Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 4067554, 13 pages
https://doi.org/10.1155/2022/4067554

Research Article

Correlation Analysis of Network Big Data and Film Time-Series Data Based on Machine Learning Algorithm

Na Li[1] and Langbo Xia[2]

[1]*Department of Information Technology, ZhengZhou Vocational College of Finance and Taxation, Zhengzhou, Henan 450048, China*
[2]*Sichuan University of Media and Communications, Chengdu, Sichuan 611730, China*

Correspondence should be addressed to Langbo Xia; xialangbo@scmc.edu.cn

Received 14 April 2022; Revised 10 June 2022; Accepted 13 June 2022; Published 26 June 2022

Academic Editor: Wen-Tsao Pan

To expand the application of machine learning in movie data, in order to explore the correlation between network big data and film time-series data, based on the machine learning algorithm, the correlation and multifractal characteristics of happiness index (HI) and film box office (BO) were studied and described by introducing multifractal crossover method. On this basis, some indicators are introduced to optimize the neural network model so that the optimization model can describe and predict the box office and other related information well. The results show that the critical values of the happiness index and box office show a linear change trend with the increase of freedom, and the corresponding change curves of the happiness index and box office show obvious nonlinear characteristics, which can be divided into slow increase stage, steady increase stage, and approximately gentle stage. With the increase of iteration parameter q value, the change trend of the long-term and short-term curves of the generalized Hurst function is basically the same, and the difference between the two is getting smaller and smaller, while the difference between the two curves is getting bigger and bigger with the increase of q value of Renyi function. The changing trend of the dynamic Hurst index in the sliding window period all shows that it first rises rapidly to a certain value, then fluctuates rapidly with the increase of time, then drops rapidly to a constant value, and finally continues to show repeated small range fluctuation. Under the influence of time-series parameter α, the original sequence changes the most, the replacement sequence changes the medium, and the corresponding rearrangement sequence changes the least. The overall distribution of box office prediction data conforms to the characteristics of linear variation. The prediction index of the optimized HI-LSTM (Happiness Index-Long term short term memory neural network) model is higher in the box office, indicating that the model has better performance in describing and predicting the box office. This study can provide a theoretical basis for the correlation study of network big data and film data.

1. Introduction

This paper mainly addresses machine learning algorithms in computing network big data and movie data. Correlation algorithms are widely used in film analysis and prediction: a good film analysis model was proposed in [1]. The film analysis model can classify and analyze film reviews so that readers can obtain relevant film information more accurately. A neural network algorithm can be used to optimize the model to make it more effective and timely push relevant information. There is a large amount of film-related data on the Internet, and a model that can analyze film data was constructed in [2]. The model will analyze the sources of various movie data, reduce it to specific movie indicators, and analyze the indicators to predict the relevant content of the movie. The model uses different algorithms, such as artificial intelligence and machine learning, to predict box office earnings by monitoring and analyzing indicators with different characteristics. The number of movie audiences is the primary factor for the sales of the film industry. The existing models are mainly used to predict the relevant properties of the film market, and an optimization model based on the analysis of the number of movie audiences was proposed in [3]. This model indirectly describes and

represents the relevant content of the film market by analyzing indicators such as the number of moviegoers and the score of movie reviews. This method can avoid the correlation error caused by the influence of the film market itself. In order to verify the accuracy of the model, several films are used to analyze and verify the optimization model. The research shows that the model has a good performance in the analysis of film time-series data. In order to better research network data with the big movie, related links between five types of machine learning algorithms were used to predict and calculate the corresponding cross-correlation [4] based on public data from the Internet of things. Algorithm analysis shows that for different algorithms for network data associated with the film, the correlation between the calculation result is also different. We should combine with actual needs when selecting an algorithm. Aiming at a series of problems existing in the film recommendation process on the network, a new optimization algorithm based on machine learning is proposed. By optimizing the original model, the optimization model based on machine learning was obtained in [5]. In this model, a content-based filtering method is adopted to provide users with recommended movies with high similarity, and a new integrated learning algorithm is adopted to improve the system's performance. The results show that the system is effective for a film recommendation. In order to solve the problems existing in the process of film investment, a model is established using different machine learning algorithms [6]. The model can predict the box office return of the film according to the data before the release of the film so as to propose a certain investment direction for investors. To verify the accuracy of the model, the model is used to analyze the relevant data, and the results show that the accuracy of the model can reach 85%. The main purpose of the movie recommendation system is to let the computer automatically learn and adjust the related movie activities. However, the quality and accuracy of search results of existing methods are low. To solve this problem, a recommendation model of a cloud platform environment based on a machine learning algorithm was constructed in [7], which can greatly improve the quality and accuracy of movie searches. At the same time, on the basis of the research, the selected features can be analyzed by using a hierarchical clustering algorithm. Then, the trust ranking algorithm is used to sort these movie clusters. Finally, the movie data is evaluated and analyzed according to relevant indicators.

Machine learning algorithms have a wide range of applications in film classification. Aiming at the problems existing in film classification, a classifier model algorithm using feature extraction and feature sorting training was proposed in [8]. The model can examine the emotional expression and classification of a given film review on the scale of negative and positive sentiment analysis. In order to verify the accuracy of the model, the optimization model was compared with the existing learning model by using the experimental verification method. The research shows that the model can describe and analyze the data related to the film well. Relevant studies show that social network has a high influence on the way of film evaluation. Existing studies have a series of problems in film evaluation, and the Newton algorithm and finite memory gradient method are used to solve these problems [9]. Firstly, the model needs to overlap the continuous samples, and the L-BFGS gradient method is used to estimate the relevant parameters in the model, and then the evaluation results are put into the correlation algorithm. Finally, the artificial neural network model is used to evaluate the proposed data. In order to analyze and verify the reliability of this method, the performance of the model is evaluated by using big data. Experimental results show that this optimization model has higher performance in film recommendation. As a relatively popular spiritual entertainment, film entertainment is gradually attracting people's attention. However, with the rapid development of the film industry, the production of films is also increasing year by year. How to quickly and accurately find users' favorite films from the massive film data has become an urgent problem to be solved. A film optimization model based on artificial intelligence and machine learning technology was proposed in [10]. The model is based on computer vision and machine learning technology, and the original model is optimized by introducing relevant algorithms such as big data. A large number of movies were accurately pushed through machine learning, and the accuracy of the model was verified through model analysis and prediction. The film industry is affected by different factors. With the development of machine learning technology, it is very necessary to analyze the factors that affect the film. A mixed feature prediction model based on social media data features was proposed in [11]. The model predicts the quality and content of films by adopting various algorithms so as to achieve the standardized management of the film market. In order to verify the accuracy of the model, the method of comparative analysis is used to extract and analyze the relevant data of the model. The results show that the model is very important to improve the film market order and has a good application in the related evaluation of film data. Massive film resources provide people with rich content, but the resources and information people need are increasingly difficult to find. To solve this problem, a film resource information mining model is constructed by combining the fuzzy neural network algorithm with dynamic data correlation technology [12]. In this model, dynamic data technology is used to preprocess and screen the movie resource information. On the basis of the above processing, the movie information mining model is used to conduct data mining for the related information and indicators of the movie information. The superiority and correctness of the film data mining model are verified by numerical simulation.

The influence of network big data on the film market is shown in Figure 1. The analysis shows that first, the information source of the movie is transmitted to consumers in the form of signals. After arousing consumers' motivation to watch movies, they can obtain relevant movie signals and browse, analyze, read, and discuss the movie signals. Then the feedback of relevant film information is given under the joint action of consumer decision and the film market. Therefore, in order to analyze the relationship between network big data and film timing, this paper adopts a

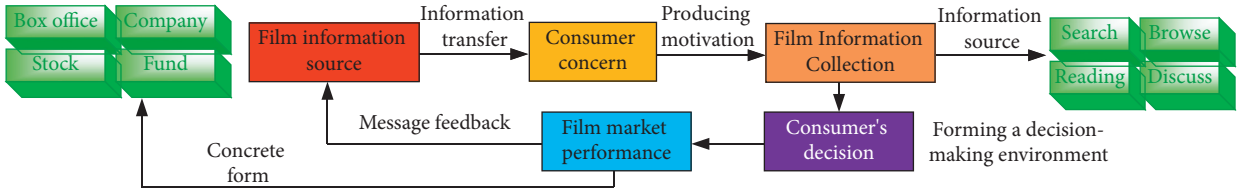

Figure 1: Frame diagram of the influence of network big data on the film market.

machine learning algorithm to conduct correlation analysis on the relationship between HI and BO by introducing the calculation method of multifractal. On this basis, the relevant algorithm is used to optimize the original neural network model so as to obtain the optimized neural network. Through verification, the model can well describe the change relationship between network data and film timing and can further predict and analyze the film market. The results of this study can help film investors understand the relationship between happiness and box office fluctuations and provide some guidance for making corresponding investment portfolio decisions and film schedule selection.

2. Machine Learning

2.1. Multifractal Cross-correlation. In the early study of fractal theory, an analysis method of rescaling range was proposed and widely used in the field of hydrology. In order to better analyze the volatility of relevant data, the multifractal theory is introduced into the theory of detrended fluctuation analysis (DFA), and then the multifractal detrended fluctuation analysis (MF-DFA) is proposed [13, 14]. In order to carry out relevant research on complex data, multifractal detrending cross-correlation analysis (MF-DCCA) was proposed by combining MF-DFA with DCCA (detrending cross-correlation analysis) [15]. The MF-DCCA method can effectively eliminate the influence of local trends on the time-series scale and observe the multifractal of time series at different time scales.

MF-DCCA method firstly inputs the corresponding time-series data of the two columns to obtain the corresponding cumulative deviation and then obtains the corresponding q-order wave function by solving the correlation function of the local covariance [16]. On the basis of the wave function obtained, on the one hand, the Hurst exponential function is optimized by introducing the variable representing the multifractal characteristics. On the other hand, the Hurst function is used to solve the Renyi index, and then the corresponding singular function and multifractal spectrum function are solved. The process of the MF-DCCA method is shown in Figure 2. The relevant judgment basis is as follows: (1) When H is greater than 0.5, it indicates that the model has long range cross-correlation; (2) When H is equal to 0.5, it indicates that the model belongs to a random walk; (3) When H is less than 0.5, it indicates that the model has negative long range cross-correlation.

The basic steps of MF-DCCA method are as follows:

(1) Construct two new time series:

$$X(t) = \sum_{k=1}^{i} [x(k) - \overline{x}];$$
$$Y(t) = \sum_{k=1}^{i} [y(k) - \overline{y}], \tag{1}$$

where $x(t)$ and $y(t)$ are time series, $\overline{x}$ and $\overline{y}$ are the mean values of time series $x(t)$ and $y(t)$.

(2) Divide the newly constructed single time series into $Ns = \text{int}(N/s)$ time windows of length s. Since the time-series length N is not necessarily an integer multiple of the corresponding time scale s, in order to make full use of the whole time series, the same processing is done for the reverse order of the time series. Therefore, we get $2Ns$ nonoverlapping time Windows.

(3) The local trend functions $X_\lambda(i)$ and $Y_\lambda(i)$ are obtained by the least square fitting of the time series within each window λ, and the local covariance function is thus obtained:

$$\begin{cases} X_\lambda(i) = a_k i^m + \cdots + a_1 i + a_0 Y_\lambda(i) = b_k i^m + \cdots + b_1 i + a_0 \\ F^2(s,\lambda) = \dfrac{1}{S}\displaystyle\sum_{i=1}^{s}\{X[(\lambda-1)s+i] - X_\lambda(i)\} \times \{Y[(\lambda-1)s+i] - Y_\lambda(i)\} \\ F^2(s,\lambda) = \dfrac{1}{S}\displaystyle\sum_{i=1}^{s}\{X[N-(\lambda-N_s)s+i] - X_\lambda(i)\} \times \{Y[N-(\lambda-N_s)s+i] - Y_\lambda(i)\} \\ i = 1,2,\ldots,s; \lambda = 1,2,\ldots,2 \quad \text{Ns}; m = 1,2,\ldots,s. \end{cases} \tag{2}$$

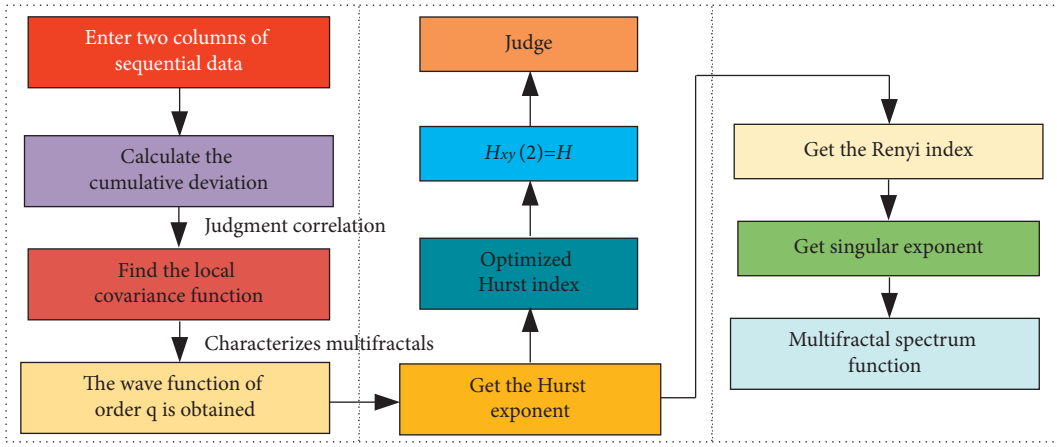

Figure 2: MF-DCCA flow chart.

(4) The q-order wave function $F_q(s)$ was obtained by averaging the local covariance of $2Ns$ windows. The scaling relationship between $F_q(s)$ and scale s is shown as follows:

$$\begin{cases} F_q(s) = \left\{ \frac{1}{2\mathrm{Ns}} \sum_{\lambda=1}^{2N} \left[F^2(s,\lambda) \right]^{q/2} \right\}^{1/q}, & q \neq 0, \\ F_0(s) = \left\{ \frac{1}{4\mathrm{Ns}} \sum_{\lambda=1}^{2N} \ln\left[F^2(s,\lambda) \right] \right\}, & q = 0. \end{cases} \tag{3}$$

When $q = 2$, the MF-DCCA method degenerates into the DCCA method.

$$\log F_q(s) = H_{\mathrm{xy}}(q) \log(s) + \log C. \tag{4}$$

The scale index $H_{xy}(q)$ is called the generalized Hurst index. If $H_{xy}(q)$ depends on q value, it indicates that there is multifractal between two columns of time-series data; otherwise, it is a single fractal.

The cross-correlation coefficient ρ_{DCCA} is used to qualitatively test the degree of cross-correlation between two nonstationary time series, which is defined as follows:

$$\rho_{\mathrm{DCCA}} = \frac{F^2_{\mathrm{DCCA}}(s)}{F_{\mathrm{DFA}\{Xi\}}(s) F_{\mathrm{DFA}\{Yi\}}(s)}. \tag{5}$$

The range of ρ_{DCCA} value is [−1,1], and the corresponding indicator meaning of the specific correlation degree is as follows: (1) When the value of ρ_{DCCA} is 1, the correlation degree is completely positive; (2) When the value of ρ_{DCCA} is 0, there is no correlation; (3) When ρ_{DCCA} value is −1, the correlation degree is completely negative.

In this paper, the method of multicross fractal correlation is used to analyze the data, in which cross-correlation statistics are mainly used to qualitatively analyze whether two-time series have cross-correlation, thus providing support for the definition of cross-correlation function. The cross-correlation statistic $Q_{\mathrm{cc}}(m)$ is defined as follows:

$$Q_{\mathrm{cc}}(m) = N^2 \sum_{i=1}^{m} \frac{X_i^2}{N - i}, \tag{6}$$

where the cross-correlation function X_i is defined as follows:

$$X_i^2 = \frac{\sum_{k=i+1}^{N} x_t y_{t-i}}{\sum_{t=1}^{N} x_t^2 \sum_{t=1}^{N} y_t^2}. \tag{7}$$

2.2. Long and Short Term Memory Model. Long-short term memory is a special RNN model, which is proposed to solve the problem of gradient dispersion in RNN model. In traditional RNN, the training algorithm uses BPTT. When the time is long, the residual to be returned will decline exponentially, resulting in a slow update of network weight. It cannot reflect the long-term memory effect of RNN, so a memory unit is needed to store the memory so as to propose the LSTM model. The number of neurons in the input layer of the Long-Term and Short-Term Memory Model (LSTM) is determined by feature vectors, and the output space depends on the number of neurons in the output layer. A single hidden layer of LSTM is a unit with three gates, as shown in Figure 3. By inputting different data, relevant signals in the model are used for calculation and analysis so as to export the calculated relevant data from the model.

The model switch is controlled by two functions, sig and tanh, and their corresponding equations are as follows:

$$\begin{cases} \mathrm{sig}(x) = \frac{1}{1 + \exp(-x)}, \\ \tanh(x) = \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)}. \end{cases} \tag{8}$$

The corresponding activation function is shown in Figure 4. With the increase of x value, the corresponding y value of the two functions is shown in three stages: (1) Stable stage I, in which y value keeps constant with the increase of x value, which is shown as a horizontal straight line in the

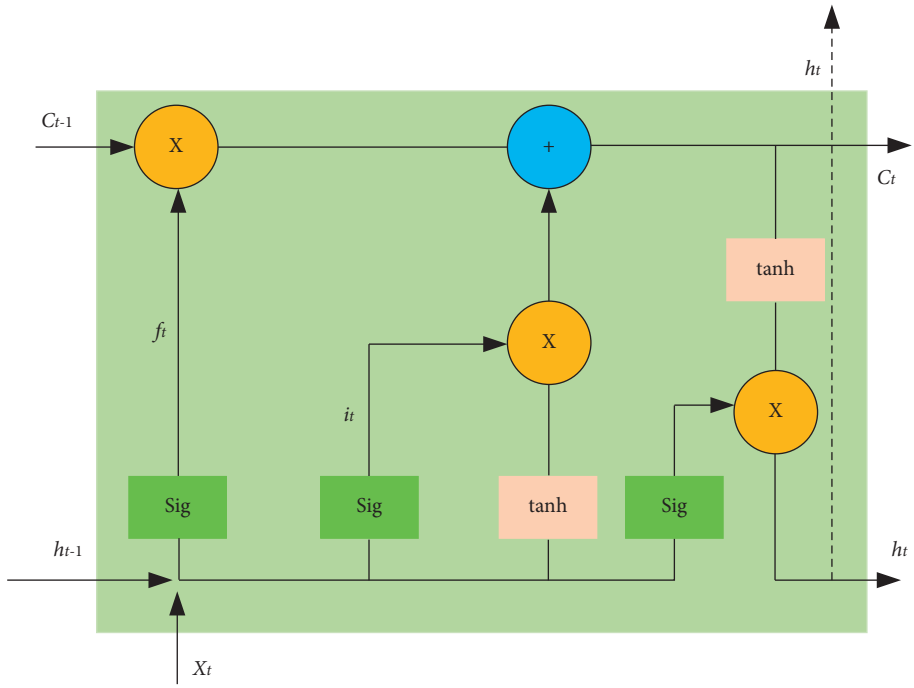

Figure 3: LSTM recurrent neural network frame diagram.

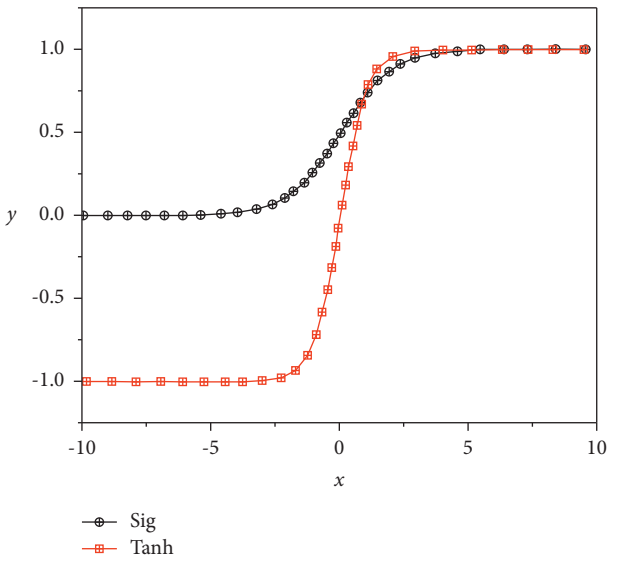

Figure 4: Different kinds of activation functions.

figure. Sig function is higher than tanh function. (2) Rapid increase stage, during which both functions show a trend of rapid increase. The corresponding curve slope increases slowly at first, then keeps constant, and finally gradually approaches 0. With the increase of x value, sig function shows a trend of higher than tanh function at first and lower than tanh function at last. (3) In the stable stage II, the corresponding curves of the two functions are approximately coincident, indicating that the change trend and rule of the two functions in this stage are basically the same. On the whole, the variation range of tanh function is higher than that of sig function, indicating that the description range of tanh function is also higher than that of sig function. X_t and h_t represent the input and output vectors at time t.

LSTM designed three gate switches to effectively control the unit state for long-term memory information. The following details how these three gate switches effectively control information processing.

(1) Calculating f_t of forgetting gate:

$$f_t = \text{sig}\left[(h_{t-1}, X_t) \times W_f + b_f\right]. \tag{9}$$

(2) Determining the stored information:

$$\begin{cases} C_t' = \tanh\left[(h_{t-1}, X_t) \times W_c + b_c\right], \\ i_t = \text{sig}\left[(h_{t-1}, X_t) \times W_i + b_i\right]. \end{cases} \tag{10}$$

(3) Calculate the value of the output gate:

$$\begin{cases} o_t = \text{sig}\left[(h_{t-1}, X_t) \times W_0 + b_0\right], \\ h_t = o_t \tanh\left(C_t\right). \end{cases} \tag{11}$$

Thus, the updated data is as follows:

$$C_t = \text{ft} \times C_{t-1} + C_t' \times i_t, \tag{12}$$

where W_f, W_i, W_c, and W_o represent the weight matrix, b_f, b_i, b_c, and b_o represent bias matrices.

3. Correlation Analysis

3.1. Test Data Acquisition. In order to study the correlation between network big data and film time-series data, it is necessary to carry out a series of data collection aiming at the relevant characteristics of films [17]. In order to more accurately study film time-series data, this paper adopts

relevant data from Facebook social platform for analysis [18, 19]. The data of this platform can further ensure the accuracy of experimental data [20]. In this paper, total national happiness (The index reflects the quality of life and happiness of the people and is made up of four parts: good government governance, economic growth, cultural development, and environmental protection.) is taken as network big data, and the specific index of total national happiness is represented by happiness index. The happiness index of Facebook is summarized in Table 1.

The data collected in this paper includes two parts: HI and BO. The data of the Happiness Index comes from the relevant website of Facebook, while the data of the movie Box Office comes from the website of the movie Box Office. The relevant data for the happiness index are shown in Figure 5.

As can be seen from Figure 5, with the increase of time, the happiness index of Facebook was relatively large from 2010 to 2014, while when in 2015, the corresponding happiness index data was relatively small. From 2015 to 2020, the data are a relatively steady increase, indicating that the selection of experimental data has certain randomness, which can be used as experimental data to carry out relevant research. It shows that the change trend of the happiness index and time is approximately linear. In addition, it can be seen that the overall distribution of the data conforms to the rule of linear distribution by fitting the happiness index with the first-order function.

The corresponding box office data are shown in Figure 6. As can be seen from the box office data, with the increase of time, the distribution of box office data is generally random without obvious regularity. In order to more accurately describe the distribution law of film box office data, the data distribution can be divided into four parts: (1) from 2010 to August 2011, the overall distribution of film box office data is between 2×10^8–10×10^8, with a relatively small range of change, and the corresponding time is relatively small; (2) From August 2011 to November 2012, the box office data showed a large change range in a small period of time, ranging from 1.8×10^8 to 15.5×10^8. From the perspective of time and change range, the box office data at this stage was not stable. (3) From November 2012 to 2015, the data corresponding to the film box office tended to be stable on the whole, with a relatively small change range, and remained around 4×10^8 on the whole. The data at this stage was relatively stable. (4) From 2015 to 2020, at this stage, the film data present a fluctuation change trend that corresponds to the film industry and also shows certain fluctuations; thus, data show that the phase stability is poorer, but the change of the phase for a long time, so you need to film the relevant data for further analysis.

In order to better analyze the relationship between happiness index and movie box office, the relevant data between them are summarized, as shown in Table 2. It can be seen from the statistical table that the difference between the maximum value and minimum value of the happiness index is about 0.53, while the corresponding mean value is 6.13, which is basically in the center of the data, which is basically consistent with the above research. The overall dispersion of film box office data is large and has high volatility, which indicates that film box office has poor stability in a certain period of time.

3.2. Analysis of Test Results

3.2.1. Correlation Analysis of Happiness Index and Box Office. In order to better describe the corresponding relationship between happiness index and film box office, two variables, $Q_{cc}(m)$ [21, 22] and degree of freedom m [23, 24], are used to study the correlation between happiness index and film box office [25, 26]. The logarithmic curve of statistics and degree of freedom is shown in Figure 7. The critical value represents the Chi-square distribution (at a significance level of 5%). Through the correlation analysis between the two variables, it can be seen that the critical values of the two variables show a linear change trend with the increase of the degree of freedom, and the change range is about 3.1. The change curves of the corresponding Happiness Index (HI) and Box Office (BO) show typical nonlinear characteristics, which can be divided into three stages according to their characteristic curves: (1) Slow increase stage. In this stage, the corresponding variation range of Hi-BO data in 0–0.5 degree of freedom is about 0.6–1.1, indicating that the variation range of the curve in this stage is small. At the same time, the slope of the corresponding curve increases gradually with the increase of the degree of freedom, indicating that the statistic has obvious acceleration characteristics in this stage. (2) The steady increase stage, when the degrees of freedom increased from 0.51 to 1.68, the corresponding statistics increased from 1.1 to 8, the phase curve slope, although slight fluctuations, overall remained at a constant data, and the stage of initial value and the critical value corresponding to the curve intersection, with the increase of the degree of freedom, the difference between the two is bigger and bigger, when the degree of freedom reached 1.68, the difference reached the maximum, about 5.8. (3) In the nearly flat stage, with the gradual increase of the degree of freedom, the variation range of the HI-BO curve gradually decreases, and the slope of the curve also gradually declines until it tends to 0, indicating that when the degree of freedom is higher, the variation range of the corresponding statistic is lower, and the degree of influence on the statistic also decreases. It is worth noting that at the beginning and end of this stage, the difference between the two curves is approximately similar, indicating that with the increase of degrees of freedom, the difference between the two curves has little influence.

The long term and short term are described by different time scales, respectively, while the multifractal behavior is measured by the Hurst index and Renyi index. In order to better study the cross-correlation between happiness index and film box office, analyze the long-term and short-term effects on cross-correlation and draw the multidistribution characteristic curves between happiness index and film box office, as shown in Figure 8.

As can be seen from the multifractal characteristic diagram of the happiness index and box office (HI-BO), with the increase of q value (from −10 to 0), the change trend of

Table 1: Facebook's happiness index.

Data	Method	Application area
Stock yield	Correlation coefficient method	International stock market
Share index	Nonlinear test	Stock market
Sunspot number	Linear regression model	Solar motion
Economic search index	MF-DCCA	Monetary market
International equity returns	*T* test	Stock market
US stock index	Vector autoregression	The US stock market
UK stock market index	Nonlinear coefficient	UK equity markets
Volatility index	Linear and nonlinear	Stock market
Movie ranking	Regression model	Film market
Happiness report data	Linear index	Field of life

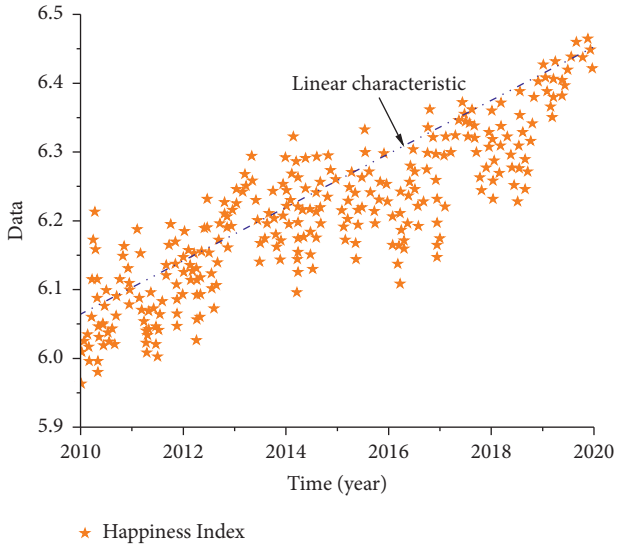

Figure 5: Statistical graph of happiness index from 2010–2020.

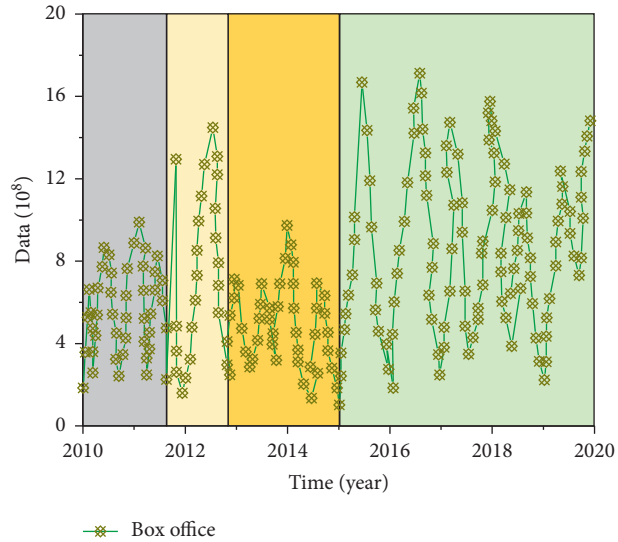

Figure 6: Box office figures for the last decade.

Table 2: Happiness index and movie box office statistics.

Index of correlation	Happiness index	Box office
Sample size	3645	3645
Minimum value	5.96	171230000
Maximum value	6.49	1769540000
Mean value	6.13	1024000420
Median	6.23	1047520042
Standard deviation	0.053	1001467250

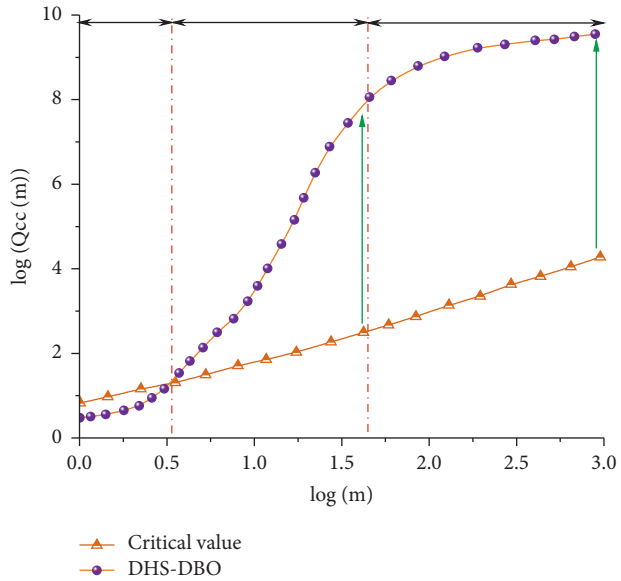

Figure 7: Correlation analysis between happiness index and movie box office.

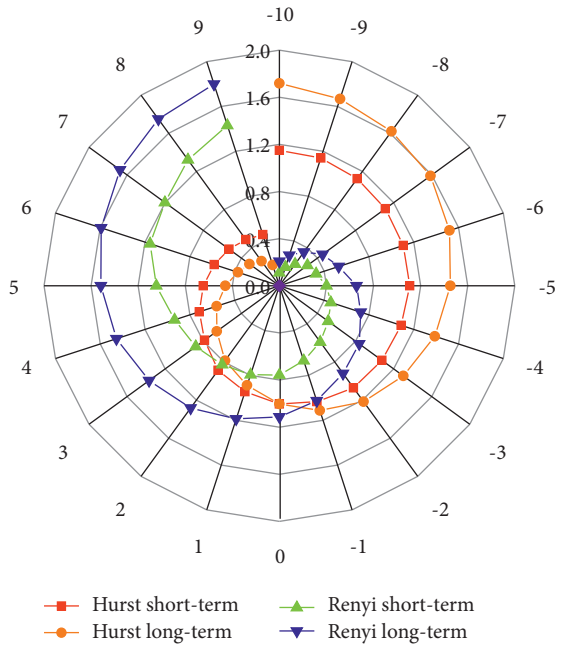

Figure 8: Multifractal characteristics of the happiness index and movie box office.

long-term and short-term curves of the generalized Hurst function is basically the same. However, the difference between them is getting smaller and smaller, indicating that q value can promote the uniformity of the Hurst curve. When q value changes in a positive way, the short-term curve of the generalized Hurst function is higher than the long-term curve of the generalized Hurst function. However, compared with the curve with a negative q value, the Hurst difference of the two corresponding generalized curves when q value is positive is smaller. From the perspective of the Renyi function, with the increase of q value, the difference between the two curves becomes larger and larger, indicating that q value plays a role in promoting the difference between the two curves. The larger q value is, the two curves of Renyi function also gradually increase.

3.2.2. Multifractals of Happiness and Box Office. Relevant studies show that time series will have a certain impact on the multiple analysis characteristics of sample data. In order to explore the change rules of different sequences in detail, multifractal characteristic curves of different sequences are drawn, as shown in Figure 9, and corresponding data statistics are shown in Table 3.

It can be seen from Figure 9 that, with the increase of a value, the three different sequences all show a slow increase to the maximum value and then a slow decline, and the curves show a symmetrical trend of change. Through research, it is found that the three curves approximately conform to the change rule of the quadratic function. The variation range of the original sequence is the smallest, the variation range of the rearranged sequence is the second, and the variation range of the alternative sequence is the largest, indicating that under the influence of independent variables, the influence degree of the alternative sequence is the highest, the influence degree of the rearranged sequence is the lowest, and the corresponding original sequence has the lowest influence degree.

According to the statistics table between the happiness index and the box office, the data varies by sequence: $\alpha_{\max}/\alpha_{\min}/\triangle\alpha$: original sequence >replacement sequence >rearranged sequence. This indicates that under the influence of the influencing factor α, the original sequence has the largest variation range, the replacement sequence has a low variation range, and the corresponding rearrangement sequence has the smallest variation range.

In order to study the dynamic evolution of cross-correlation and local correlation, the influence of external events on power-law cross-correlation is explored by using the sliding window analysis method [27, 28]. The length of the sliding window is fixed at 200 days, and the number of sliding steps is 1 day. The function of calculating the time series of two columns in each window period is the Hurst index, thus obtaining the dynamic change trend diagram of the Hurst index, and the specific change rule is shown in Figure 10.

It can be seen from the changing trend of the dynamic Hurst index during the sliding window period: HI-BO curves show the typical repeated change trend, the change of each phase, are first rapid rise to a certain value, and then increases with the increase of time rapid fluctuations, but the corresponding fluctuation amplitude is small, then quickly fell to a constant value, finally continued to perform as the repeated small scale fluctuation. With the increase of time, the corresponding Q value repeats the above change rule. As can be seen from the maximum value of the curve, the maximum value of the HI-BO curve rises rapidly first and then decreases with the increase of time, and then presents a slow rising change, which lasts for a long time. From the minimum value of the curve, it can be seen that q value drops rapidly at first and then fluctuates with the increase of time, and finally maintains a trend of slow rise.

4. Model Prediction based on Machine Learning

4.1. Model Prediction Theory and Optimization Algorithm. A backpropagation neural network is a machine learning algorithm. Its network structure consists of an input layer, hidden layer, and output layer [29, 30]. For the weights at each synapse, follow these steps to update: (1) The input excitation and response error are multiplied to obtain the gradient of weight; (2) Multiply the gradient by a ratio and invert it and add it to the weight. Each neuron in each layer of the network is a node, and the two layers are connected by a weight coefficient. The corresponding neural network structure is shown in Figure 11, and its main principles and steps in time-series prediction are as follows:

(1) Network initialization

Determine the speed of network learning and neuronal excitation function.

(2) Step 2: Calculate the output of the hidden layer

$$H_j = f\left(\sum_{i=1}^{n} w_{\mathrm{ij}} y_i - a_j\right), \quad j = 1, 2, \ldots, l, \tag{13}$$

where l is the number of nodes in the hidden layer, and f is the excitation function of Sig.

(3) Calculate the output layer

$$\widehat{y_k} = \sum_{j=1}^{l} H_j w_{\mathrm{jk}} - b_k, \quad k = 1, 2, \ldots, m. \tag{14}$$

(4) Calculation error

$$e_k = y_k - \widehat{y_k} \quad k = 1, 2, \ldots, m. \tag{15}$$

(5) Weight update

$$w_{\mathrm{jk}} = w_{\mathrm{jk}} + \eta H_j e_k, \quad j = 1, 2, \ldots, l; k = 1, 2, \ldots, m, \tag{16}$$

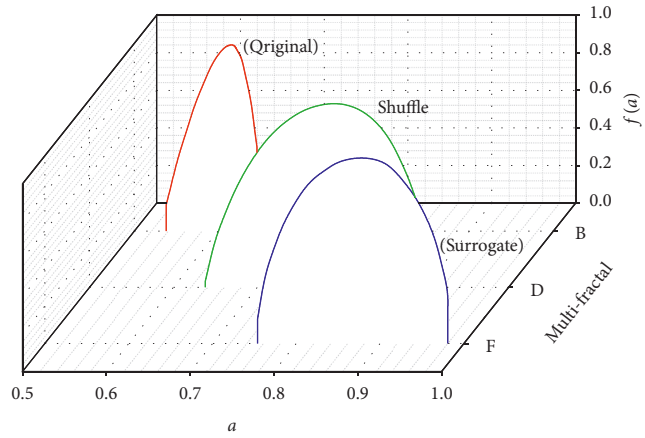

Figure 9: Multifractal features of different sequences.

Table 3: Parameter summary of HI and BO.

Different sequence	α_{max}	α_{min}	$\triangle\alpha$	h_{max}	h_{min}	$\triangle h$
The original sequence	0.945	0.487	0.458	0.947	0.426	0.521
Rearrange the sequence	0.752	0.358	0.394	0.912	0.356	0.556
Alternative sequence	0.914	0.462	0.452	0.896	0.512	0.384

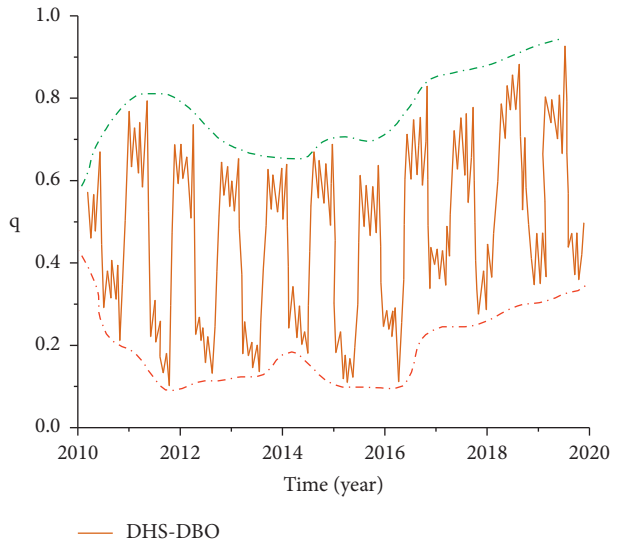

Figure 10: Dynamic Hurst exponent during the sliding window period.

where η is the learning rate.

(6) Threshold update

$$\begin{cases} a_j = a_j + \eta H_j\left(1 - H_j\right)\sum_{k=1}^{m} w_{jk} e_k, & j = 1, 2, \ldots, l, \\ b_k = b_k + e_k, & k = 1, 2, \ldots, m. \end{cases} \tag{17}$$

Optimization algorithms for machine learning include (1) Bat Algorithm (BA), (2) Particle Swarm Optimization (PSO), (3) Genetic Algorithm (GA), (4) Cuckoo Search Algorithm (CSA), and (5) Antlion Algorithm Optimization (ALO). Particle swarm optimization (PSO) is a kind of evolutionary algorithm. It starts from the random solution and finds the optimal solution through iteration. It also evaluates the quality of the solution through fitness. But it is simpler than the rules of the genetic algorithm, and it does not have the "crossover" and "mutation" operations of the genetic algorithm. It seeks the global optimal value by following the currently found optimal value. This algorithm has attracted the attention of academic circles for its advantages of easy implementation, high precision, and fast convergence, and it has demonstrated its superiority in solving

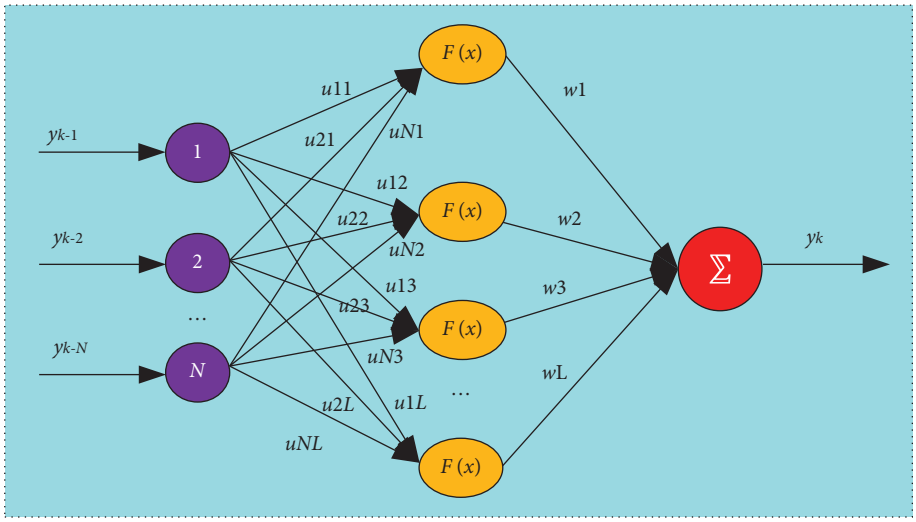

Figure 11: Structure diagram of wavelet neural network.

practical problems. In order to study the convergence speed and performance of the above five swarm intelligence optimization algorithms, this chapter adopts five commonly used minimum benchmark functions to comprehensively test the above swarm intelligence optimization algorithms to evaluate the performance of these optimization algorithms.

Each benchmark function optimization test was independently run for 30 times, and the results were output when each iteration reached the maximum number of iterations. The operation output corresponding to the minimum value obtained from the 30 independent operation results was taken as the test results, as shown in Figure 12.

Through different optimization algorithm of the iterative graph can be seen that with the increase of the number of iterations, the CSA algorithm corresponding to the first slow decline curve performance and then remain constant, and the state kept for a long time, with the further increase of the number of iterations, the curve of further rapid decline finally still tends to constant change trend. The curve corresponding to the GA algorithm decreases slowly at first and then keeps constant with the increase of the number of iterations. The slope of this curve also decreases slowly at first and gradually approaches 0 at last. The curve corresponding to the BA algorithm drops rapidly in a short time. When it drops to a certain value, the corresponding curve keeps the data constant with the further increase in the number of iterations. Compared with the above algorithms, the curve corresponding to the ALO algorithm shows a drop type change, which has a relatively large change range, indicating that the number of iterations has the highest impact on the ALO algorithm. The curve corresponding to the PSO algorithm and the curve corresponding to the GA algorithm have basically the same change rule, but the curve corresponding to the PSO algorithm has the largest change range.

4.2. Model Prediction and Analysis of Film Data. In order to further verify whether the Facebook happiness index can improve movie box office prediction performance, Long-

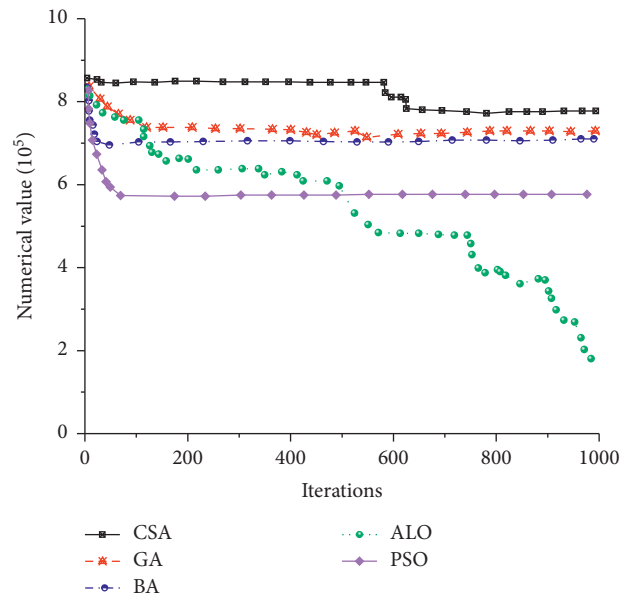

Figure 12: Iterative graph of different optimization algorithms.

Term Short-Term Memory Neural Network (LSTM) is used to predict movie box office. Select the time-series data affecting the total box office of the next day as the input feature. An optimizer is used to train the model. The sample size of each batch is 28, the learning rate is 0.001, the number of iterations is 20,000, and the window size is set to seven days. Normalization is conducted on each vector of the time series so as to obtain the prediction of the movie box office. Specific forecast data and curves are shown in Figure 13.

As can be seen from the box office prediction chart, the number of HI-LSTM data decreases gradually with the increase of income. When the income is 0–4 dollars, the total amount of HI-LSTM data is large, while when the income is 4–8 dollars, the HI-LSTM data decreases rapidly. In addition, by drawing the upper and lower curves of the data, it can be seen that the overall distribution of the data presents a

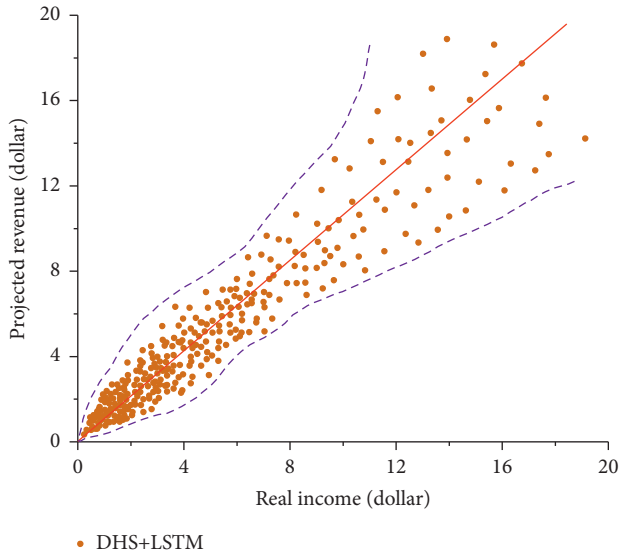

Figure 13: Box office forecast chart.

U-shaped change. Through linear fitting, it can be seen that the overall distribution of the data conforms to the characteristics of linear change. Finally, the prediction index of the HI-LSTM model in the box office is 0.924, which shows that the model can describe and predict the box office well and is very important for improving the accuracy of box office prediction.

5. Conclusion

(1) From the correlation analysis between the happiness index and the box office, it can be seen that the critical values of the two variables show a linear change trend with the increase of the degree of freedom, and the change range is about 3.1. The curve of Happiness Index (HI) and Box Office (BO) can be divided into a slow increasing stage, a steady increasing stage, and an approximately flat stage.

(2) It can be seen from the multifractal characteristic diagram of Happiness Index and Box Office (HI-BO) that, with the increase of q value, the change trend of long-term and short-term curves of generalized Hurst function is basically the same, but the difference between them becomes smaller and smaller, indicating that q value can promote the uniformity of Hurst curve. According to the Renyi function, with the increase of q value, the difference between the two curves becomes larger and larger, indicating that q value plays a role in promoting the difference between the two curves.

(3) With the increase of a value, the three different sequences all conform to the variation rule of the quadratic function. The variation range of the original sequence is the smallest, the variation range of the rearranged sequence is the second, and the variation range of the alternative sequence is the largest, indicating that under the influence of independent variables, the influence degree of the alternative sequence is the highest, the influence degree of the rearranged sequence is relatively low, and the corresponding original sequence has the lowest influence degree.

(4) As can be seen from the prediction chart of the movie box office, the overall distribution of data presents a U-shaped change, and the distribution of data conforms to the characteristics of linear change. The prediction index of the HI-LSTM model in film box office is 0.924, indicating that the model has good performance in describing and predicting film box office.

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.

Acknowledgments

This work was supported by a phased research result of the second industry-university-research collaborative education project: Practical Innovation and Creative Education of Documentary Creation (no.202102022076) of the Ministry of Education in 2021.

References

[1] K. Karishma and M. Parmar, "Sentiment analysis based on movie reviews using various classification techniques: a review," *International Journal of Scientific Research in Computer*

Science, Engineering and Information Technology, vol. 105, no. 18, pp. 197–208, 2021.

[2] J. Sanyam, M. Rohan, and K. Varsha, "Analyzing and predicting the success of box office collection of a movie using machine learning," *International Journal of Advanced Research in Science, Communication and Technology*, vol. 128, no. 16, pp. 325–331, 2021.

[3] H. Jung and C. Lim, "Predicting movie audience with stacked generalization by combining machine learning algorithms," *Communications for Statistical Applications and Methods*, vol. 28, no. 10, pp. 217–232, 2021.

[4] S. M. R. Abidi, Y. Xu, J. Ni, X. Wang, and W. Zhang, "Popularity prediction of movies: from statistical modeling to machine learning techniques," *Multimedia Tools and Applications*, vol. 79, no. 47-48, pp. 35583–35617, 2020.

[5] S. Sridevi and C. Murnal, "Implementation of movie recommendation system using machine learning," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 158, no. 12, pp. 587–593, 2020.

[6] Q. I. Mahmud, N. Z. Shuchi, F. M. Tawsif, A. Mohaimen, and A. Tasnim, "A machine learning approach to predict movie revenue based on pre-released movie metadata," *Journal of Computer Science*, vol. 16, no. 6, pp. 749–767, 2020.

[7] K. Indira and M. K. Kavithadevi, "Efficient machine learning model for movie recommender systems using multi-cloud environment," *Mobile Networks and Applications*, vol. 24, no. 6, pp. 1872–1882, 2019.

[8] C. Sachin, S. Deepak, and S. Niranjan, "Analysis of sentiment based movie reviews using machine learning techniques," *Journal of Intelligent and Fuzzy Systems*, vol. 41, no. 10, pp. 1–8, 2021.

[9] G. Shital and K. Amin, "Multi-batch quasi-Newton method with artificial neural network for movie recommendation," *Journal of the Institution of Engineers: Serie Bibliographique*, vol. 102, no. 4, pp. 729–742, 2021.

[10] Y. Wan and M. Ren, "New visual expression of anime film based on artificial intelligence and machine learning technology," *Journal of Sensors*, vol. 2021, no. 5, pp. 1–10, 2021.

[11] N. Iqbal, R. Ahmad, and J. Faisal, "Hybrid features prediction model of movie quality using multi-machine learning techniques for effective business resource planning," *Journal of Intelligent and Fuzzy Systems*, vol. 18, no. 6, pp. 10–32, 2021.

[12] J. Duan and R. Gao, "Research on English movie resource information mining based on dynamic data stream classification," *Security and Communication Networks*, vol. 20, no. 12, pp. 1–10, 2021.

[13] H. Sun, H. Shi, and A. Alterazi Hassan, "The use of neural network in defense audit nonlinear dynamic processing under the background of big data," *Fractals*, vol. 30, no. 2, pp. 224–242, 2021.

[14] J. Yue, "Retracted article: big data-based twin network target tracking and local regional economic development," *Personal and Ubiquitous Computing*, vol. 25, no. S1, p. 43, 2021.

[15] P. Yuan and k. Dere, "Machine learning methods for rockburst prediction-state-of-the-art review," *International Journal of Mining Science and Technology*, vol. 29, no. 4, pp. 44–49, 2019.

[16] D. Blackburn Landen, F. Tuttle Jacob, and M. Powell Kody, "Real-time optimization of multi-cell industrial evaporative cooling towers using machine learning and particle swarm optimization," *Journal of Cleaner Production*, vol. 271, no. 36, pp. 162–175, 2020.

[17] V. Lopez-Vazquez, J. M. Lopez-Guede, and S. Marini, "Video image enhancement and machine learning pipeline for underwater animal detection and classification at cabled observatories," *Sensors*, vol. 20, no. 3, pp. 145–168, 2020.

[18] C. Lian, C. Kang, and J. Lee, "The effect of rating dispersion on purchase of experience goods based on the Korean movie box office data," *Asia Marketing Journal*, vol. 21, no. 5, pp. 124–139, 2019.

[19] H. Ha, H. Han, and S. Mun, "An improved study of multilevel semantic network visualization for analyzing sentiment word of movie review data," *Applied Sciences*, vol. 9, no. 12, pp. 148–167, 2019.

[20] C. A. Rice, B. Beekhuizen, V. Dubrovsky, S. Stevenson, and B. C. Armstrong, "A comparison of homonym meaning frequency estimates derived from movie and television subtitles, free association, and explicit ratings," *Behavior Research Methods*, vol. 51, no. 3, pp. 1399–1425, 2019.

[21] M. Shalini, E. Tamma, and R. Talla, "Multi-genre movie data analysis using pearson's correlation," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 15, no. 8, pp. 881–887, 2019.

[22] P. Mohapatra, R. Singh, and S. Pandey, "Sentiment classification of movie review and twitter data using machine learning," *International Journal of Computer & Organization Trends*, vol. 9, no. 5, pp. 568–579, 2019.

[23] S. Styawati and K. Mustofa, "A support vector machine-firefly algorithm for movie opinion data classification," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 13, no. 3, p. 219, 2019.

[24] M. Ventura, H. Saulo, V. Leiva, and S. Monsueto, "Log-symmetric regression models: information criteria and application to movie business and industry data with economic implications," *Applied Stochastic Models in Business and Industry*, vol. 35, no. 4, pp. 963–977, 2019.

[25] S. Narayana, S. B. Chandanapalli, and S. Rao Mekala, "Ant cat swarm optimization-enabled deep recurrent neural network for big data classification based on map reduce framework," *The Computer Journal*, vol. 56, no. 12, pp. 135–154, 2021.

[26] M. Simakovic and Z. Cica, "Detection and localization of failures in hybrid fiber–coaxial network using big data platform," *Electronics*, vol. 10, no. 23, pp. 457–476, 2021.

[27] J. Fang, "Research on automatic cleaning algorithm of multi-dimensional network redundant data based on big data," *Evolutionary Intelligence*, vol. 48, no. 16, pp. 476–495, 2021.

[28] H. Jiang, L. Li, and H. Xian, "Crowd flow prediction for social internet-of-things systems based on the mobile network big data," *IEEE Transactions on Computational Social Systems*, vol. 48, no. 8, pp. 456–478, 2021.

[29] A. A. Dennis and P. Chenniappan, "Extended and optimized deep convolutional neural network-based lung tumor identification in big data," *International Journal of Imaging Systems and Technology*, vol. 16, no. 8, pp. 3511–3526, 2021.

[30] B. Seung and Na Su, "Sports field, school sports department, sports department leader Me Too keyword analysis using social network big data," *Korean Society For The Study Of Physical Education*, vol. 26, no. 15, pp. 133–144, 2021.