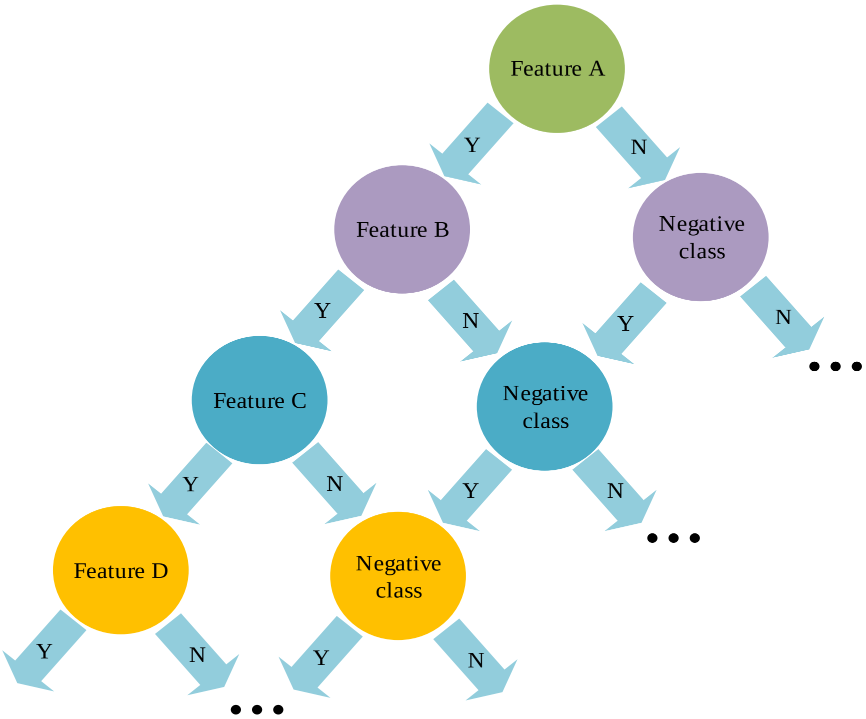


Fig. 1. Schematic diagram of decision tree algorithm

As shown in Figure 1, the decision tree model is also regarded as a classification rule group, and the model can perform classification processing and prediction on huge data sets through internal rules. Another advantage of the decision tree algorithm is that the model trained by this algorithm is extremely simple, and a lot of work is relatively reduced when interpreting. When processing data, it is possible to consider both categorical and numerical characteristics of the data without the need for data standardization. The division of nodes is very important in the training of the decision tree algorithm. The key to node division is how to choose a suitable division attribute so that the divided nodes belong to the same class. Commonly used node division methods include the information gain, gain rate, and Gini index [15].

Information gain is the earliest basis used in the decision tree algorithm for node division. Information gain refers to the degree to which the uncertainty of the information generated by the class is continuously reduced under the condition that Y the eigenvalues are determined. X Given a set, the D proportion of samples of the p_k first k class in the set is D. Therefore D, the definition of information entropy in this set is shown in equation (1) [16].

$$H(D) = -\sum_{k=1}^{K} p_k \log_2(p_k) \tag{1}$$

the condition that the A discrete eigenvalues are determined, if there A is n a possible value in the discrete eigenvalues, the sample set D can be divided into n different branch nodes according to this possible value. After calculating the information entropy and weight of each branch node respectively $\left|D^i\right|\big/D$, the empirical conditional entropy can be calculated, as shown in equation (2) [17].

$$H\left(D\middle|A\right)=\sum_{i=1}^{n}\frac{\left|Di\right|}{D}H\left(Di\right)=-\sum_{i=1}^{n}\frac{\left|Di\right|}{D}\sum_{k=1}^{K}p_{ik}\log_2\left(p_{ik}\right) \tag{2}$$

Then, according to the discrete eigenvalues A, the division information gain of the already divided data set is calculated, as shown in equation (3).

$$Info\_Gain\left(D,A\right)=H\left(D\right)-H\left(D\middle|A\right) \tag{3}$$

It can be known from equation (3) that the optimal criterion for dividing features is shown in equation (4).

$$A_* = \underset{A=\{A_1,A_2,...,A_m\}}{\arg\max}\ Info\_Gain\left(D,A\right) \tag{4}$$

When deciding on the information gain rate for division, the selection of features with more values can be prevented, unlike with the information gain method. The information gain rate is divided as shown in equation (5) [18].

$$Info\_Gain\_Ratio\left(D\middle|A\right)=\frac{Info\_Gain\left(D,A\right)}{H_A\left(D\right)} \tag{5}$$

In Equation (5), it represents $H_A\left(D\right)$ the entropy of the sample set D with respect to discrete eigenvalues, as shown in Equation (6).

$$H_A\left(D\right)=-\sum_{i=1}^{n}\frac{\left|D_i\right|}{\left|D\right|}\log_2\frac{\left|Di\right|}{\left|D\right|} \tag{6}$$

As shown in Equation (6), the optimal method for dividing features is to select the feature with the highest gain rate. In the classification and regression tree (classification and regression tree, CART) algorithm for feature division, the Gini index method is selected. The Gini index of the sample set D is shown in equation (7).

$$Gini\left(D\right)\sum_{k=1}^{K}p_k\left(1-p_k\right) \tag{7}$$

In equation (7), it represents p_k the proportion of the first class samples in the k sample set. D If the discrete eigenvalues A have n different values, then when the discrete eigenvalues are A determined, the Gini index of the sample set D is shown in equation (8) [19].

$$Gini\_index\left(D,A\right)=\sum_{i=1}^{n}\frac{\left|D_i\right|}{D}Gini\left(D_i\right) \tag{8}$$

As shown in Equation (8), the characteristic standard divided according to the Gini index is shown in Equation (9) [20].

$$A^* = \underset{x=X}{\arg\min}\,Gini\_index\left(D,A\right) \tag{9}$$

3.2 Random forest algorithm based on decision tree

Random forest algorithm (RF) was first proposed in the early 21st century. This algorithm is a statistical learning theory based on ensemble theory. This algorithm uses the self-help sampling technique to extract different data sets from the original data set for training, and then uses the random subspace method to model the decision tree of the selected data set to form a random forest [20]. The random forest algorithm exhibits strong scalability when processing large quantities of data. In artificial intelligence technology, as the amount of data increases and the complexity of problems increases, algorithms that can efficiently process large-scale data are needed. The parallel computing and low computational complexity of the random forest algorithm make it a choice suitable for large-scale data. Using the random forest algorithm requires preparing data, performing feature selection, parameter tuning, model training and evaluation, and interpreting and monitoring the results of the model. Adhering to these requirements guarantees the successful implementation of random forest algorithms in specific tasks. The random forest algorithm under artificial intelligence technology focuses more on feature selection and engineering. When processing large-scale and high-dimensional datasets, it is necessary to select the most predictive features to improve model performance. Applications lacking AI technology may lack feature selection and engineering steps, or simply employ basic features. The training and prediction process of the random forest algorithm is shown in Figure 2.

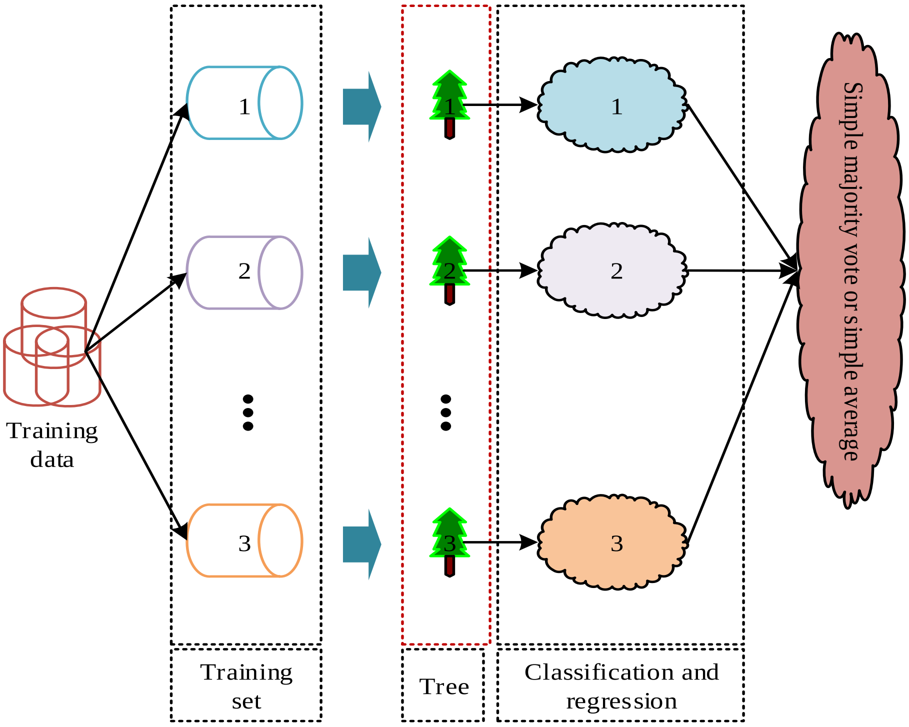

Training data
1
2
3
Training set
Tree
Classification and regression
Simple majority vote or simple average

Fig. 2. Random forest training and prediction process

As shown in Figure 2, before constructing a decision tree, it is first necessary to extract k a training data set, each of which contains N a sample. Then, the classification and regression tree (CART) is used to construct the base classifier model, and m a different feature is selected from the input features at each node as the split feature set of the decision tree. Then it selects the optimal splitting features and splitting points from these sets, and the Gini index and the minimum criterion classification method must be followed in the selection process. It is judged whether the termination condition is met, if yes, it will jump out of the loop, otherwise continue the loop. A random forest, comprising multiple decision trees, demonstrates superior generalization ability and accuracy than a single decision tree algorithm. Diversity measurement methods are usually used to measure the degree of diversity of individual classifiers. Common diversity measurement methods include paired diversity measurement and unpaired diversity measurement [18].

Pairwise diversity measures are more used to measure the similarity between two classifiers. Among them, the common measures are Q statistic, correlation coefficient, discordance measure, and Kappa statistic. The Q statistic belongs to the definition in statistics, and h_i, h_j the Q statistic of the two classifiers is shown in the equation (10).

$$Q_{ij} = \frac{N^{11}N^{00} - N^{10}N^{01}}{N^{11}N^{00} + N^{10}N^{01}} \tag{10}$$

In equation (10), N^{11} represents h_i, h_j the number of samples predicted as positive by the two classifiers, the number of samples predicted as positive N^{10} by h_i the classifier, the number of samples predicted as positive N^{01} by h_j the classifier, and the number of samples predicted as positive N^{00} by the h_i, h_j two classifiers at the same time. It is a negative number of samples. And the statistic takes value in the range [-1,1], if and only if the two classifiers h_i, h_j predict the same result, the statistic value is 1. The statistic value is -1 when the prediction results of the two classifiers are h_i, h_j completely inconsistent, and the statistic value is 0 when the two classifiers h_i, h_j are independent of each other.

Like the Q statistic, the correlation coefficient also belongs to the category of statistics, and h_i, h_j the correlation coefficient of the two classifiers is shown in equation (11).

$$\rho ij = \frac{N^{11}N^{00} - N^{10}N^{01}}{\sqrt{\left(N^{11} + N^{10}\right)\left(N^{10} + N^{00}\right)\left(N^{00} + N^{01}\right)\left(N^{01} + N^{11}\right)}} \tag{11}$$

When the correlation coefficient Q statistic is the same as the correlation coefficient, and both are used as a measure of diversity, then the modulus of the Q statistic is less than or equal to the modulus of the correlation coefficient. The disparity metric, however, differs from these two measures by placing greater emphasis on samples with varying predictive outcomes of the classifier, as presented in Equation (12).

$$dis_{ij} = \frac{N^{10} + N^{01}}{N} \tag{12}$$

equation (12), dis_{ij} the value is in the range of [0, 1]. The larger the value, the richer the diversity among the classifiers. Kappa statistic is more comprehensive, it can be used to measure the accuracy of the classifier; It can also indicate the consistency of the prediction results of different models, as shown in equation (13).

$$K = \frac{p_r(a) - p_r(e)}{1 - p_r(e)} \tag{13}$$

In equation (13), $p_r(a)$, $p_r(e)$ respectively represent the actual and theoretical consistency of the two classifiers, as shown in equation (14) and equation (15).

$$p_r(a) = \frac{N^{11} + N^{00}}{N} \tag{14}$$

$$p_r(E) = \frac{\left(N^{11} + N^{10}\right)\left(N^{11} + N^{01}\right) + \left(N^{01} + N^{00}\right)\left(N^{10} + N^{00}\right)}{N^2} \tag{15}$$

The K value is in the range of [-1,1], the higher the K value, the stronger the consistency. When the two classifiers h_i, h_j predict the same results, $K = 1$; when the agreement is the same as expected by chance, $K = 0$; when the agreement is smaller than expected by chance, $K < 0$. Usually the K value is greater than zero.

3.3 Improvement of Random Forest Algorithm

Since there is a need to enhance the classification effectiveness and diversification of the decision tree in the random forest algorithm, this research will pursue further improvement based on the decision tree. Currently, enhancing the random forest algorithm involves primarily extracting precise decision trees from the original cluster and utilizing a clustering algorithm to divide these trees into diverse clusters for extracting differentiated decisions. Due to practical considerations, this study will be improved using the classification accuracy metric and the diversity metric.

First, use the data in the validation set to calculate the AUC value of each tree in the original forest, and use the calculated AUC value as the classification accuracy of the corresponding decision tree. Sort them from high to low according to the classification accuracy, and select the top 60% of the data to form a sub-forest. However, discrepancies in noise and features present in various datasets lead to low accuracy of most of the data, while some contain more precise data [21]. Therefore, the number of decision trees selected in this study is not fixed, in order to find a sub-forest with a higher average classification accuracy than a single decision tree in the original forest $SubF$, as shown in Equation (16).

$$SubF = \{t_i : Auc_i \geq A\} \left| A = \frac{1}{|K|} \sum_{j=1}^{|K|} Auc_j \right. \tag{16}$$

As shown in equation (16), if $SubF$ the number of trees in the tree exceeds 60% of the trees in the original forest, it will be selected $SubF$ as the sub-forest to be clustered, and if not, the selection criteria will be lowered. As shown in equation (17).

$$SubF = \{t_i : Auc_i \geq A - \delta\} \tag{17}$$

As shown in Equation (17), the standard deviation of the classification accuracy of all trees is first calculated in the original forest, and then select $A-\delta$ a decision tree with an accuracy greater than or equal to, and finally obtain a new sub-forest. To sum up, when building a new sub-forest, the training subsets are selected from the training set by sampling; then the decision tree generation algorithm training and classification model are used to process these training subsets; Then it performs classification and calculates the corresponding classification accuracy; finally, according to all the classification accuracy, average and standard deviation for comparative analysis, a suitable decision tree is selected to form a high-precision sub forest $SubF$.

While obtaining high-precision sub-forests, the overall diversity and richness of the decision tree model will be greatly reduced. Therefore, a clustering algorithm is selected $SubF$ to cluster high-precision sub-forests. In this step, the classification results of the validation set obtained by the high-precision sub-forest obtained above are first regarded as a new data set, and another data set is selected from these data sets K as the first cluster center; All data in the set are calculated. The distance from the data to the cluster center, re-form each data in the data set into a cluster according to the center with the closest distance; It summarizes the obtained several cluster centers, and calculate each cluster center; It judges whether the cluster center is Change or whether the number of iterations reaches the maximum value, if not, loop the above operation, if so, go to the next step; then calculate K the clustering silhouette coefficient of different data until the optimal K value is finally determined, and the corresponding clustering is found. Result: Finally, the optimal decision tree is selected from the divided clusters to form the final random sub-forest. The improved random forest algorithm is shown in Figure 3.

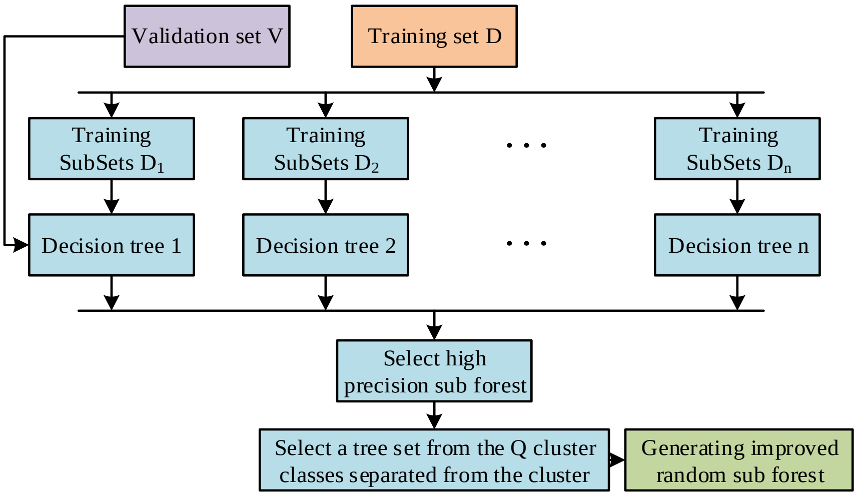


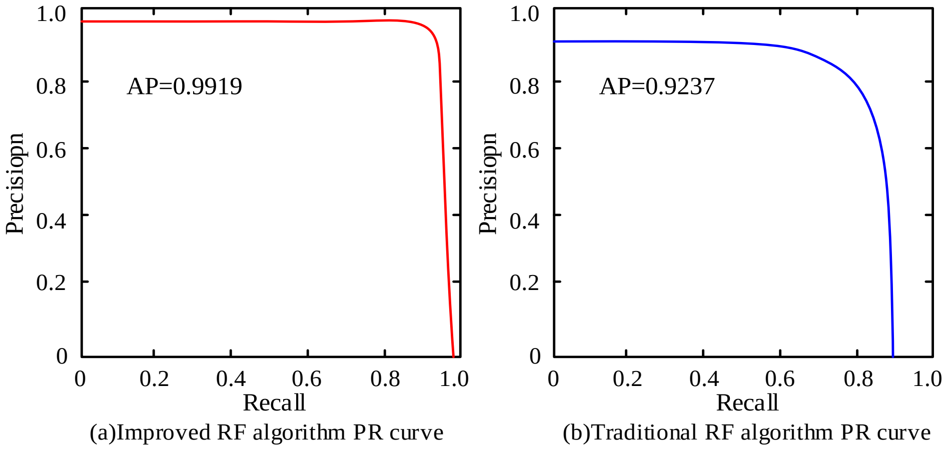
Fig. 3. Flowchart of the improved random forest algorithm

As shown in Figure 3, when the random forest algorithm is improved, it uses the classification accuracy of the decision tree and the similarity between the numbers to reduce the decision tree in the forest, thereby removing the unqualified decision tree model. The retained decision trees form a new sub-forest. These new sub-forests can reduce the model storage and prediction time. The algorithm has two stages: decision tree generation with precision and clustering.

4. Experimental results and analysis

This study selected accuracy, detection speed, Precision Recall Curve, and F1 score as evaluation indicators to comprehensively evaluate the performance of the algorithm. Among them, Precision Recall Curve is an indicator used to evaluate the performance of classification algorithms. Average Precision is an accurate measure of the average precision of prediction outcomes, and higher values indicate increasingly accurate performance. F1 score is a weighted harmonic average of accuracy and recall, used to comprehensively evaluate the performance of classification models.

In the experiment, the threshold for dividing the positive and negative cases of the predicted value must be set in advance, and the threshold is usually set to 0.5. Samples displaying a predicted value that is greater than or equal to 0.5 are classified as positive examples, while those displaying a predicted value that is less than 0.5 are classified as negative examples. Through the constant changes of the thresholds we set, the Precision and Recall values also change continuously. According to different test results, PR curves are formed. The PR curves of the test results of the improved DE algorithm and the traditional DE algorithm are shown in Figure 4.

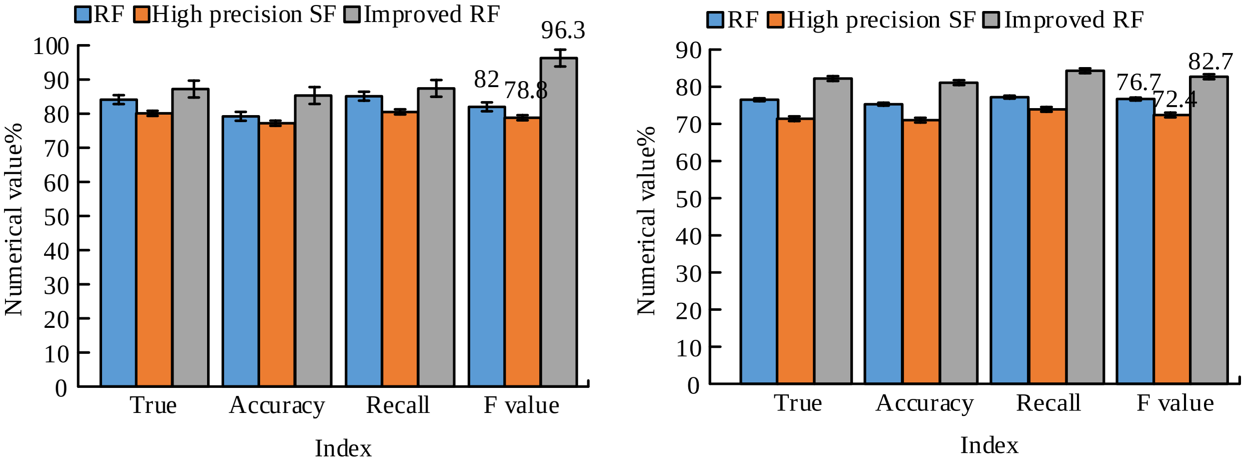


(a)Improved RF algorithm PR curve

(b)Traditional RF algorithm PR curve

Fig. 4. PR curves of the two algorithms

As shown in Figure 5, Figure 5(a) is the PR curve of the improved random forest algorithm, and Figure 5(b) is the PR curve of the traditional random forest algorithm. The results demonstrate that the improved recommendation algorithm exhibits superior precision and detection speed compared to its previous version. The AP value of the improved random forest algorithm is 0.9919, which is significantly higher than that of the RF algorithm before the improvement of 0.9237. There is a superior equilibrium and greater proficiency comparatively. In order to compare the performance of the random forest algorithm, the high-precision sub-forest and the improved random forest algorithm, four datasets, Greman Credit, Bank Marketing, Default of Credit Clients and Income, were introduced in this study to conduct the analysis of the three algorithms. test. The test results are grouped in pairs, as shown in Figure 5 and Figure 6.

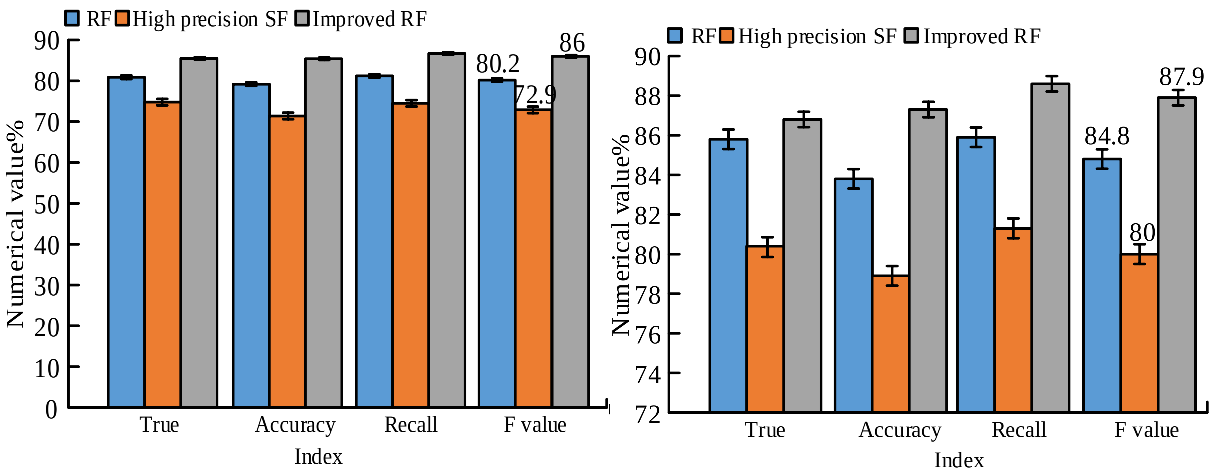


(a) Greman credit dataset test results

(b) Bank marketing dataset test results

Fig. 5. Test results of Greman Credit dataset and Bank Marketing dataset

As shown in Figure 5, Figure 5(a) is the test result of the Greman Credit dataset, and Figure 5(b) is the test result of the Bank Marketing dataset. In both test sets, no matter which indicator is the improved random forest algorithm, the test value is higher. In the Greman Credit dataset, the F-value of the random forest algorithm is as high as 96.3%, and the F-value of the other three algorithms is around 82%. In the Greman Credit dataset, the accuracy, precision, and recall rates of the three algorithms are all stable at around 80%. For the Bank Marketing dataset, the accuracy, precision, and recall rates of the three algorithms fall within the range of 70% to 75% with some fluctuation.



(a)Default of credit clients dataset test results

(b)Test results of incoming dataset

Fig. 6. Test results of Default of Credit Clients dataset and Income dataset

As shown in Figure 6, Figure 6(a) is the test result of the Default of Credit Clients dataset, and Figure 6(b) is the test result of the Income dataset. Firstly, the enhanced random forest model yields improved test values for both datasets. However, in the Income dataset, there is a 5% variation between the accuracy, precision, recall, and F value of the high-precision sub-forest and the random forest algorithm. Above, the fluctuation is more obvious. In order to further evaluate the predictive ability of the model, the classification regression tree and support vector machine model were added in this study to compare with the RF prediction model established in this research. In the support vector machine model added this time, it is divided into different kernel function models, including linear kernel function (liner), polynomial kernel function (polymial), and radial basis function (radial). The three models were implemented in the R language, and the comparison of their prediction results is depicted in Figure 7.

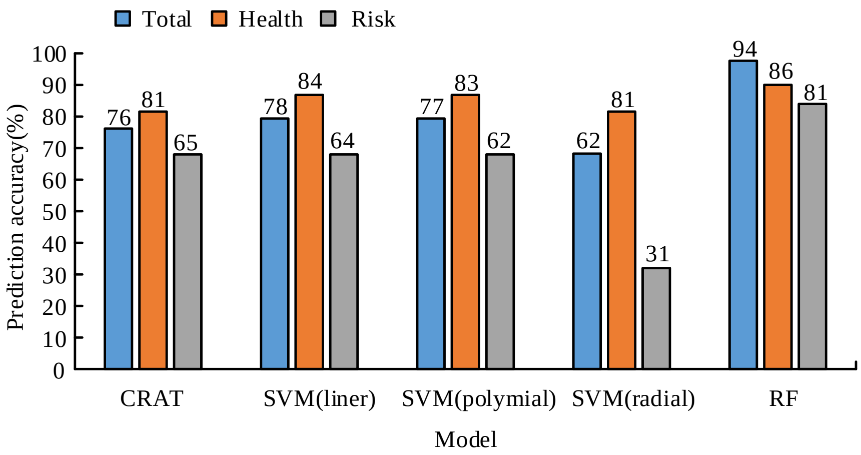
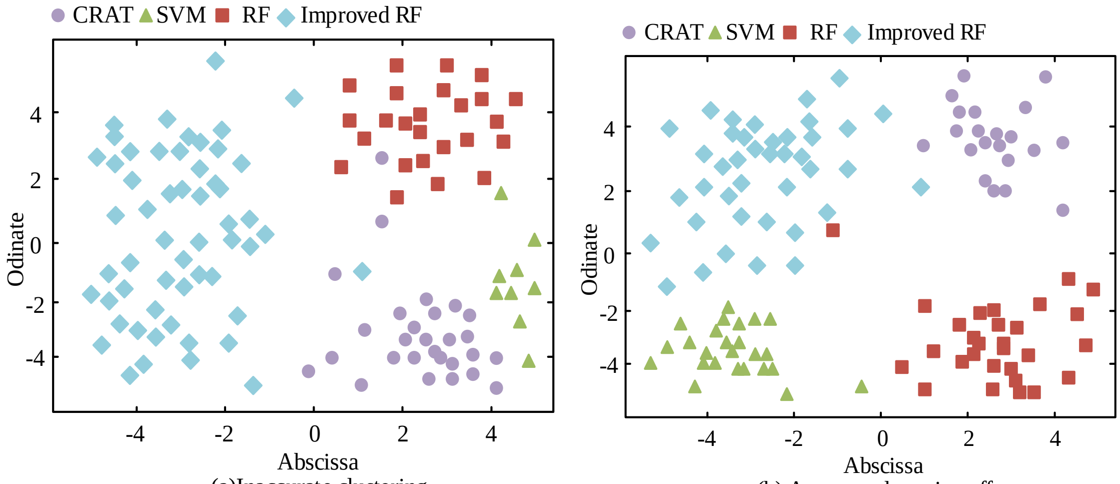


Fig.7. Comparison of the prediction accuracy of the three models

As shown in Figure 7, among the three models, the classification and regression tree model have an accuracy of 81% for healthy enterprises and 65% for risky enterprises. However, when utilizing the linear kernel function, the prediction accuracy for healthy enterprises increases significantly to 84%, yet decreases to 64% for risky enterprises. By contrast, when the polynomial kernel function is employed, the prediction accuracy for healthy enterprises is 83%, but drops to 62% for risky enterprises. It can be seen that the prediction capabilities of the three kernel functions are not much different, and the prediction ability when the kernel function is selected is the strongest; in the random forest algorithm model, the prediction accuracy rate of healthy enterprises is as high as 86%, and the prediction accuracy of risky enterprises is as high as 86%. The accuracy rate is 81%, and the overall accuracy rate is 94%. Compared to the other two models, the RF model demonstrates superior predictive ability in both healthy and venture enterprises. In order to test the improved random forest algorithm, the testSet dataset is introduced to test it. Figure 8 is a set of simulation results of the improved random forest algorithm under the testSet data set.



(a)Inaccurate clustering rendering of the testSet dataset

(b) Accurate clustering effect diagram for the testSet dataset

Fig. 8. TestSet dataset test results

As shown in Figure 8, Figure 8(a) is the inaccurate clustering effect of the testSet dataset, and Figure 8(b) is the accurate clustering effect of the testSet dataset. The figure illustrates that the clustering accuracy of the four algorithms is relatively high. The clustering accuracy of the unimproved RF algorithm is above 52%, the clustering effect of the SVM algorithm is around 37%, the clustering accuracy of the CRAT algorithm is around 51%, and the clustering accuracy of the improved random forest algorithm. around 81%. Through the comparison of the three improved algorithms, it is found that the accuracy of the improved random forest algorithm in the testSet data set has been greatly improved compared with that before the improvement. In this research, the enhanced random forest algorithm underwent further optimization through the integration of Spark distributed computing framework. Figure 9 illustrates the reduction in algorithmic running time achieved after the implementation of this framework.

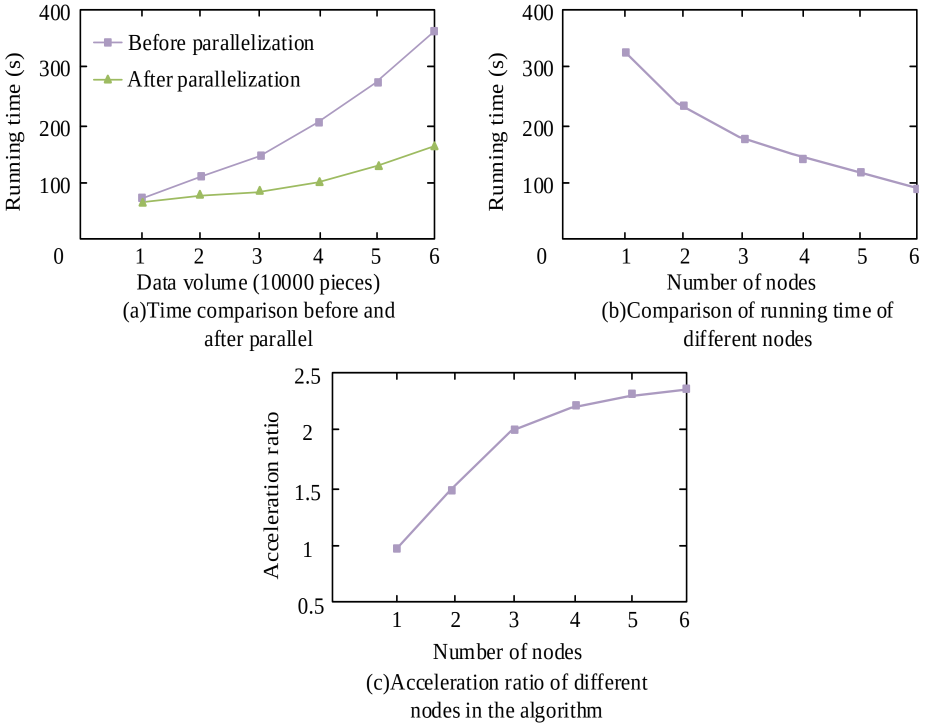


Fig. 9. Algorithm running time after adding Spark computing framework

As shown in Figure 9, Figure 9(a) is the comparison of the running time of the algorithms before and after parallelization, Figure 9(b) is the comparison of the running time of the algorithms under different nodes, and Figure 9(c) is the speedup ratio of the algorithms under different nodes. It can be seen from the algorithm before and after parallelization that when the sample data is less than 20000, the difference between the algorithm running time before and after parallelization is small. When the sample data continues to increase, the gap between the two gradually increases; Figure 9(b) demonstrates that the running time of the algorithm is also influenced by the number of nodes. Allocating the algorithm's tasks to various nodes for parallel computing as the number of nodes increases continuously shortens the algorithm's running time. However, as the number of nodes increases, the speedup ratio does not continue to grow at its current rate. Instead, the growth trend of the speedup ratio is decreasing.

After the financial risk management prevention model is constructed, ensure that the model can be widely used. In order to evaluate and verify the model, the sample data of a company is predicted, and the results are shown in Table 1 by comparing with the expert evaluation prediction.

Table 1
Comparison results between AI models and experts

Type _	Model	N number of accuracy	Accuracy
Regional access	Expert	12	37.5%
	AI	19	59.4%
Enterprise access	Expert	26	56.5%
	AI	32	69.6%
Enterprise credit	Expert	112	55.2%
	AI	133	65.5%

As shown in Table 1, it can be seen that the AI model and experts are used to evaluate and predict three types of regional access, enterprise access and enterprise credit. The results show that in the three types of regional access, enterprise access, and enterprise credit, the accuracy rates of expert prediction results are 37.5%, 56.5%, and 55.2%, respectively; while the AI model prediction results are 59.4%, 69.6%, and 65.6. The AI model has demonstrated a considerable enhancement in both prediction quantity and accuracy. The prediction accuracy of the AI model is significantly higher than the expert prediction and evaluation results.

5. Conclusion

In this era of intertwined finance and the Internet, in order to establish an efficient and safe enterprise financial risk prevention and management framework, this research introduces a new AI algorithm to discuss it. The experimental results show that from the PR curve, it can be seen that the improved recommendation algorithm is better than that before the improvement in terms of accuracy and detection speed. The AP value of the improved random forest algorithm is 0.9919, which is significantly higher than that of the RF algorithm before the improvement, which is 0.9237. The improved algorithm has achieved a better balance between recall rate and performance. Four datasets are introduced to test these three algorithms. In the Greman Credit dataset, the F value of the random forest algorithm is as high as 96.3%, and the F value of the other three algorithms is around 82%. In the Greman Credit dataset, the accuracy, precision, and recall rate of the three algorithms are all stable at about 80%; in the Default of Credit Clients dataset and the Income dataset, the test values of the improved random forest algorithm are higher. However, there are noticeable differences exceeding 5% with apparent fluctuations in the accuracy, precision, recall, and F value of the high-precision sub-forest and random forest algorithms between the Default of Credit Clients dataset and the Income dataset. Finally, through the analysis and prediction of a company’s financial data sample, it is found that in the three types of regional access, enterprise access, and enterprise credit, the accuracy rates of expert prediction results are 37.5%, 56.5%, and 55.2%, respectively. To sum up, the improved random forest algorithm based on decision tree has a good predictive effect on the financial risk of financial enterprises, and also provides a theoretical basis for the new financial technology enterprise financial risk management framework. Future research can further combine algorithms with practical application scenarios, and verify the feasibility and effectiveness of algorithms in practical enterprises through empirical research. This will provide strong support for practical applications and further improve the financial risk management framework of fintech enterprises.

The improved random forest algorithm based on decision trees has made the following contributions to the new financial risk management framework for fintech enterprises:

A new AI algorithm has been developed that can accurately predict the financial risks of financial enterprises. By introducing this new algorithm, research has provided a feasible solution for fintech enterprises.

Improved prediction accuracy and detection speed: The enhanced recommendation algorithm exceeds its predecessor in terms of accuracy and detection speed. This means that fintech companies can more accurately predict and manage financial risks, and can make corresponding response measures more quickly.

Balancing recall and performance: The improved algorithm achieves a better balance between recall and performance. This means that fintech companies can increase the coverage of risk prediction without sacrificing performance, thereby more comprehensively assessing and managing financial risks.

Verified the effectiveness of the algorithm: The study validated the improved random forest algorithm's effectiveness in multiple financial scenarios by testing it on four datasets. In the Greman Credit dataset, the F-value of this algorithm is as high as 96.3%, and it has also achieved good performance in other datasets.

Funding

This research received no external funding.

Data Availability Statement

Data sharing is not applicable to this article.

Conflicts of Interest

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

Acknowledgement

This research was not funded by any grant.

Reference

[1] Radhachandran, A., Garikipati, A., Zelin, N., Pellegrini, E., Ghandian, S., Hoffman, J., Mao, Q., & Das, R. (2021). A Gradient-Boosted Decision-Tree Algorithm for the Prediction of Short-Term Mortality in Acute Heart Failure Patients. Cardiovascular Revascularization Medicine, 28, S19.

[2] Singh, J., & Tripathi, P. (2021). Sentiment analysis of Twitter data by making use of SVM, Random Forest and Decision Tree algorithm. In 2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT) (pp. 193-198). IEEE. https://doi.org/10.1109/CSNT51715.2021.9509679

[3] Si, Y. (2022). Construction and application of enterprise internal audit data analysis model based on decision tree algorithm. Discrete Dynamics in Nature and Society, 2022. https://doi.org/10.1155/2022/4892046

[4] Liu, Q., & Xia, X. (2022). Construction of classification model of academic library websites in Jiangsu based on decision tree algorithm and link analysis method. Open Access Library Journal, 9(1), 1-9. https://doi.org/10.4236/oalib.1108324

[5] Wu, F., Liu, X., Wang, Y., Li, X., & Zhou, M. (2022). Research on evaluation model of hospital informatization level based on decision tree algorithm. Security and Communication Networks, 2022, 1-9. https://doi.org/10.1155/2022/3777474

[6] Wang, L., Wang, H., Zhang, H., Akemujiang, N., & Xiao, A. (2020). Somatotype identification of middle-aged women based on decision tree algorithm. International Journal of Clothing Science and Technology, 33(3), 402-420. https://doi.org/10.1108/IJCST-12-2019-0193

[7] Ariyati, I., Rosyida, S., Ramanda, K., Riyanto, V., & Faizah, S. (2020, November). Optimization of the decision tree algorithm used particle swarm optimization in the selection of digital payments. In Journal of Physics: Conference Series (Vol. 1641, No. 1, p. 012090). IOP Publishing. https://doi.org/10.1088/1742-6596/1641/1/012090

[8] Muditomo, A., & Broto, A. S. (2021). IPO performance prediction during Covid-19 pandemic in Indonesia using decision tree algorithm. Jurnal Keuangan dan Perbankan, 25(1), 132-143. https://doi.org/10.26905/jkdp.v25i1.5137

[9] Saadoon, Y. A., & Abdulamir, R. H. (2021, May). Improved random forest algorithm performance for big data. In Journal of Physics: Conference Series (Vol. 1897, No. 1, p. 012071). IOP Publishing. https://doi.org/10.1088/1742-6596/1897/1/012071

[10] Ning, F., Cheng, Z., Meng, D., & Wei, J. (2021). A framework combining acoustic features extraction method and random forest algorithm for gas pipeline leak detection and classification. Applied Acoustics, 182, 108255. https://doi.org/10.1016/j.apacoust.2021.108255

[11] Song, M., Jung, H., Lee, S., Kim, D., & Ahn, M. (2021). Diagnostic classification and biomarker identification of Alzheimer's disease with random forest algorithm. Brain Sciences, 11(4), 453. https://doi.org/10.3390/brainsci11040453

[12] Zhang, Z., & Cai, Z. (2021). Permeability prediction of carbonate rocks based on digital image analysis and rock typing using random forest algorithm. Energy & Fuels, 35(14), 11271-11284. https://doi.org/10.1021/acs.energyfuels.1c01331

[13] Liu, X. (2021). Empirical analysis of financial statement fraud of listed companies based on logistic regression and random forest algorithm. Journal of Mathematics, 2021, 1-9. https://doi.org/10.1155/2021/9241338

[14] Sonza, R. L., & Tumibay, G. M. (2020). Decision tree algorithm in identifying specific interventions for gender and development issues. Journal of Computer and Communications, 8(2), 17-26. https://doi.org/10.4236/jcc.2020.82002

[15] Mao, L., & Zhang, W. (2021). Analysis of entrepreneurship education in colleges and based on improved decision tree algorithm and fuzzy mathematics. Journal of Intelligent & Fuzzy Systems, 40(2), 2095-2107. https://doi.org/10.3233/JIFS-189210

[16] Munawar, H. S., Mojtahedi, M., Hammad, A. W., Kouzani, A., & Mahmud, M. P. (2022). Disruptive technologies as a solution for disaster risk management: A review. Science of the total environment, 806, 151351. https://doi.org/10.1016/j.scitotenv.2021.151351

[17] Hentzen, J. K., Hoffmann, A., Dolan, R., & Pala, E. (2022). Artificial intelligence in customer-facing financial services: a systematic literature review and agenda for future research. International Journal of Bank Marketing, 40(6), 1299-1336. https://doi.org/10.1108/IJBM-09-2021-0417

[18] Dinesh, T., & Rajendran, T. (2021). Higher classification of fake political news using decision tree algorithm over naive Bayes algorithm. Revista Geintec-Gestao Inovacao E Tecnologias, 11(2), 1084-1096.

[19] Zhang, Y., Balochian, S., Agarwal, P., Bhatnagar, V., & Housheya, O. J. (2014). Artificial intelligence and its applications. Mathematical problems in Engineering, 2014, Article ID 840491. https://doi.org/10.1155/2014/840491

[20] Li, J., Li, X., & Zhang, Z. (2021, April). Dynamic prediction model of bridge project life cycle cost investment based on decision tree algorithm. In IOP Conference Series: Earth and Environmental Science (Vol. 760, No. 1, p. 012049). IOP Publishing. https://doi.org/10.1088/1755-1315/760/1/012049

[21] Toker, D., Sommer, F. T., & D'Esposito, M. (2019). The chaos decision tree algorithm: A fully automated tool for the experimental study of chaotic dynamics. arXiv.