Journal of Network Intelligence ©2024 ISSN 2414-8105
Taiwan Ubiquitous Information Volume **9**, Number **3**, August **2024**

Financial Intelligence Forecasting Model on Regression Analysis and Support Vector Machine

Dan Wang
Inner Mongolia Technical College of Mechanics & Electrics,
Hohhot 010070, P. R. China
1214360775@qq.com

Li-Xin Chen
University of Central Lancashire,
Preston, Lancashire PR1 2HE, United Kingdom
LChen30@uclan.ac.uk

*Corresponding author: Dan Wang

Received June 28, 2023, revised October 19, 2023, accepted January 4, 2024.

Abstract. *The world today is gradually moving into the era of smart economy, and the application of artificial intelligence is bound to trigger huge changes in all walks of life. In the context of the current smart economy, how to use machine learning technology in artificial intelligence to improve the accuracy of enterprise financial analysis has become a hot direction for current research. To address the above issues, this paper proposes a financial intelligence forecasting model based on machine learning models. By establishing a mathematical model to analyse the annual financial statement data published by listed companies, and to determine whether there is fraud according to the model forecasting results. Firstly, from a large number of financial indicators, 60 financial indicators with high frequency of use were selected as variables of the model by using frequency statistics method. Secondly, as financial statement fraud is a typical classification problem, Twin Support Vector Machine (TSVM), a machine learning technique, was chosen and combined with K-Nearest Neighbor (KNN) in order to further improve the forecasting speed and accuracy. In addition, as the data samples for financial statement fraud forecasting are typically unbalanced data, the data are oversampled, undersampled and downsampled in this paper. Finally, for the judgement of model effectiveness, five indicators are selected for analysis in this paper. The experimental results show that compared with other single models, the KNN-TSVM model under the undersampling method has the highest Recall and can effectively identify the fraud samples.*

Keywords: Machine learning; TSVM; KNN; Unbalanced data; Data dimensionality reduction

1. **Introduction.** As a breakthrough in the new round of technological revolution and industrial change, the smart economy is a modern form of economic development based on emerging information technologies such as cloud computing, big data, the Internet of Things and the mobile Internet. The core factors of production in the smart economy are knowledge and data [1,2]. Today's world is gradually moving into a new era of smart economy with artificial intelligence as the core driver. Machine learning as an important part of AI technology is driving changes in various industries [3,4].

With the rapid development of the economy, people are paying more and more attention to the stock market and the research on the stock market never declines. When selecting stocks, in addition to macro factors, it is necessary to pay more attention to the profitability of the company. Financial statements can fully reflect the recent profitability of listed companies [5,6]. Financial statements summarise the overall operating results and financial position of a company over a certain period of time. Publicly released financial statements are the most important basis for investors to understand the company's size, operating conditions, profit potential and other comprehensive levels [7,8]. As we all know, data is the most convincing. If there are errors in the data in financial statements, they can easily be detected by relatively simple calculations. As a result, investors pay a great deal of attention to financial statements [9,10]. However, with the continuous development of the market economy and modern technology, some listed companies have engaged in fraud financial statements for personal gain, and the means of fraud have become increasingly hidden and sometimes difficult to be detected in time. The potential harm of financial statement fraud is enormous. Such dishonest behaviour can cause investors to receive false information and thus make wrong judgments, ultimately leading to great losses of economic interests [11,12]. Therefore, there is an urgent need for an effective method to predict financial fraud. In the context of the current smart economy, how to use machine learning technology in artificial intelligence to predict financial fraud has become a hot direction of current research [13,14].

The purpose of this study is therefore to analyse data from annual financial statements published by listed companies by building mathematical models and to determine whether they are fraud based on the model forecastings. Machine learning allows both top-down verification of hypotheses and bottom-up conclusions from the data without hypotheses [15]. Therefore, this paper applies machine learning methods to the forecasting of financial fraud. As financial statement fraud is a typical classification problem, a support vector machine (SVM) model, which is more accurate in machine learning for classification and forecasting, is used. The samples obtained have a high dimensionality and happen to be classical dichotomous problems. The SVM model has a better advantage in handling data with these characteristics. In this paper, data on financial statements of fraud companies in different years and financial statements of non-fraud listed companies in the corresponding years are collected [16,17,18]. A portion of the acquired data is used to build the model and another portion is used to test the model. Since the number of fraud companies is relatively small in the overall population of listed companies, i.e. since the data sample for financial statement fraud forecasting is typically unbalanced data. Therefore, different data processing methods are used in this paper, including over-sampling methods and under-sampling methods [19]. For the judgement of model effectiveness, five assessment indicators are selected for analysis in this paper.

1.1. **Related Work.**

According to International Standards on Auditing (ISA) 240, financial fraud is the intentional provision and release of false information that results in a material misstatement of financial statements, including the overstatement of assets, sales and profits, or the understatement of liabilities, expenses and losses [20].

In recent years, financial fraud has begun to emerge and continues to grow rapidly, significantly shaking investor confidence and threatening the economic stability [21] of entire countries and even the world. There is a growing demand for greater transparency and consistency and the inclusion of more information in financial statements. Due to the existence of widespread financial fraud, it is important to improve the forecasting of fraud. There are four theories of financial fraud, including the iceberg theory [22], the triangle theory [23], the GONE theory [24] and the risk factor theory [25]. The triangle theory suggests that fraud is caused by a combination of three factors: pressure, opportunity and excuses, as shown in Figure 1. This is currently the more popular theory. Similarly, the GONE theory suggests that fraud consists of the four factors of greed, opportunity, need and exposure, as shown in Figure 2. But it is difficult to detect fraud by management when the above four theories are used with normal audit procedures. In addition, the over-reliance on the experience of professionals results in a fraud detection system that cannot be automated.

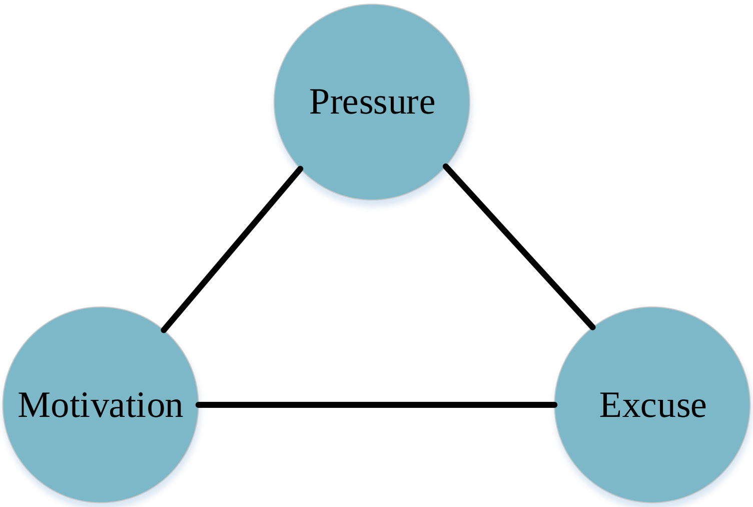

Figure 1. Triangle theory

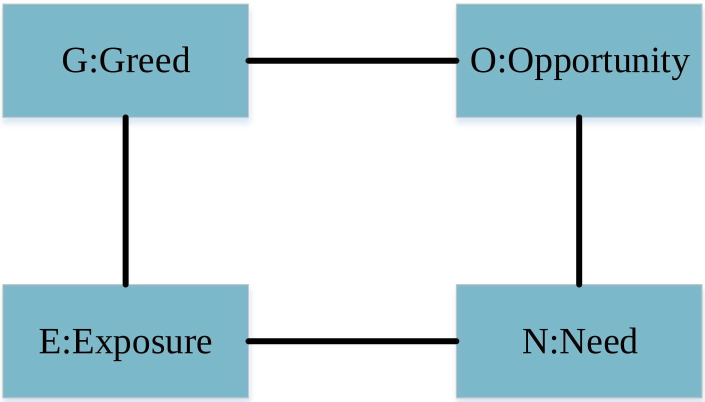

Figure 2. GONE theory

With the advancement of machine learning technology, various computer-aided classification and identification applications have become a hot spot for innovative research in various industries. At this stage, there is widespread interest in how to achieve corporate financial crisis forecasting through machine learning. In order to achieve an automated malpractice detection system, researchers have proposed financial fraud forecasting models. The commonly used forecasting techniques are mainly based on various classification algorithms, including logistic statistical detection, regression analysis, neural networks, decision trees, Bayesian networks and so on. Of these, the main theoretical approaches used fall into two categories: regression analysis and neural networks. Regression analysis is commonly used to discover hidden data information. Neural networks generally outperform regression models in terms of forecasting effectiveness and accuracy. In general, the forecasting effectiveness of the models is better than unaided auditor detection. A comparison of financial fraud forecasting models is shown in Table 1.

Table 1. Comparison of financial forecasting models

Forecasting models	Disadvantages	Advantages	Precision
Logistic statistical testing	Highly subjective; lack of automatic access to rules	Very good logical reasoning skills and	Low
Regression analysis	Poor applicability of a single model	Simple models and fast forecastings	High
Bayesian Networks	Highly subjective	Greater flexibility	High
Decision Trees	High requirements for data reliability	No requirement for sample size	Low
Neural networks	Slower forecasting	High precision	High

Regression analysis forecasting models are mainly used to analyse data from a sample of companies through various classifier algorithms to obtain financial crisis forecasting models. For example, Engels et al. [26] developed a Linear Probability Model (LPM) and a logistic regression model for 48 fraud and 92 non-fraud companies respectively, both of which had a misclassification rate of less than 36%. Yang [27] conducted a comparative study of decision tree, K-Nearest Neighbor (KNN) and neural network based risk forecasting models for securities companies. Jaramillo et al. [28] proposed a fraud forecasting model based on SVM to identify the risks assessed by listed companies. The experimental results showed that the SVM model was able to successfully predict financial statement fraud.

Neural network forecasting models are mainly obtained by training the neural network architecture and optimising its various parameters through the BP algorithm to obtain a financial crisis forecasting model. For example, Uthayakumar et al. [29] proposed a financial crisis forecasting method based on ant colony neural network, which uses ant colony algorithm to find the best structure and parameters of the neural network model so as to improve the forecasting accuracy. Zhou et al. [30] used particle swarm algorithm to construct a neural network forecasting model, which improved the global optimisation finding ability of the forecasting model, thus obtaining a better enterprise financial risk forecasting results.

Compared to neural network forecasting models, regression analysis forecasting models balance operational efficiency and forecasting accuracy. Therefore, the research direction chosen for this paper is a financial fraud forecasting model based on regression analysis. By analysing the above research, it is found that KNN has the problem of slow computational efficiency when faced with high-dimensional data in forecasting tasks, while SVM exhibits poor accuracy in sample point classification. As one of the most widely used algorithms for data dimensionality reduction, Principal Component Analysis (PCA) of multivariate statistics has excellent performance in feature extraction from high-latitude data. In addition, Twin Support Vector Machine (TSVM) is a new type of machine learning method based on statistical learning theory [31,32].

1.2. **Motivation and contribution.** As a variation of the traditional SVM, TSVM inherits its excellent learning capability, but runs four times more efficiently than the traditional SVM. Therefore, this paper attempts to combine KNN and TSVM to build a financial fraud forecasting model. The main innovations and contributions of this paper include:

(1) As the data sample for financial forecasting is typically unbalanced data. Therefore, this paper uses different data processing methods, including oversampling and undersampling. Also, in order to map the high-dimensional data into a low-dimensional space while

retaining the main data feature information, this paper utilises the PCA principal component extraction function of the spss1 9.0 software to perform a dimensionality reduction of the data.

(2) A combination of KNN and TSVM is used to complete the financial forecasting. By combining the classification methods for forecasting, the problem of slow KNN recognition is solved and the model forecasting efficiency is improved. In addition, the optimal weight assignment is used to improve the accuracy of the forecasting algorithm. The experimental results verified the effectiveness and accuracy of the combined model.

2. **Support vector machines.**

2.1. **Classification hyperplane and maximum interval.**

Support vector machines (SVMs) are generalised linear classifiers for the binary classification case [31], which can find the global optimal solution from a certain number of samples, are sparse and stable, and can classify and predict high-dimensional non-linear objectives using kernel functions. SVMs are widely used in finance, biology, environment, industry, medicine and other fields because they overcome the problems of "over-learning" and "dimensional catastrophe" to a large extent.

The basic principle of SVM is to find an optimal classification hyperplane that satisfies the classification requirements. The hyperplane is able to maximise the blank area on either side of it while ensuring classification accuracy. Theoretically, SVM can achieve optimal classification of linearly separable data. Taking binary classification data as an example, assume a training sample set is $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$ and class labels is $y_i \in \{-1, +1\}$. Find a classification hyperplane in the sample space that separates the two classes of samples in D. There may be many such classification hyperplanes. What we need is the classification hyperplane that maximizes the separation between the two classes of training samples, called the "optimal classification hyperplane". The optimal classification hyperplane is the most fault-tolerant to local perturbations of the training samples and produces the most robust classification results.

In the sample space, the classification hyperplane can be expressed as the following linear equation.

$$\omega^T \mathbf{x} + b = 0 \tag{1}$$

where $\boldsymbol{\omega}$ is the direction of the classification hyperplane and b is the distance between the classification hyperplane and the origin. The classification hyperplane $(\boldsymbol{\omega}, b)$ can be determined from these two parameters. r represents the distance from the sample point $\boldsymbol{x}$ to the classification hyperplane $(\boldsymbol{\omega}, b)$.

$$r = \frac{\left|\omega^T \mathbf{x} + b\right|}{\|\omega\|} \tag{2}$$

To find the categorical hyperplane with the maximum "interval", we look for parameters $\boldsymbol{\omega}$ and b that maximize r.

$$\max_{\mathbf{w}, b} \frac{2}{\|\omega\|} \text{s.t.} y_i \left(\omega^T \mathbf{x}_i + b\right) \geq 1, i = 1, 2, \ldots, m \tag{3}$$

2.2. **Lagrange multipliers and the dual problem.**

From the above analysis, it can be seen that the problem of finding the optimal classification hyperplane for the SVM is in fact a constrained optimal solution problem. We add Lagrange multipliers to each constraint in Equation (3).

$$L(\omega, b, \alpha) = \frac{1}{2}\|\omega\|^2 + \sum_{i=1}^{m} \alpha_i \left(1 - y_i \left(\omega^T \mathbf{x}_i + b\right)\right) \tag{4}$$

The solution to the constrained optimization problem is determined by the saddle point of the Lagrangian function. At the same time, the bias derivatives of the parameters $\boldsymbol{\omega}$ and b are satisfied at the saddle point to be zero. The problem can be transformed into the corresponding dyadic problem.

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j, s.t. \sum_{i=1}^{m} \alpha_i y_i = 0, \alpha_i \geq 0 \tag{5}$$

In order to obtain the optimal solution, the optimal parameters $\boldsymbol{\omega}^*$ and b^* need to be calculated to obtain the optimal classification hyperplane $\boldsymbol{\omega}^{*T}\boldsymbol{x} + b^* = 0$.

$$\omega^* = \sum_{j=1}^{m} \alpha_j^* y_j \mathbf{x}_j^T \tag{6}$$

$$b^* = y_i - \sum_{i=1}^{m} y_j \alpha_j^* \left(\mathbf{x}_j^T \cdot \mathbf{x}_i \right) \tag{7}$$

2.3. **Feature space and kernel functions.** The above SVM theory is based on the assumption that the training samples are linearly divisible, however the actual data may be linearly indivisible.

In this case, the main idea of SVM is to map the samples from the original space to a higher dimensional feature space. The mapping allows the samples to be made linearly divisible within this feature space, and thus the optimal classification hyperplane can be constructed in that feature space. The choice of feature space is very important. The kernel function can be approximately equivalent to the feature space as shown in Equation (2), so the choice of the kernel function becomes the key to the construction of the SVM model. At this stage, there are five main types of kernel functions: linear kernel, polynomial kernel, Radial Basis Function (RBF) kernel, Laplace kernel and Sigmoid kernel. The more commonly used kernel is the Sigmoid kernel.

$$\kappa \left(\mathbf{x}_i, \mathbf{x}_j \right) = \tanh \left(\beta \mathbf{x}_i^T \mathbf{x}_j + \theta \right) \tag{8}$$

where *tanh* is the hyperbolic tangent function.

2.4. **Twin support vector machines.** As an improved version of the traditional SVM, the TSVM finds a pair of non-parallel hyperplanes, and therefore has a much better classification capability and is well suited to solving approximate types of sample classification problems [33]

In addition, TSVM performs two SVM-type problem solving and is therefore more computationally efficient than traditional SVMs. The time complexity of the standard SVM is about $O(m^3)$, while the time complexity of TSVM is $O(2 * (m/2)^3)$ and m denotes the number of samples, which shows that the computational overhead is about 1/4 of that of the standard SVM.

Depending on the characteristics of the financial data, most of the data samples are nonlinear classification problems, and when linearly inseparable, this requires the introduction of kernel function solutions to solve the problem. Assume that the training sample set in the n-dimensional real space R^n is $(x_j^i, y_j), \quad i = 1, 2, \quad j = 1, 2, \ldots, m.$ The total number of samples is $m = m_1 + m_2$, where m_1 is the number of positive class sample points and m_2 is the number of negative class sample points. Then the method for finding the nonlinear hyperplane [34] shown as follows:

$$K(x^T, C^T)u_1 + b_1 = 0, \quad K(x^T, C^T)u_2 + b_2 = 0 \tag{9}$$

where u is the normal vector of the hyperplane, b is the offset, and the subscript symbols of both denote positive and negative class samples respectively. K is the kernel function. $C^T = [AB]^T$, A is the sample matrix composed of positive class samples and B is the matrix composed of negative class samples.

Similarly, the plane that divides the positive and negative classes is obtained by solving two quadratic programs.

$$\min_{u_1 b_1 \xi} \frac{1}{2}\left\|K\left(A, C^T\right)u_1 + e_1 b_1\right\|^2 + c_1 e_2^T \xi, \text{s.t.} - \left(K\left(B, C^T\right)u_1 + e_2 b_1\right) + \xi \geq e_2, \xi \geq 0 \quad (10)$$

$$\min_{u_2 b_2 \xi} \frac{1}{2}\left\|K\left(B, C^T\right)u_2 + e_2 b_2\right\|^2 + c_2 e_1^T \xi, \text{s.t.} \left(K\left(A, C^T\right)u_2 + e_1 b_2\right) + \xi \geq e_1, \xi \geq 0 \quad (11)$$

To further simplify Equation (10) and Equation (11), they are pairwise transformed as follows:

$$\max_{\alpha} e_2^T \alpha - \frac{1}{2}\alpha^T R(S^T S)^{-1} R^T \alpha, \quad s.t. \quad 0 \leq \alpha \leq c_1 e_2 \quad (12)$$

$$\max_{\gamma} e_1^T \gamma - \frac{1}{2}\gamma^T S(R^T R)^{-1} S^T \gamma, \quad s.t. \quad 0 \leq \gamma \leq c_2 e_1 \quad (13)$$

$$R = \left[K(B, C^T)e_2\right], S = \left[K(A, C^T)e_1\right] \quad (14)$$

Solving Equation (12) and Equation (13) gives:

$$(u_1^T, b_1)^T = -(S^T S)^{-1} R^T \alpha \quad (15)$$

$$(u_2^T, b_2)^T = -(R^T R)^{-1} S^T \gamma \quad (16)$$

The hyperplane of the classification can be obtained from u_1, u_2, b_1 and b_2. The classification decision function is shown as follows:

$$classlabel = \arg \min_{k=+1,-1} \left[K(x^T, C^T)u_k + b_k\right] \quad (17)$$

3. **Selection of indicators for forecasting models.**

In order to construct an effective financial fraud forecasting model, it is first necessary to establish the most objective and representative system of financial indicators in order to fully reveal the financial situation of the company. As there are missing values in the downloaded data, a preliminary treatment of the data is made before further analysis.

Firstly, indicators with missing values greater than 50% were removed and the remaining data was filled in with missing values, depending on the actual data. This is done because to eliminate the entire data directly would result in a much smaller sample size. In cases where there is not enough data, this may lose a lot of important information that is hidden or lead to changes in the characteristics and distribution of the data. As the padding is our subjective estimate of the missing values (which does not necessarily match the actual situation, and inappropriate padding may add noise), we have uniformly padded the missing values to zero.

Secondly, as each company is of a different size and the data in the company's balance sheet, income statement and cash flow statement are absolute values that would have a significant impact on the forecasting results if left untreated, the relevant indicators in these three tables are divided by the total assets of that sample company. This relative value is used to replace the corresponding indicator in the original sample. At the same time, the other ratio financial indicators remain unchanged. Finally, we add class labels to each sample. The fraud sample is 1 and the non-fraud sample is 0.

In addition, for each financial indicator, we classify it according to its attributes. As the selection and classification of financial indicators are subjective, and the data we obtained includes both absolute and ratio indicators, all indicators are divided into five categories: size indicators, liquidity indicators, operating indicators, profitability indicators and growth indicators. On the basis of the existing research results, a system of financial indicators was constructed by using frequency statistics to select 60 frequently used financial indicators from a large number of financial indicators, which mainly reflect the company's long-term and short-term solvency, operating capacity, development capacity and profitability.

4. **KNN-TSVM based financial fraud forecasting model.**

4.1. **PCA-based dimensionality reduction.**

In order to map the high-dimensional aero-engine fault data into a low-dimensional space while retaining the main data features, this paper uses the PCA principal component extraction function of SPSS 9.0 software to perform dimensionality reduction of the acquisition data.

As one of the most commonly used linear dimensionality reduction methods, PCA is able to reduce the dimensionality of the original features by projection while keeping the "amount of information intact". Assume that the model sample $X = (X_1, X_2, ..., X_m)^T$ is a sample of data consisting of, for example, industrial system failure characteristics, each with n features, and that the main influencing factors need to be extracted from these explanatory variables. The training sample can be denoted as $x_1, x_2, \ldots, x_m$ and the standard deviation is denoted as $S_1, S_2, \ldots, S_m$. The normalised transformation is then shown as follows:

$$Y_j = a_{j1}x_1 + a_{j2}x_2 + \ldots + a_{jm}x_m, \quad j = 1, 2, \ldots, m \tag{18}$$

where a_{jm} is the coefficient factor corresponding to the training sample x_m. First, if the value of Y_1 is equal to the value of the orthogonal unit of the corresponding eigenvalue and the variance of Y_1 is the largest, then Y_1 can be identified as the first principal component.

Secondly, if the value of Y_2 is equal to the value of the orthogonal unit of the corresponding eigenvalue, the covariance between Y_1 and Y_2 is zero and the variance of Y_2 is the largest, then Y_2 can be identified as the second principal component.

By the same token, multiple principal components can be obtained by analogy.

The contribution rate η_i of the i-th principal component Y_i in the cumulative contribution rate calculation is shown as follows:

$$\eta_i = \frac{\lambda_i}{\sum_{j=1}^{k} \lambda_j}, \quad i = 1, 2, \ldots, k \tag{19}$$

The total contribution of the first m principal components, CPV, is shown as follows:

$$CPV = \sum_{i=1}^{m} \frac{\lambda_i}{\sum_{j=1}^{k} \lambda_j} \times 100\%, \quad m < k \tag{20}$$

In general, it is necessary to ensure that the value of CPV is greater than 85%.

4.2. **Primary classification model based on TSVM.** KNN is a classical non-parametric regression classification algorithm. KNN has the advantage of high forecasting accuracy when facing classification tasks in complex environments, but is computationally intensive when dealing with high-dimensional data.

TSVM, on the other hand, also has a good performance of being fast in classification problems. Therefore, a combination of KNN and TSVM is used to accomplish financial fraud forecasting. By combining classification methods for forecasting, the problem of slower KNN recognition is solved and the model forecasting efficiency is improved. In addition, the optimal weight assignment is used to improve the accuracy of the forecasting algorithm. The experimental results validate the effectiveness and accuracy of the combined model. Firstly, TSVM was used to classify the original data samples to form a smaller number of sub-sample sets, thus generating the primary classification model. Then, KNN was used to perform secondary classification on the sub-sample set to finally obtain the forecasting results.

The essential goal of the TSVM is classification algorithm is to obtain the optimal classification hyperplane for both classes of samples and therefore requires a high-dimensional linear transformation. The decision function typically used is shown below.

$$\text{plabel} = \operatorname{sgn}\left(\sum_{i=1}^{N} y_i \alpha_i K(x_i, x) + b\right) \tag{21}$$

where plabel denotes the decision function, x denotes the sample to be predicted by the label, the subscript i is the sample number, y_i is the value of the sample, α_i denotes the parameter coefficients, x_i is the support vector sample, $K(x_i, x)$ is the kernel function, and b is the constant term of the hyperplane.

$$K(x_i, x_j) = \phi(x_i^T)\phi(x_j) \tag{22}$$

where $\phi(\cdot)$ denotes the non-linear function.

The size indicator, liquidity indicator, operating indicator, profitability indicator and growth indicator are selected as the feature vectors and TSVM is used to classify the historical financial statement data samples of listed companies into several sub-data sample sets.

(1) Selection of the kernel function. In this paper, the RBF kernel function is selected to construct the expressions of the decision function.

$$\text{plabel} = \operatorname{sgn}\left(\sum_{i=1}^{n} y_i \alpha_i \exp\left(-\gamma \|x_i - x\|^2\right) + b\right) \tag{23}$$

where (x_i, y_i) represents a training sample data point.

(2) Cross-validation method to obtain optimal penalty coefficients and parameters γ .

(3) Generation of sub-data sample sets through classification labels.

4.3. **KNN-based secondary classification model.** This paper focuses on a sample of two types of company financial statements (fraud and non-fraud).

The category of the fraud sample is "1" and the category of the non-fraud sample is "0". Let the test sample set be $x = \{x_1, x_2, ..., x_n\}$ and n denote the number of companies to be assessed. The training sample set is $y = \{y_1, y_2, ..., y_m\}$ and m represents the number of training samples. Each training sample y_j has a known category, i.e. y_j is known to be "0" or "1".

Let x_{ia} be the predictor a for company x_i and y_{ja} be the predictor a for the training sample y_j, then $d(x_i, y_j)$ is the Euclidean distance between x_i and the training sample y_j.

$$d(x_i, y_j) = \sqrt{\sum_{a=1}^{S} (x_{ia} - y_{ja})^2} \tag{24}$$

where S indicates the number of indicators in the financial statement fraud forecasting model.

When $j = 1, 2, ..., m$ is available, sort $d(x_i, y_j)$ in order from smallest to largest and select the top K for analysis.

Assuming that the number of training samples belonging to the non-fraud category in the top K is δ and the number of training samples belonging to the fraud category is θ, then $H(x_i)$ is the forecasting model for financial statement fraud crisis of listed companies.

$$H(x_i) = \begin{cases} 0, & \delta > \theta \\ 1, & \delta < \theta \\ \text{False}, & \delta = \theta \end{cases} \tag{25}$$

where $\delta + \theta = K$.

4.4. **Optimal weight assignment.** In order to improve the forecasting accuracy for the test sample points, the neighbour with the higher contribution needs to be selected, so an optimal weight assignment method is introduced.

The weights for each nearest neighbour are calculated as shown below.

$$w_i = \frac{1/d_i}{d} \tag{26}$$

where d_i represents the Euclidean distance between the i-th nearest neighbour and the current data.

$$d = \sum_{i=1}^{K} \frac{1}{d_i} \tag{27}$$

where d denotes the sum of the reciprocal of all distances. Then the final financial crisis classification forecasting model for the company x_i.

$$\hat{H} = \sum_{i=1}^{K} w_i H(x_i) \tag{28}$$

4.5. **Forecasting process.** The flow of the KNN-TSVM financial statement fraud forecasting model is shown in Figure 3.

5. Example forecastings and analysis of results.

5.1. **Sample data selection and pre-processing.** We looked up the list of companies that received administrative penalties between 2013 and 2018 on the official website of the China Securities Regulatory Commission(CSRC).

There are various reasons for listed companies to be administratively punished, such as fraud issuance, illegal information disclosure, insider trading, market manipulation, misappropriation of company assets, etc. We have selected from the above companies a total of 63 companies that have been administratively penalised for committing financial statement fraud. The stock codes of the financial fraud companies (in part) are shown in Table 2. It can be observed that there is a certain lag in listed companies being penalised by the CSRC for financial fraud. It generally takes two years (or even ten years) before

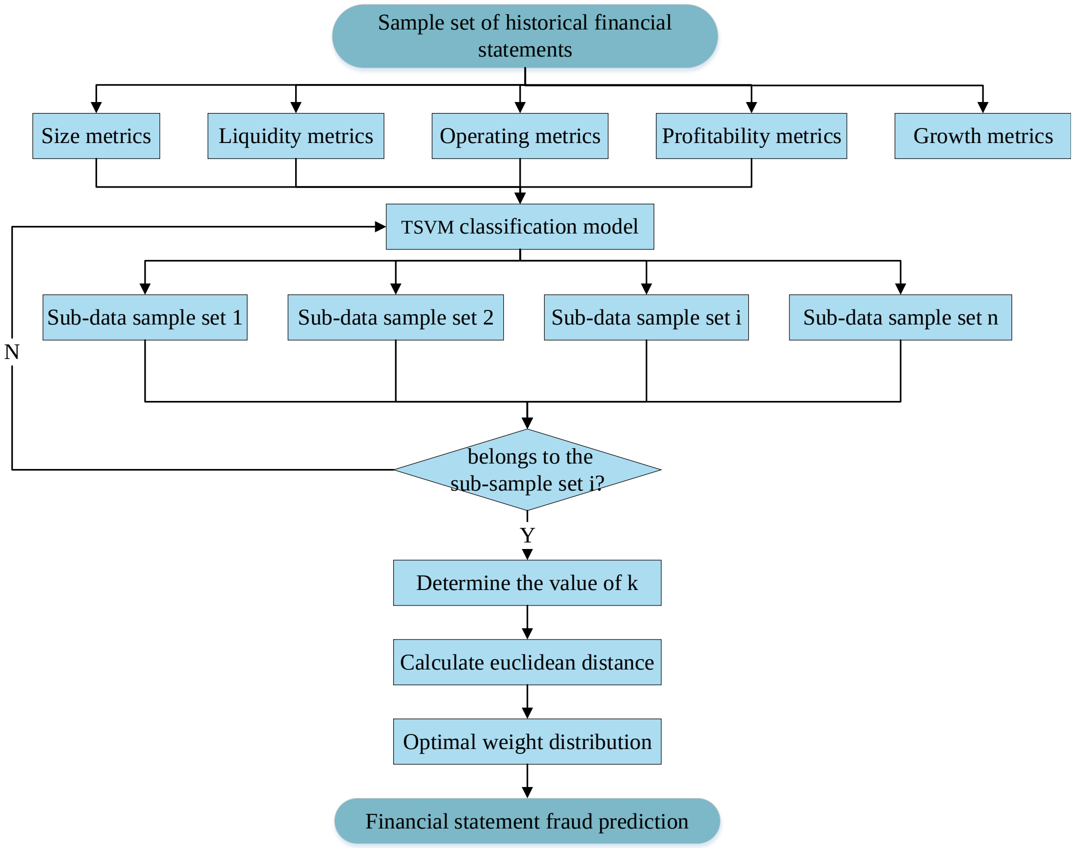
Sample set of historical financial statements
Size metrics
Liquidity metrics
Operating metrics
Profitability metrics
Growth metrics
TSVM classification model
Sub-data sample set 1
Sub-data sample set 2
Sub-data sample set i
Sub-data sample set n
N
belongs to the sub-sample set i?
Y
Determine the value of k
Calculate euclidean distance
Optimal weight distribution
Financial statement fraud prediction

Figure 3. Flow of the financial intelligence forecasting model

a listed company is found to have committed financial fraud. This is because the CSRC has to go through the process of investigation and evidence collection, case hearing and hearing and review in order to perform its duties.

When 60 financial indicators were selected for forecasting analysis on the test set. The financial indicators were normalized in order to eliminate the effect of numerical units in the financial indicators.

$$V_i' = \frac{V_i - V_{\min}}{V_{\max} - V_{\min}} \tag{29}$$

where $V_{\max}$ is the maximum value of the indicator and $V_{\min}$ is the minimum value of the indicator.

Before the sample t-test, the consistency of variance between the two groups of data for fraud and non-fraud companies needs to be examined. In this paper, a two-sample variance F-test was used, yielding $F = 0.07$, $P = 0.3 > 0.05$, meaning that the 60 financial indicators were not statistically significantly different. A two-sample variance t-test was then used, yielding $t = 0.454$ and $P = 0.656$. Therefore, there is no significant variability between the samples of the two types of companies, thus validating the reliability of the selected indicators.

5.2. **Performance evaluation methods for forecasting models.** In order to objectively assess the performance of a forecasting model, we have to choose appropriate evaluation metrics for the classifier. In the evaluation of binary classification problems, we generally use metrics such as *Accuracy*, *Precision*, *Recall* and *F-Score*. For binary classification problems, a confusion matrix can be derived from the test set samples according to different combinations, as shown in Table 3. Based on the confusion matrix the above

Table 2. Stock codes of financial fraud companies.

Serial number	**Stock code**	**Time of being punished**	**Time of financial fraud**
1	000611	2016/6/30	2012-2013
2	300372	2016/715	2013
3	601519	2016/7/20	2013
4	600598	2016/8/18	2011
5	002330	2016/9/27	2014
6	002608	2016/10/24	2013
7	000995	2016/12/12	2015
8	300126	2016/12/21	2015
9	600810	2016/12/23	2014-2015
10	600247	2017/3/13	2012
11	300117	2017/4/25	2008
12	600800	2017/5/16	2006-2012
13	02715	2017/5/31	2010
14	600656	2017/6/29	2011
15	600281	2017/7/5	2014
16	600318	2017/9/18	2015
17	002490	2017/9/21	2015
18	000798	2017/11/15	2015
19	000922	2017/12/1	2013
20	002323	2017/12/14	2015
21	002288	2017/12/18	2014
22	000511	2017/12/19	2015
23	000693	2018/1/23	2013
24	600806	2018/2/5	2013
25	002194	2018/2/28	2016
26	300028	2018/3/1	2014
27	600680	2018/3/21	2014
28	300208	2018/4/11	2014-2015
29	600610	2018/4/12	2015
30	002070	2018/5/31	2016-2017
31	300267	2018/6/15	2015
32	002473	2018/8/30	2014-2016
33	000519	2018/10/30	2014-2016
...	...	...	...
63	300269	2018/12/17	2014-2016

common metrics can be calculated.

Table 3. Example of confusion matrix.

		Real Category	
		1(True)	0(False)
Forecasting Category	1(Positive)	TP(True Positive)	FP(False Positive)
	0(Negative)	FN(False Negative)	TN(True Negative)

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{30}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{31}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{32}$$

$$F = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{33}$$

5.3. **Experimental configuration and parameter settings.** The experimental hardware environment is: Intel i7 6700k CPU, 8 GB RAM, 300G hard disk; the experimental software environment is: Window7 64-bit, MATLAB 10. The matlab function used for the cross-validation is crossval.

The optimization diagram of the penalty coefficient c and the parameter γ for TSVM classification in the forecasting model is shown in Figure 4. It can be seen that the optimal penalty coefficients c and parameters γ take values of 0.25 and 0.1758 respectively. The size of the K value directly affects the accuracy of the KNN-TSVM forecasting model. The effect of K value on the forecasting accuracy of the model was obtained after several iterations of experiments as shown in Figure 5.

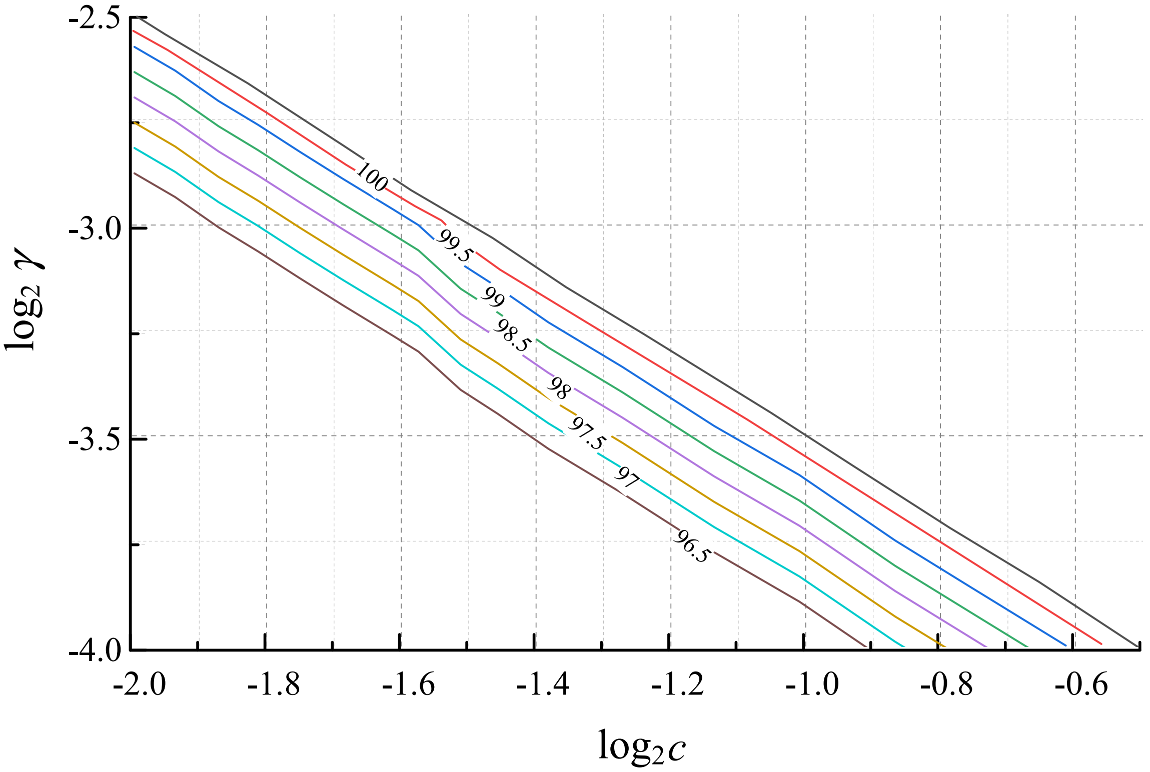

Figure 4. Parametric optimization schematic

It can be seen that the Mean Absolute Error (MAE) is minimised when K is taken to be [2,4]. Therefore, $K = 3$ was chosen so that the $\delta = \theta$ case does not occur.

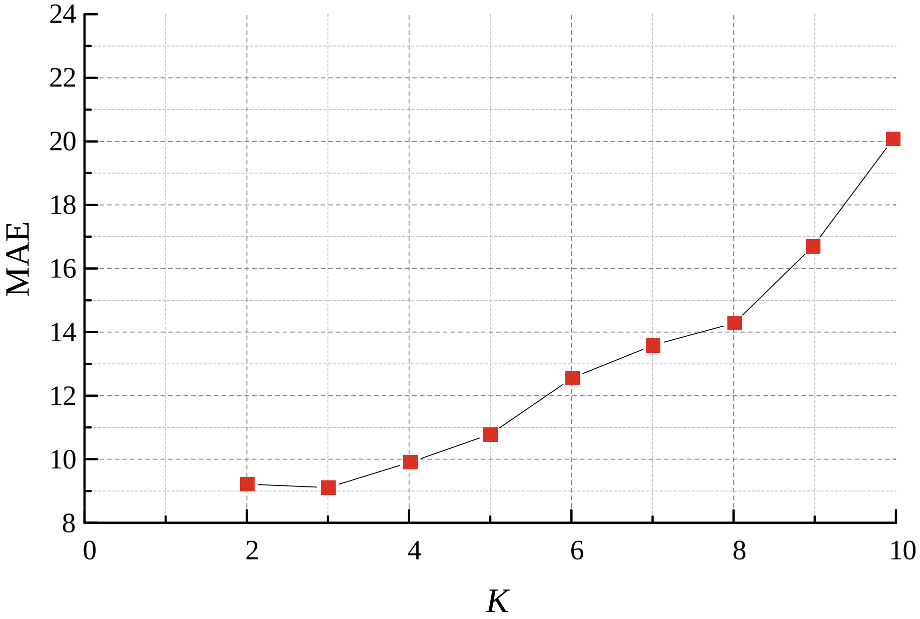

Figure 5. Effect of K values on model forecasting accuracy

5.4. **Under-sampling and over-sampling of unbalanced data samples.** The original data were divided into a training set and a test set in chronological order.

The training set was the first 70% (84) fraud samples and the first 70% (6446) non-fraud samples. The test set was the second 30% (36) fraud samples and the second 30% (2762) non-fraud samples. Since the sample size of financially fraud companies in the actual data is much smaller than the overall number of listed companies (typical of unbalanced data), over-sampling and under-sampling methods are used to process the original training set respectively.

The advantage of under-sampling is that it is relatively simple to perform, while the disadvantage is that important information is easily lost. a typical approach to under-sampling is the EasyEnsemble algorithm [35]. The EasyEnsemble algorithm uses each training subset to train a base classifier, and then integrates all base classifiers to obtain the final classifier. Over-sampling is the repeated extraction of minority class samples. The advantage of over-sampling is that important information from the original dataset is preserved, while the disadvantage is that it tends to over-fit. a typical approach to over-sampling is the SMOTE algorithm. The SMOTE method applies interpolation to create new minority class samples, rather than simply copying samples or assigning weights, thus dealing with unbalanced data.

5.5. **Comparative evaluation of model forecasting results.** The forecasting results of the undersampled model under the EasyEnsemble method are shown in Table 4.

Table 4. Forecasting effects of under-sampling models

Models	Accuracy	Precision	Recall	F-Score	Time (s)
Logistic	0.5972	0.0194	0.6111	0.0376	0.22
Decision Trees	0.5993	0.0206	0.6257	0.0389	0.23
SVM	0.6052	0.0228	0.6537	0.0426	0.22
KNN-TSVM	0.6061	0.0233	0.7222	0.0451	0.20
BP Neural Networks	0.6773	0.0252	0.7189	0.0485	0.41

Under the SMOTE method, the forecasting results of the over-sampling model are shown in Table 5. Among the five models, the KNN-TSVM forecasting model has the highest Recall, which indicates that its ability to identify fraud samples is the best. This is due to the use of SVM to divide the historical financial statement sample set into several sub-sample sets, which improves the classification accuracy and generalisation ability. In addition, the KNN-TSVM forecasting model achieved excellent results on the Recall metric. This is due to the use of optimal weight assignment to solve the problem of distributing contribution values among traditional state vectors, resulting in more stable accuracy performance. Collectively, the KNN-TSVM forecasting model under the undersampling approach has the highest check-all rate compared to the over-sampling model. In addition, the running time of the KNN-TSVM forecasting model under the undersampling method is significantly reduced, that is, the forecasting speed is improved.

Table 5. Forecasting effects of over-sampling models

Models	Accuracy	Precision	Recall	F-Score	Time (s)
Logistic	0.4772	0.0181	0.5211	0.0286	0.24
Decision Trees	0.4793	0.0193	0.5357	0.0299	0.25
SVM	0.4852	0.0215	0.5637	0.0336	0.23
KNN-TSVM	0.4861	0.022	0.6322	0.0361	0.21
BP Neural Networks	0.5573	0.0239	0.6289	0.0395	0.45

6. **Conclusion.** In this paper, we propose to combine KNN and TSVM to construct a fraud forecasting model for corporate financial statements. Sixty representative financial indicators were selected and normalized. TSVM was used to divide the sample set of historical corporate financial statements of listed companies into several sub-sample sets. A KNN was used to sub-classify the sub-data sample sets. In addition, the KNN-based model was optimised using the nearest neighbour weight assignment. As the data samples for financial statement fraud forecasting are typically unbalanced data. Therefore, different data processing methods are used in this paper, including over-sampling methods and under-sampling methods. The experimental results validate the effectiveness and sophistication of the KNN-TSVM forecasting model. However, there are still limitations in the selection of indicators, and more indicators such as financial or operational conditions will be incorporated for follow-up research.

REFERENCES

[1] T.-Y. Wu, F. Kong, Q. Meng, S. Kumari, and C.-M. Chen, "Rotating behind security: an enhanced authentication protocol for IoT-enabled devices in distributed cloud computing architecture," EURASIP Journal on Wireless Communications and Networking, vol. 2023, no. 1, 3701, 2023.

[2] T.-Y. Wu, Q. Meng, Y.-C. Chen, S. Kumari, and C.-M. Chen, "Toward a Secure Smart-Home IoT Access Control Scheme Based on Home Registration Approach," Mathematics, vol. 11, no. 9, 2123, 2023.

[3] T.-Y. Wu, L. Wang, and C.-M. Chen, "Enhancing the Security: A Lightweight Authentication and Key Agreement Protocol for Smart Medical Services in the IoHT," Mathematics, vol. 11, no. 17, 3701, 2023.

[4] A. Kokkinou, G. M. Korres, and E. Papanis, "Blue smart economy-A current approach towards growth," Smart Cities and Regional Development Journal, vol. 2, pp. 81-90, 2018.

[5] J. M.-T. Wu, Q. Teng, S. Huda, Y.-C. Chen, and C.-M. Chen, "A Privacy Frequent Itemsets Mining Framework for Collaboration in IoT Using Federated Learning," ACM Transactions on Sensor Networks, vol. 19, no. 2, pp. 1-15, 2023.

[6] M.-E. Wu, H.-H. Tsai, W.-H. Chung, and C.-M. Chen, "Analysis of Kelly betting on finite repeated games," Applied Mathematics and Computation, vol. 373, 125028, 2020.

[7] R. Runanto, M. F. Mislahudin, F. A. Alfiansyah, M. K. M. Taqiyyah, and E. T. Tosida, “Potential classification of Smart Village–Smart Economy with Deep Learning methods,” International Journal of Quantitative Research and Modeling, vol. 2, no. 3, pp. 147-162, 2021.

[8] X. Huang, H. Xiong, J. Chen, and M. Yang, “Efficient Revocable Storage Attribute-based Encryption with Arithmetic Span Programs in Cloud-Assisted Internet of Things,” IEEE Transactions on Cloud Computing, vol. 11, no. 2, pp. 1273-1285, 2023.

[9] H. Xiong, X. Huang, M. Yang, L. Wang, and S. Yu, “Unbounded and Efficient Revocable Attribute-Based Encryption with Adaptive Security for Cloud-Assisted Internet of Things,” IEEE Internet of Things Journal, vol. 9, no. 4, pp. 3097-3111, 2022.

[10] H. Xiong, T. Yao, H. Wang, J. Feng, and S. Yu, “A Survey of Public-Key Encryption with Search Functionality for Cloud-Assisted IoT,” IEEE Internet of Things Journal, vol. 9, no. 1, pp. 401-418, 2022.

[11] Y. A. Fatimah, K. Govindan, R. Murniningsih, and A. Setiawan, “Industry 4.0 based sustainable circular economy approach for smart waste management system to achieve sustainable development goals: A case study of Indonesia,” Journal of Cleaner Production, vol. 269, 122263, 2020.

[12] Y.-J. Chen, W.-C. Liou, Y.-M. Chen, and J.-H. Wu, “Fraud detection for financial statements of business groups,” International Journal of Accounting Information Systems, vol. 32, pp. 1-23, 2019.

[13] B. Subramani, and M. Veluchamy, ”MRI brain image enhancement using brightness preserving adaptive fuzzy histogram equalization. ” International Journal of Imaging Systems and Technology, vol. 28, no. 3, pp. 217-222, 2018.

[14] Z. Zhuang, N. Lei, A. N. Joseph Raj, and S. Qiu, ”Application of fractal theory and fuzzy enhancement in ultrasound image segmentation. ” Medical & Biological Engineering & Computing, vol. 57, pp. 623-632, 2019.

[15] K. Mayathevar, M. Veluchamy, and B. Subramani, ”Fuzzy color histogram equalization with weighted distribution for image enhancement,” Optik, vol. 216, 164927, 2020.

[16] Y.-H. Huang, and D.-W. Chen, ”Image fuzzy enhancement algorithm based on contourlet transform domain,” Multimedia Tools and Applications, vol. 79, no. 47-48, pp. 35017-35032, 2020.

[17] A. K. Bhandari, S. Shahnawazuddin, and A. K. Meena, ”A novel fuzzy clustering-based histogram model for image contrast enhancement. ” IEEE Transactions on Fuzzy Systems, vol. 28, no. 9, pp. 2009-2021, 2019.

[18] H. G. Daway, E. G. Daway, and H. H. Kareem, ”Colour image enhancement by fuzzy logic based on sigmoid membership function. ” International Journal of Intelligent Engineering and Systems, vol. 13, no. 5, pp. 238-246, 2020.

[19] M. Liu, Z. Zhou, P. Shang, and D. Xu, ”Fuzzified image enhancement for deep learning in iris recognition,” IEEE Transactions on Fuzzy Systems, vol. 28, no. 1, pp. 92-99, 2019.

[20] J. Arnal, M. Chillarón, E. Parcero, L. B. Súcar, and V. Vidal, ”A parallel fuzzy algorithm for real-time medical image enhancement. ” International Journal of Fuzzy Systems, vol. 22, pp. 2599-2612, 2020.

[21] S. Mandal, S. Mitra, and B. U. Shankar, ”FuzzyCIE: fuzzy colour image enhancement for low-exposure images,” Soft Computing, vol. 24, no. 3, pp. 2151-2167, 2020.

[22] R. Chandrasekharan, and M. Sasikumar, ”Fuzzy transform for contrast enhancement of nonuniform illumination images,” IEEE Signal Processing Letters, vol. 25, no. 6, pp. 813-817, 2018.

[23] M. Veluchamy, and B. Subramani, ”Fuzzy dissimilarity color histogram equalization for contrast enhancement and color correction. ” Applied Soft Computing, vol. 89, 106077, 2020.

[24] A. Slowik, and H. Kwasnicka, ”Evolutionary algorithms and their applications to engineering problems,” Neural Computing and Applications, vol. 32, pp. 12363-12379, 2020.

[25] K. Li, R. Chen, G. Fu, and X. Yao, ”Two-archive evolutionary algorithm for constrained multiobjective optimization,” IEEE Transactions on Evolutionary Computation, vol. 23, no. 2, pp. 303-315, 2018.

[26] C. Engels, K. Kumar, and D. Philip, “Financial literacy and fraud detection,” The European Journal of Finance, vol. 26, no. 4-5, pp. 420-442, 2020.

[27] L. Yang, “Risk prediction algorithm of social security fund operation based on RBF neural network,” International Journal of Antennas and Propagation, vol. 2021, pp. 1-8, 2021.

[28] J. Jaramillo, J. D. Velasquez, and C. J. Franco, “Research in financial time series forecasting with SVM: Contributions from literature,” IEEE Latin America Transactions, vol. 15, no. 1, pp. 145-153, 2017.

[29] J. Uthayakumar, N. Metawa, K. Shankar, and S. Lakshmanaprabu, “Financial crisis prediction model using ant colony optimization,” International Journal of Information Management, vol. 50, pp. 538-556, 2020.

[30] H. Zhou, G. Sun, S. Fu, J. Liu, X. Zhou, and J. Zhou, “A big data mining approach of PSO-based BP neural network for financial risk management with IoT,” IEEE Access, vol. 7, pp. 154035-154043, 2019.

[31] S. Ding, J. Yu, B. Qi, and H. Huang, “An overview on twin support vector machines,” Artificial Intelligence Review, vol. 42, pp. 245-252, 2014.

[32] M. A. Kumar, and M. Gopal, “Least squares twin support vector machines for pattern classification,” Expert Systems with Applications, vol. 36, no. 4, pp. 7535-7543, 2009.

[33] Y. Fan, and Z. Sun, “CPI big data prediction based on wavelet twin support vector machine,” International Journal of Pattern Recognition and Artificial Intelligence, vol. 35, no. 04, 2159013, 2021.

[34] Y. Du, ”Vision system of apple picking robot based on twin support vector machine,” in Journal of Physics: Conference Series. IOP Publishing, 2021, pp. 336-348.

[35] Y. Guo, X. Jiang, L. Tao, L. Meng, C. Dai, X. Long, F. Wan, Y. Zhang, J. Van Dijk, and R. M. Aarts, “Epileptic seizure detection by cascading isolation forest-based anomaly screening and EasyEnsemble,” IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 30, pp. 915-924, 2022.