Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 6922554, 11 pages
https://doi.org/10.1155/2022/6922554

Research Article

Integration of SOM and PCA for Analyzing Sports Economic Data and Designing a Management System

Lijing Cao

Department of Sports, Tianjin Vocational Institute, Tianjin 300410, China

Correspondence should be addressed to Lijing Cao; 000697@tjtc.edu.cn

Received 30 March 2022; Revised 25 April 2022; Accepted 30 April 2022; Published 19 May 2022

Academic Editor: Wen-Tsao Pan

Currently, sports economic data have attracted more attention because they normally exist with a high dimensionality manner that reflects the historical behavior and the potential decision trend of users or players. Traditional analysis techniques dealing with such kinds of data rely heavily on the empirical knowledge of the manager. With the development of data science, traditional experience-based knowledge barely meets the requirements of multiple features and high-dimensional data analysis. In this regard, machine learning-based data analysis techniques nowadays can give more importance to the process of extracting latent information hidden in chaotic data, which helps users make decisions and take appropriate actions. In this paper, we integrate principal component analysis (PCA) and a self-organizing map (SOM) to exploit the hidden features in the high-dimensional data. Specifically, PCA considers an orthogonal transformation operation to linearly transform the observed data to the low-dimensional one. SOM clusters the data by constructing a two-layered neural network without manual intervention and knowing the category in the training stage. The integration of PCA and SOM helps promote the research on pattern recognition and visualization of high-dimensional data. The experimental results obtained from economic data indicate the effectiveness of the combination of PCA and SOM.

1. Introduction

The main purpose of data analysis is to exploit the information latent, which normally originates from chaotic data [1, 2]. In real applications, data analysis enables users to make appropriate decisions and take effective responses, which first involves user data and then analyzes data to capture more useful information [3, 4]. It is treated as the basic procedure of the management system for quality assessment. In the entire procedure of product generation, the data analysis process should be properly and technically governed, which is a crucial way to improve the effectiveness of decision-making. A representative example can be found in the process of designation that, prior to initializing a new proposal, designers and other staff will analyze the users' habitats obtained from investigations and surveys to select the appropriate line of design. In this regard, the data analysis process plays a crucial and indispensable part in our current society [5].

The development of big data triggers multiple lines of application and academic research [6, 7]. Among these lines, sports and economic data become two major parts of data analysis. The former one involves the process that utilizes sports-related data to explore useful features and transmit them to the management staff or other decision-makers in the form of graphics, reports, and so on [8, 9]. The representative features behind the sports data include athlete statistics, media contracts, ticket and commodity sales, and license agreements. By collecting and analyzing such data, the results obtained from the analyzing process can provide competitive advantages for teams or individuals. The latter involves exploiting the actual economic data in the past or now and finding the potential anomalous transactions, and proposing new economic strategies [10, 11]. Such economic data usually appear in the form of time series, that is, cross-sectional data covering multiple periods or one time period, or also are captured from surveys of individuals and companies with high dimensions. Specifically, they can be

normally expressed in nominal or real values, including various alternative indicators of output, orders, trade, labor, confidence, prices, and financial series.

With the development of technology in the past several decades, data can be obtained in a relatively easy behavior, leading to the fast development of artificial intelligence or machine learning methods [12, 13]. Machine learning is an important data analysis tool that has been considered in different data processing fields. By acquiring loads of training samples, machine learning methods can be effectively optimized and trained to exploit intrinsic data representations and users' historic habits for the purpose of predicting future trends. Compared to empirical knowledge-based traditional data analysis strategies, machine learning-based methods enable researchers to explore latent information within the observed dataset by using statistical and optimization techniques. At present, machine learning methods are categorized as three main directions, i.e., unsupervised learning, supervised learning [14, 15], and semisupervised learning. Unsupervised learning, also known as the clustering method, does not need training samples but only interprets the data by exploring the structure and correlation information between the input data, including K-means [16], ISODATA [17], DBSCAN [18], Fuzzy C-Means [19], etc. Supervised learning needs a group of training samples to train the model, which has obtained the optimal model parameters, and then transplants the trained model to the test samples to observe the behaviors of test samples, including sparse/collaborative representation [20], ensemble learning [21], support vector machine [22]. Unlike unsupervised and supervised learning, semisupervised learning introduces some unlabeled samples into the training process for the sake of improving the robustness of method.

Biological research shows that the organizing principle of neurons is orderly arranged in the sensory channel of the human brain [23]. When the external specific spatiotemporal information is captured by the brain, its specific regions will be activated. In this case, if similar information is obtained by the brain, they will be mapped to the same region. For example, if some patterns simultaneously simulate several receiving neurons in the retina, some neurons in the cerebral cortex will be subsequently excited. More specifically, such response phenomena are not innate but formed by learning and self-organization manners [24]. In this regard, self-organizing map (SOM) based neural networks show specific learning modes and network structures [25]. The network structure of SOM normally contains two layers in which, the first is the input layer, and the second is the competition layer. Each neuron uses a two-way connection to link two network layers without considering hidden layers. Sometimes there are horizontally connected two neuron neighbors in the competitive layer. Unlike multilayer neural networks, this manner simulates the dynamic principle of information processing of excitation, coordination, inhibition, and competition between biological neurons. The intuitive motivation of SOM is that each neuron in the network competition layer is competing to maximize the opportunity of responding to the input information. Only one neuron will be treated as the winner in the competition process, which determines the final classification label of the input mode. Currently, SOM has been widely discussed in another research field such as environmental monitoring [26, 27], medicine analysis [28], accident analysis [29], speech recognition [30], and so on.

Owing to the main fact that the collected data nowadays exist with high dimensionalities, the most important pretreatment step of analyzing latent information is to reduce the dimension of the data on the premise of ensuring the essence of the data as much as possible. Dimensionality reduction is an important preprocessing technology within the field of data analysis, which is often used prior to applying classification algorithms to the dataset. The main reason for using dimensionality reduction methods is that this process can remove some redundant information and noise from the data to improve the data processing speed and reduce time and cost burdens. Currently, principal components analysis (PCA) [31] aims to reduce the information redundancy and the dimensionality of data, which is a widely used dimensionality reduction technique in the field of statistical analysis such as hyperspectral data analysis [32, 33], medical image processing [34], and so on. It uses an orthogonal transformation to linearly transform the observed related variables to linearly unrelated variables. Specifically, the principal component can be regarded as a linear equation, which contains a series of linear coefficients to indicate the projection direction. In this regard, a small number of principal components can be used to represent the main characteristics of the original data.

This paper discusses the qualitative identification of sports and economic-based data by the SOM method and discusses the characteristics of the main influencing factors and explanatory factors of sports and economic-based data by the PCA method. Under these circumstances of the big data age, SOM and PCA methods will help to promote the research on pattern recognition and visualization of sports and economic-based data. At the same time, the integration of SOM and PCA has certain scientific significance and practical values for the research on the data analysis community. Additionally, this paper also provides the workflow for designing a management system for analyzing sports or economic data. The major contributions of this paper are twofold. First, this paper joints PCA and SOM methods to analyze sports economic data, where the former is used to reduce the dimensionality of original data, and the latter is utilized to analyze dimensionality-reduced data for the purpose of learning different features. Second, this paper provides the details of the construction process of designing a management system for analyzing sports economic data.

This paper is divided into six sections, which can be given as follows. Section 3 introduces the algorithmic structures of SOM and PCA in detail. Section 4 provides the workflow of

the data management system. Section 5 gives the experimental results to verify the effectiveness of the combination of SOM and PCA. Section 6 concludes with some remarks.

2. Proposed SOM-PCA Model

2.1. SOM for Pattern Recognition and Analysis. Two layers are normally considered in SOM network including the input layer and the competition layer. The first aims to simulate the process of sensing the input information. The second aims to simulate the response behavior of the cerebral cortex. The impact of the winning neuron of the SOM network on its adjacent neurons depends mainly on the distance. Therefore, in the learning process, not only the winning neuron itself should adjust the weight vector, but also the surrounding neurons should follow the same manner to varying intensities under its impact. Common adjustment methods include the Mexican straw hat function, top-hat function, and chef-hat function [27], where top-hat and chef-hat functions are two special cases of Mexican straw hat function. For the Mexican straw hat function, the winning node has the largest amount of weight adjustment, and the adjacent node has a relatively small amount of adjustment. When the distance is far, the weight adjustment is negative. On the contrary, the greater the distance from the winning node, the smaller the weight adjustment intensity. Figure 1 gives a visual description of the Mexican straw hat function. In this figure, R represents a neighborhood radius centered on the winning neuron that has the highest weight a. In the SOM learning algorithm, all neurons in the winning neighborhood adjust their weights according to their distance from the winning neuron.

SOM network can form the characteristic topological distribution of input signal on a one-dimensional or two-dimensional processing unit array behavior. To visually display the structure of SOM, Figure 2 demonstrates a two-layered SOM network. As can be seen from Figure 2, the network is composed of the input layer and output layer, in which the output layer is also called as a competitive layer. It is worth mentioning that the number of neurons in the input layer is determined by the number of vectors in the competitive layer. A one-dimensional vector is used to represent the input neurons for the sake of receiving the input signal, and a two-dimensional matrix is arranged in the output layer to provide responses. Additionally, the neurons in different layers are connected using weights.

The process of SOM has six steps:

(1) Allocate relatively small random values to each weight vector of the output layer and perform a normalization process. Let $w_i\,(i = 1, 2, \ldots, m)$, $\phi_{i(0)}$, and η, respectively, be the weight vector, initial active neighborhood, and learning rate, where m denotes the number of neurons in the output layer;

(2) Select a feature from the training dataset and perform a normalization process. Let $Z^p\,(p = 1, 2, \ldots, n)$ be the input data, where n represents the number of neurons in the input layer;

(3) Calculate the Euclidean distance between w_i and Z^p for the sake of finding active node that has the largest distance, expressed as follows:

$$d_j = \left\| Z - w_i \right\|_2^2, \tag{1}$$

where $\|\cdot\|_2^2$ represents the L_2 norm.

(4) Define an active neighborhood $\phi_i^*\,(t)$. Normally, the initial $\phi_{i(0)}$ is relatively large, and it will be shrunk during the training;

(5) Adjust the weights of all nodes in the active neighborhood $\phi_i^*\,(t)$, expressed as follows:

$$w_{ij}(t+1) \leftarrow w_{ij}(t) + \alpha(t, \phi)\left[Z_i^p - w_{ij}(t)\right], \quad i = 1, 2, \ldots, n\; j \in \phi_j^*(t), \tag{2}$$

where $w_{ij}\,(t)$ denotes the weights of i-th neuron in t moment, $\alpha(t, \phi)$ denotes a function indicating the training time and distance between j-th neuron in the neighborhood and active neuron j^*.

(6) If the learning rate is lower than a predetermined threshold, the training step will stop; otherwise, go to step 2).

A brief workflow of SOM can be seen in Figure 3.

2.2. PCA for Data Dimensionality Reduction and Component Recognition. PCA is one of the ten classical machine learning algorithms, which is a multivariate statistical dimensionality-reduction method proposed by Pearson in 1901 and subsequently developed by Hotelling in 1933.

Generally, for the given data with high dimensionalities, the most important pretreatment step of analyzing latent information is to reduce the dimension of the data on the premise of ensuring the essence of the data as much as possible. Dimensionality reduction is an important preprocessing technology within the field of data analysis, which is often used prior to applying to another algorithm because this process can remove some redundant information and noise from the data and make the data clearer and more efficient so as to improve the data processing speed and reduce time and cost burdens. Specifically, PCA transforms a set of potentially related variables to that of linearly uncorrelated variables based on an orthogonal transformation operation. Given high-dimensional data $Y \in \mathbb{R}^{L \times N}$, where L and N, respectively, represents the dimensionality of the dataset and the number of samples. PCA involves four main steps to generate principles:

(1) Calculate the mean of data and remove the mean of each dimensionality, i.e., feature:

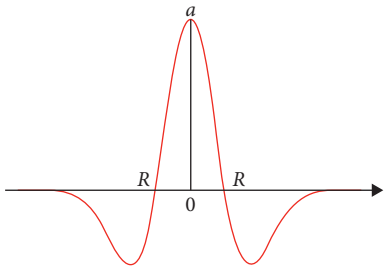

Figure 1: Illustration of Mexican straw hat function.

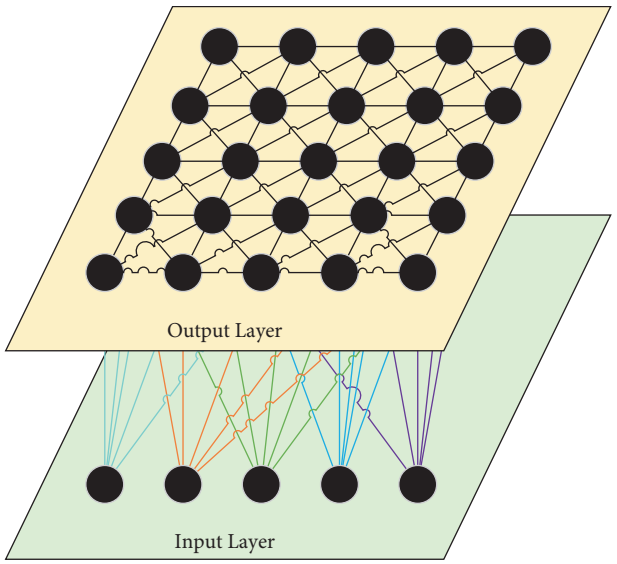

Figure 2: Neural network structure of SOM algorithm.

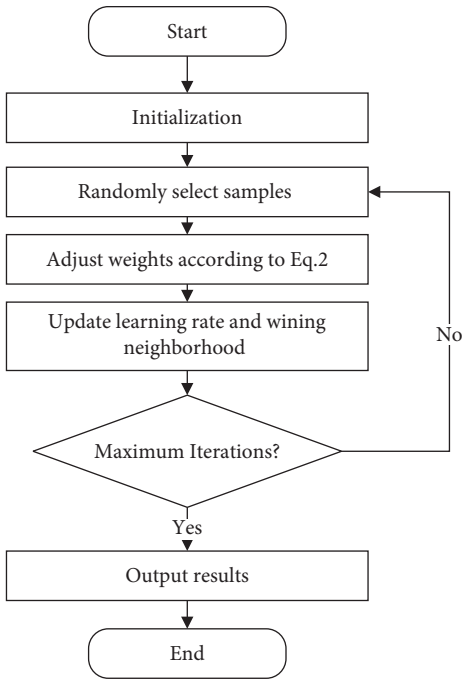

Figure 3: Workflow of SOM algorithms.

$$u = \frac{1}{N}\sum_{i=1}^{N} Y_i, \quad Y \leftarrow Y - u1_N^T, \tag{3}$$

where 1_N represents an all-ones vector with N entities and $(\cdot)^T$ represents a transpose operator.

(2) Calculate the covariance matrix:

$$C = \frac{1}{N} YY^T. \tag{4}$$

(3) Calculate the eigenvalues and corresponding eigenvectors using the singular value decomposition method:

$$[V, \sum] \leftarrow sv\ ds(C), \tag{5}$$

where V, $\sum$, U, respectively, represents the left eigenvector matrix, singular value matrix, and the right eigenvector matrix. It is worth mentioning that p represents the desirable dimensionality, which is normally determined according to the number of features in a original given data.

(4) Project the original data Y into the matrix of eigenvectors to capture the low dimensionality subspace:

$$X \leftarrow V^T Y. \tag{6}$$

It is worth noting that PCA decomposes the covariance matrix of metadata into eigenvalues to obtain a set of bases for projecting the metadata into low-dimensional space. The covariance matrix can reduce the correlation of samples and facilitate the analysis of the internal characteristics of data. SVD pays more attention to the original matrix. More specifically, SVD is the best low rank matrix estimation at the Frobenius norm level, and it is also an estimation of data in the low dimensional hyperplane.

3. Designation of the Management System for Data Analysis

3.1. System Composition and Requirement. The analysis and management of sports or economic data cannot be limited to the scattered data itself. Normally, the system needs to integrate the elements of product supporting management, development stage management, data management, data processing, data results analysis, and also provides an evaluation of the system from the perspective of the system and the requirements of the development process, so as to provide decision-making basis for policy design and product development. This is the basic design goal of the data analysis and management system. The system structure is shown in Figure 4.

The data analysis and management system shall meet the following requirements:

(1) Data management: data management includes data input and sorting, retrieval and browsing,

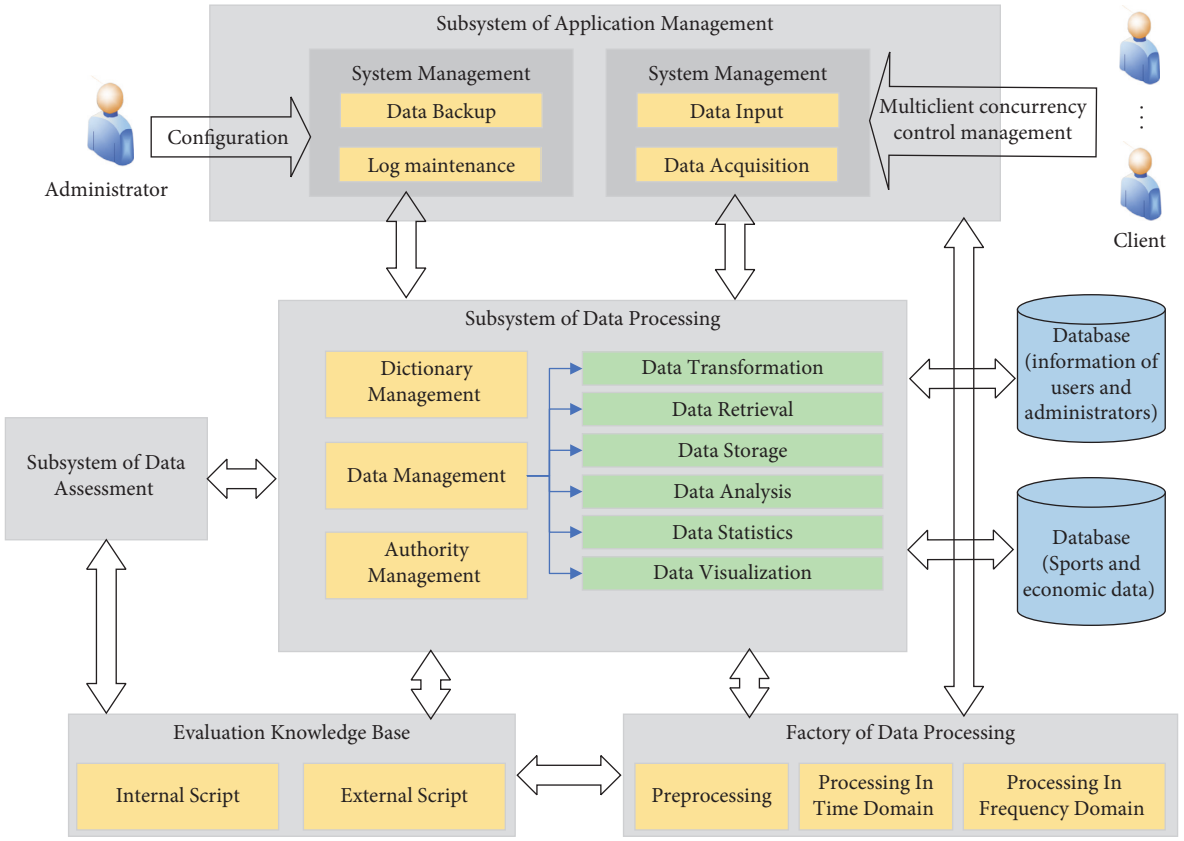

Figure 4: Diagram of system composition structure.

downloading, import and export, statistics, printing, backup, etc.

(2) Data processing: the system shall provide a platform for processing data with a built-in general processing function and provide a data interface with other processing software for special processing. The built-in processing functions include preprocessing, time-domain processing, frequency-domain processing, time-frequency domain analysis, statistical analysis, batch processing, and other specialized processing.

(3) Data analysis and evaluation: generally, users establish and improve the expert knowledge base of data analysis and evaluation, provide the reasoning mechanism model of analysis and evaluation, realize the automation of test data analysis and evaluation, and provide the basis for supporting decision-making and design improvement of test pieces.

(4) Basic data interface: in recent years, sports and economics-based basic data platform has been widely used. All units have developed various professional information management systems around the basic platform. In order to fully and conveniently utilize and share the value of data, the data analysis and management system shall have a data interface with the basic platform and professional system.

(5) Procedure management: procedure management is the premise of ensuring accurate data entry, correct processing, and accurate utilization. The procedure involves data warehousing, processing, evaluation, supporting, and downloading.

3.2. Technology Line of System Designation

3.2.1. Framework of Application Development. The application system with vitality needs to possess good scalability and easily adapt to the changing application requirements at a negligible cost. Model-view-controller (MVC) software hierarchy is a mature software architecture design pattern that has been widely used at present. Model (M) represents the data object and its business processing. View (V) represents the user manipulated display interface. The controller (C) provides process control and plays a connecting role between the model and the view. It is responsible for converting the user's request into a call to the model and calling the corresponding view to display the data. MVC structure forces the input, processing, and output of the application program to be separated, which improves the maintainability, scalability, flexibility, and encapsulation of the software. The diagram details can be found in Figure 5.

3.2.2. Designation of Software Architecture. For the development of an information management system, there are two main software design modes, i.e., C/S mode and B/S mode. In B/S mode, data storage and most business processing are run on the server side, and the application

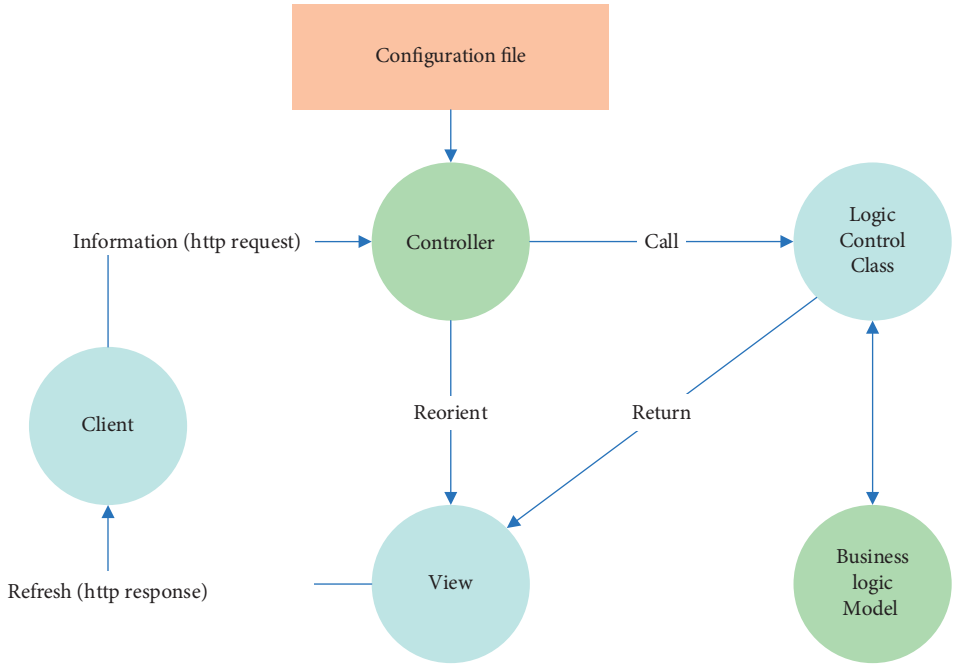

Figure 5: MVC diagram.

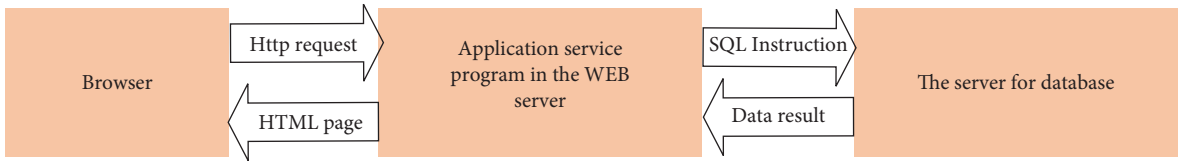

Figure 6: Diagram of B/S system.

software is deployed on the server side, so it is easy to upgrade and maintain. The B/S mode adopts a relatively independent multilayer structure. When the software function changes, only the corresponding hierarchical modules need to be modified on the server, and the client does not need to make any changes. B/S mode puts most of the work on the server, which increases the burden on the server. The client uses browser mode, which is not convenient for complex interface operations and processing of information, and the efficiency is lower than that of C/S mode.

The effective management of data and the maximum value utilization of experimental results are the primary problems of data analysis and management systems. B/S mode system architecture is adopted to facilitate the centralized and comprehensive management of test data and the release and utilization of test results by providing users with convenient functions such as data entry, query, retrieval, browsing, and download. For the complex interface operation and processing of data, we can combine the client control technology or integrate the special client analysis and processing software in the page to form a distributed processing system, and the client computer can complete some complex operations and processing, which can also reduce the processing burden of the server and network load. Figure 6 shows the diagram of the B/S system.

3.2.3. Designation of Database. As an important part of computer data processing and information management systems, database technology plays a crucial part in solving the effective organization and data storage, data security, data sharing, data retrieval, and processing, and has become an indispensable part of the organization and management of a large amount of data. After decades of development of database technology, the commonly used commercial general database management systems include Oracle, SQL Server, Ingres, Informix, Sybase, DB2, and so on. Considering the technical development, application scope, follow-up support, compatibility, performance, security and confidentiality, and other factors, Oracle and SQL servers are preferred as database management systems.

3.2.4. Development Framework for Management System. Currently, J2EE and .NET are the two most widely used and discussed frameworks in system development. The major advantage of J2EE is that it can be deployed and applied in cross-platform scenarios. Specifically, Unix, Linux, and Windows Server can be used in the database and application servers under the J2EE situation. For the .NET, both database and application servers can use Windows Server, where the database management system can resort to Oracle series and SQL Server and the application server can use IIS.

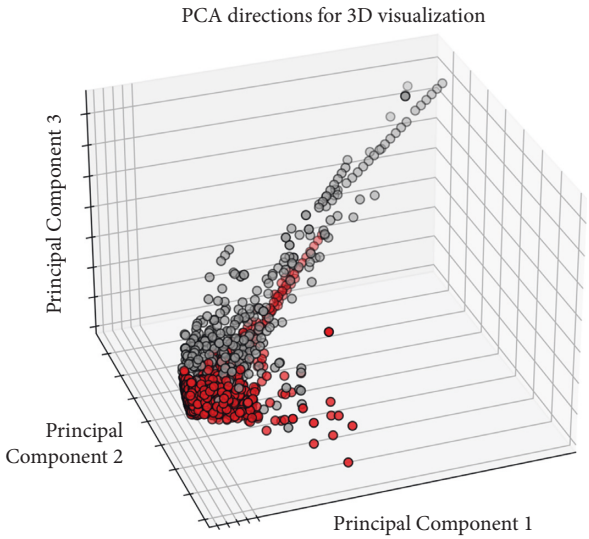

Figure 7: 3D visualization of PCA on the economic data.

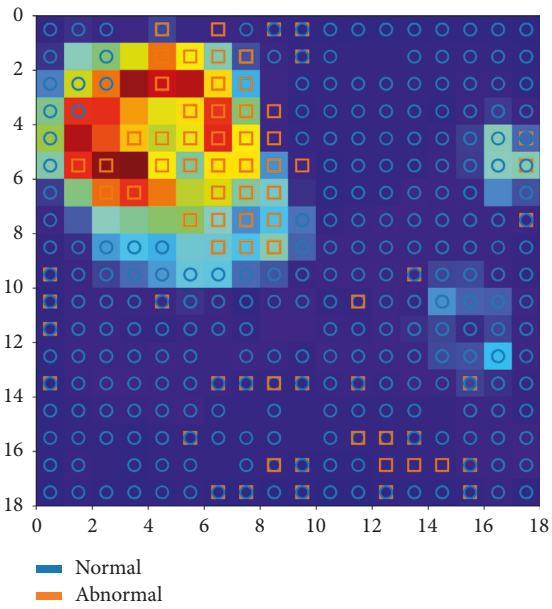

Figure 8: Label map in the output layer.

Compared to J2EE, .NET can achieve higher development efficiency. It is worth mentioning that both two frameworks can be used in the system development steps if they meet the following principals: (1) some complex functions such as data storage, data analysis, data processing, and data management, should be run on the servers; (2) some simple functions such as data retrieval and data visualization should be generated on the server with dynamic page forms and then sent to the client for subsequent operations.

4. Result Analysis and Discussion

4.1. Datasets. This paper considers a credit card fraud data set, which contains the transactions of European cardholders in September 2013 within two days. A total of 284,807 transactions are included, including 492 fraudulent transactions. The data itself is highly random and can be used for training to identify fraud models. Since the size of this dataset is too big, we, therefore, select 5000 samples, including 492 abnormal transactions, for subsequent experiments.

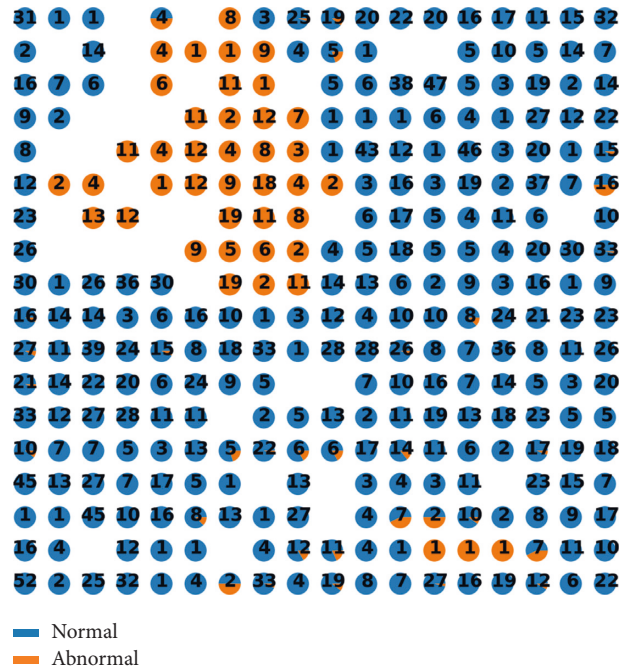

Figure 9: Visualization of pie chart in each grid, where the blue and orange denote the category, and the numbers denote the total number of samples.

4.2. Validation Metric. In order to validate the experimental performances, this paper uses four standard evaluation metrics derived from the confusion matrix for evaluation purposes, including accuracy, precision, recall, and F1 score. The above four evaluation criteria are calculated based on the confusion matrix. For the classification results of a given binary classification model, true positive (TP) is the result that the model correctly predicts the positive class. True negative (TN) denotes the result indicating the model correctly predicts the negative class. False positive (FP) denotes the result that the model incorrectly predicts the positive class. False negative (FN) denotes the result that the model incorrectly predicts the negative class. Accuracy, precision, recall, and F1 score are calculated by the following four formulas:

$$\begin{aligned} \text{Accuracy} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \\ \text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}}, \\ \text{Recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ F1 - \text{score} &= \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}}. \end{aligned} \tag{7}$$

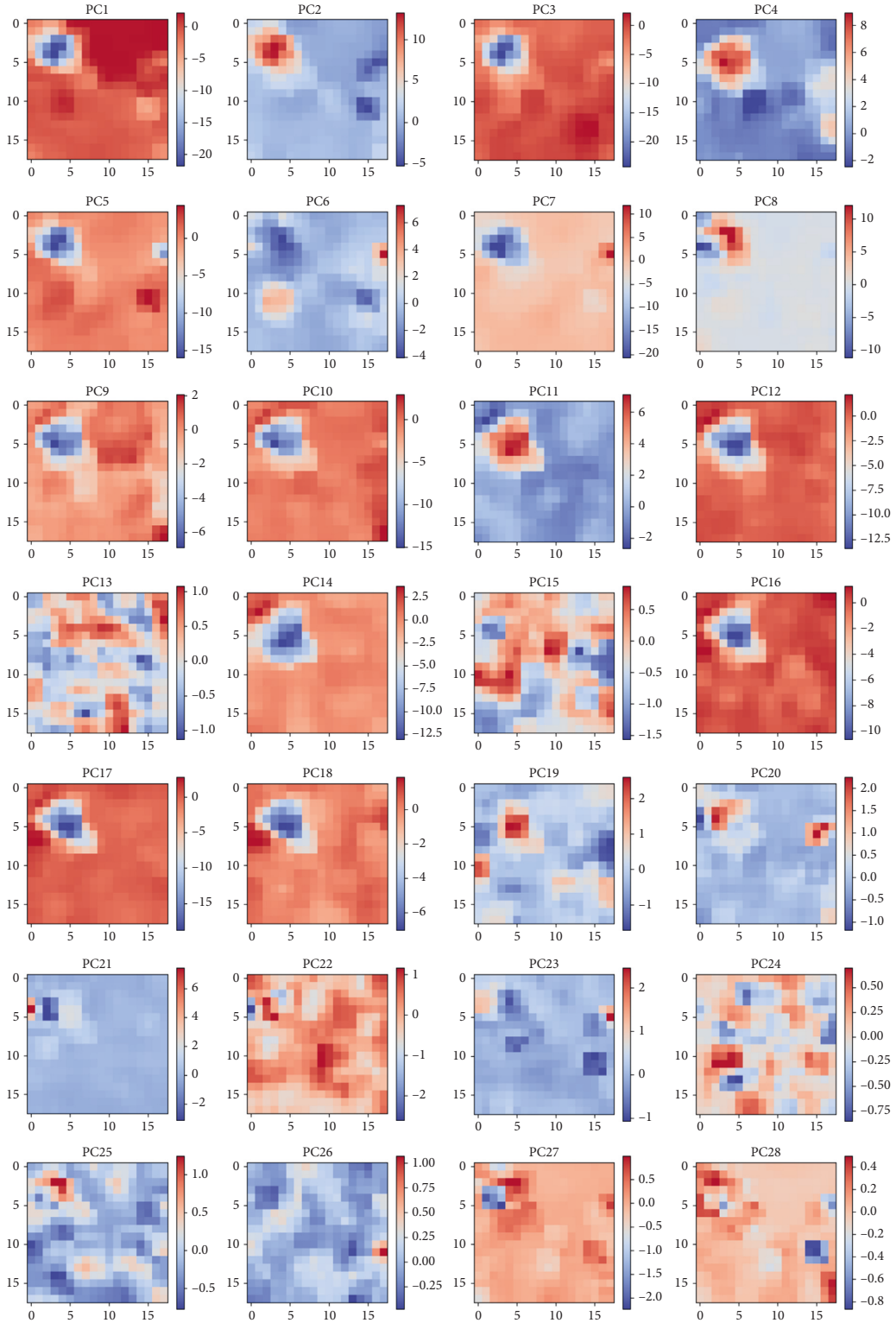

Figure 10: Weighting matrix map in each component plane.

4.3. Experimental Results. Due heavily to the high dimensionality of data, this experiment first uses PCA to decompose original data to the low dimensional space in which the first 28 principal components are retained for subsequent experiments. As can be seen from Figure 7, by plotting the first three major components, the dimensionality-reduced data show low redundancy and clear directions.

During the training process, 70% of total samples are treated as training samples and the rest of 30% of total samples are used for testing purposes. 2000 iterations are

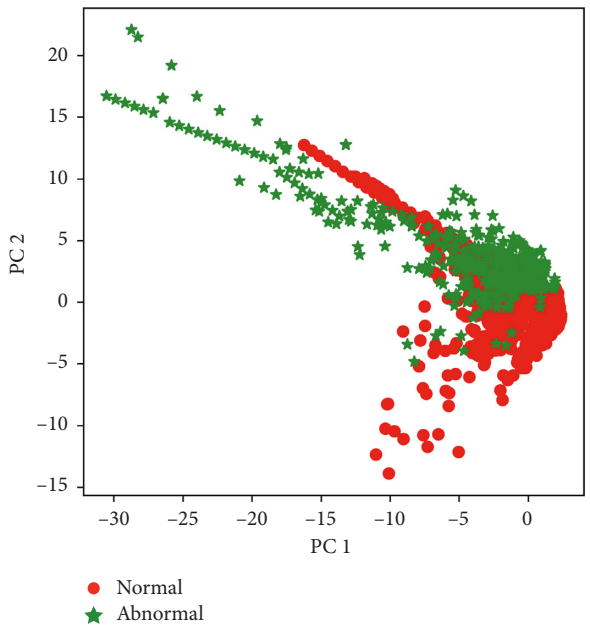

Figure 11: Scatter map of classification results obtained from SOM.

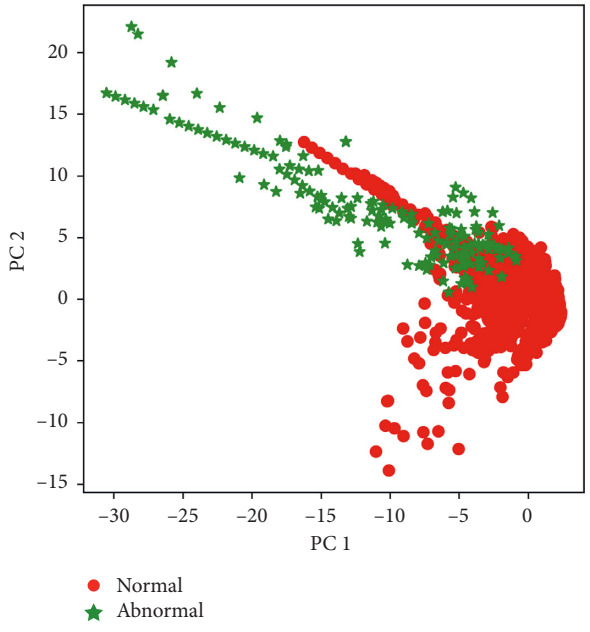

Figure 12: Scatter map of classification results obtained from k-means.

considered to train an optimal SOM. Based on the trained optimal SOM, we can obtain a weighting matrix that calculates the distance between each neuron and its adjacent neurons. Figure 8 provides a visual description of the optimal weighting matrix. As can be seen from Figure 8, it is obvious that there is a deep dividing line on the diagonal, suggesting that the data will be distributed on both sides. The samples corresponding to two categories exist in different positions on the output plane, and this dividing line roughly separates the blue sample from the orange sample.

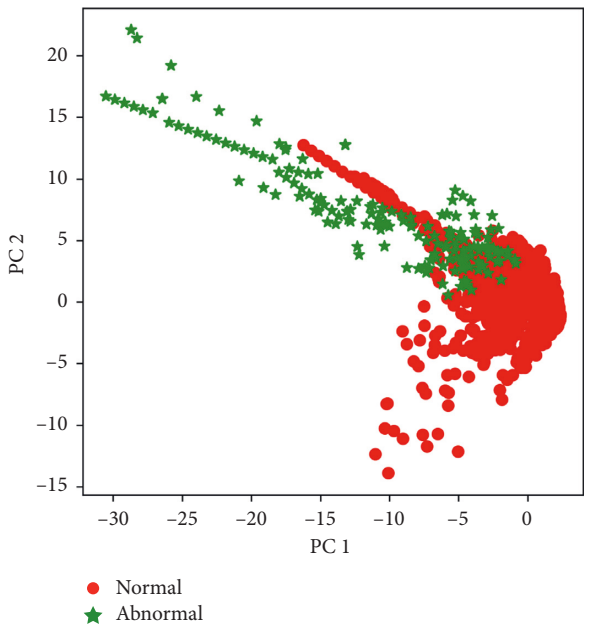

Figure 13: Scatter map of classification results obtained from MiniBatch K means.

In order to clearly observe the number of samples in each grid and the proportion of different samples in each grid, Figure 9 further shows the pie chart of each grid. And the category is represented by color and the total number of samples is represented by numbers. Now we can clearly see how many samples fall into each position, and if there are multiple categories, we can also observe their proportion.

Figure 10 visually displays the weighting matrix corresponding to each principal component. The weighting matrix obtained by the SOM is a matrix with the dimension of $k \times k$, where each dimensionality represents the weight matrix corresponding to the principal component and k represents the optimal side length of output mesh calculated by $\sqrt{5N}$ in which N denotes the number of training samples. In this experiment, k is calculated as 18. The value in the component plane indicates what value the neuron at each location is most sensitive to. It can be seen from the component plane that nearly the entire principal component plays a great role in dividing normal and abnormal transactions.

In order to verify the classification performance, SOM and two classic unsupervised clustering methods, i.e., k-means and its improved version MiniBatch K means, are conducted on entire components produced by PCA to observe the clustering results. Based on two major principal components, Figures 11–13 provide a visual clustering comparison between SOM and k-means. As can be seen from Figure 11, SOM can produce relatively accurate classification results, while k-means and MiniBatchKmeas show distinct results in part of samples [see Figures 12 and 13]. The main reason behind this fact is that SOM involves a training process that uses a competitive learning strategy to gradually optimize the network by relying on the competition between neurons, where the nearest neighbor function

Table 1: Overall results obtained from three methods on the dataset

Metrics	SOM	k-means	MiniBatchKmeans
Accuracy	0.95	0.94	0.94
Precision	**0.98**	0.93	0.94
Recall	**1**	1	1
*F*1-score	**0.99**	0.97	0.97
Time (s)	0.476	0.053	0.049

is used to maintain the topology of the input space. For the k-means, it only defines the class by iteratively calculating the relations between samples and clusters.

Table 1 provides overall results obtained from three algorithms conducted on the dataset. As can be observed from Table 1, compared to *K*-means and MiniBatchKmeans, SOM shows the best experimental results, followed by the MiniBatch *K* means, which show relatively good performances. Execution time (in seconds) is also tabulated in Table 1. Since SOM involves more complex training steps, it requires more clustering time. Among the three methods, MiniBatchKmeans has the lowest computational burden mainly because it has a fast convergence speed compared to *K*-means.

5. Conclusion

With the development of equipment in data acquisition, data can be obtained in a relatively easy behavior with multiple feature dimensions. Compared to other fields of data science, sports and economic-based data have attached more attention because they involve complex and various user-related data. Thanks to the fast development of machine learning, plenty of learnable data analysis methods can be used to exploit latent information of user habitats for the purpose of predicting future decisions. To analyze sports and economic-based data, this paper combines two well-known data science analysis techniques, i.e., principal component analysis (PCA) and self-organizing map (SOM), to exploit the hidden features in the high-dimensional data. Due to the fact that PCA can reduce the dimensionality of data by orthogonal transformation and SOM unsupervised clusters data by establishing a two-layered neural network. The combination of PCA and SOM has the ability to improve the pattern recognition accuracy of high-dimensional and multiple feature economic data. The experimental results show the effectiveness of the combination of PCA and SOM in analyzing economic data. Also, this paper provides the designation process of a management system for analyzing sports and economic data, which contains four main subsystems, including the application management subsystem, data processing subsystem, data assessment subsystem. Besides, the development details of the framework of application development, software architecture, and database designation are also given in this paper.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] R. L. Ott and M. T. Longnecker, *An Introduction to Statistical Methods and Data Analysis*, Cengage Learning, Massachusetts, MS, USA, 2015.

[2] J. Fan, F. Han, and H. Liu, "Challenges of big data analysis," *National Science Review*, vol. 1, no. 2, pp. 293–314, 2014.

[3] S. G. Heeringa, B. T. West, and P. A. Berglund, *Applied Survey Data Analysis*, Chapman and Hall/CRC, London, UK, 2017.

[4] A. Trnka, "Big data analysis," *European Journal of Science and Theology*, vol. 10, no. 1, pp. 143–148, 2014.

[5] N. Zaman, M. E. Seliaman, M. F. Hassan, and F. P. G. Márquez, *Handbook of Research on Trends and Future Directions in Big Data and Web Intelligence*, IGI Global, Hershey, PA, USA, 2015.

[6] A. Alyass, M. Turcotte, and D. Meyre, "From big data analysis to personalized medicine for all: challenges and opportunities," *BMC Medical Genomics*, vol. 8, no. 1, pp. 1–12, 2015.

[7] K. H. Coble, A. K. Mishra, S. Ferrell, and T. Griffin, "Big data in agriculture: a challenge for the future," *Applied Economic Perspectives and Policy*, vol. 40, no. 1, pp. 79–96, 2018.

[8] D. Patel, D. Shah, and M. Shah, "The intertwine of brain and body: a quantitative analysis on how big data influences the system of sports," *Annals of Data Science*, vol. 7, no. 1, pp. 1–16, 2020.

[9] R. Rein and D. Memmert, "Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science," *SpringerPlus*, vol. 5, no. 1, pp. 1410–1413, 2016.

[10] N. I. Didenko, D. F. Skripnuk, and O. V. Mirolyubova, "Big data and the global economy," in *Proceedings of the 2017 Tenth International Conference Management of Large-Scale System Development (MLSD)*, pp. 1–5, IEEE, Russia, Moscow, October 2017.

[11] H. Tao, M. Z. A. Bhuiyan, M. A. Rahman et al., "Economic perspective analysis of protecting big data security and privacy," *Future Generation Computer Systems*, vol. 98, pp. 660–671, 2019.

[12] J. Qiu, Q. Wu, G. Ding, Y. Xu, and S. Feng, "A survey of machine learning for big data processing," *EURASIP Journal on Applied Signal Processing*, vol. 2016, no. 1, pp. 1–16, 2016.

[13] O. Y. Al-Jarrah, P. D. Yoo, S. Muhaidat, G. K. Karagiannidis, and K. Taha, "Efficient machine learning for big data: a review," *Big Data Research*, vol. 2, no. 3, pp. 87–93, 2015.

[14] L. Zhou, S. Pan, J. Wang, and A. V. Vasilakos, "Machine learning on big data: opportunities and challenges," *Neurocomputing*, vol. 237, pp. 350–361, 2017.

[15] A. L'Heureux, K. Grolinger, H. F. Elyamany, and M. A. M. Capretz, "Machine learning with big data: challenges and approaches," *IEEE Access*, vol. 5, pp. 7776–7797, 2017.

[16] M. Moshkovitz, S. Dasgupta, C. Rashtchian, and N. Frost, "Explainable k-means and k-medians clustering," in *Proceedings of the International conference on Machine Learning PMLR*, pp. 7055–7065, Vienna, Austria, July 2020.

[17] Q. Wang, Q. Li, H. Liu, Y. Wang, and J. Zhu, "An improved ISODATA algorithm for hyperspectral image classification," in *Proceedings of the 2014 7th International Congress on Image and Signal Processing*, pp. 660–664, IEEE, Dalian, China, December 2014.

[18] K. Khan, S. U. Rehman, K. Aziz, S. Fong, and S. Sarasvady, "DBSCAN: past, present and future," in *Proceedings of the The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014)*, pp. 232–238, IEEE, Chennai, India, Febraury 2014.

[19] Y. Ding and X. Fu, "Kernel-based fuzzy c-means clustering algorithm based on genetic algorithm," *Neurocomputing*, vol. 188, pp. 233–238, 2016.

[20] X. Shen, W. Bao, H. Liang, X. Zhang, and X. Ma, "Grouped collaborative representation for hyperspectral image classification using a two-phase strategy," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[21] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, "A survey on ensemble learning," *Frontiers of Computer Science*, vol. 14, no. 2, pp. 241–258, 2020.

[22] M.-W. Huang, C.-W. Chen, W.-C. Lin, S.-W. Ke, and C.-F. Tsai, "SVM and SVM ensembles in breast cancer prediction," *PLoS One*, vol. 12, no. 1, Article ID e0161501, 2017.

[23] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT press, Cambridge, CA, USA, 2016.

[24] D. Miljković, "Brief review of self-organizing maps," in *Proceedings of the 2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 1061–1066, IEEE, Opatija, Croatia, May 2017.

[25] V. Bruni, M. L. Cardinali, and D. Vitulano, "A short review on minimum description length: an application to dimension reduction in PCA," *Entropy*, vol. 24, no. 2, p. 269, 2022.

[26] G. Zhu, X. Wu, J. Ge, F. Liu, W. Zhao, and C. Wu, "Influence of mining activities on groundwater hydrochemistry and heavy metal migration using a self-organizing map (SOM)," *Journal of Cleaner Production*, vol. 257, Article ID 120664, 2020.

[27] S. Clark, S. A. Sisson, and A. Sharma, "Tools for enhancing the application of self-organizing maps in water resources research and engineering," *Advances in Water Resources*, vol. 143, Article ID 103676, 2020.

[28] M. Wang, L. Li, C. Yu et al., "Classification of mixtures of Chinese herbal medicines based on a self-organizing map (SOM)," *Molecular Informatics*, vol. 35, no. 3-4, pp. 109–115, 2016.

[29] S. Kang and Y. Suh, "On the development of risk factor map for accident analysis using textmining and self-organizing map (SOM) algorithms," *Journal of the Korean Surgical Society*, vol. 33, no. 6, pp. 77–84, 2018.

[30] S. Lokesh, P. Malarvizhi Kumar, M. Ramya Devi, P. Parthasarathy, and C. Gokulnath, "An automatic Tamil speech recognition system by using bidirectional recurrent neural network with self-organizing map," *Neural Computing & Applications*, vol. 31, no. 5, pp. 1521–1531, 2019.

[31] D. Blazquez and J. Domenech, "Big Data sources and methods for social and economic analyses," *Technological Forecasting and Social Change*, vol. 130, pp. 99–113, 2018.

[32] X. Shen, W. Bao, K. Qu, and H. Liang, "Superpixel-guided preprocessing algorithm for accelerating hyperspectral endmember extraction based on spatial-spectral analysis," *Journal of Applied Remote Sensing*, vol. 15, no. 02, Article ID 026514, 2021.

[33] X. Kang, X. Xiang, S. Li, and J. A. Benediktsson, "PCA-based edge-preserving features for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 12, pp. 7140–7151, 2017.

[34] J. Reena Benjamin and T. Jayasree, "Improved medical image fusion based on cascaded PCA and shift invariant wavelet transforms," *International Journal of Computer Assisted Radiology and Surgery*, vol. 13, no. 2, pp. 229–240, 2018.