

Received 15 June 2023, accepted 29 June 2023, date of publication 3 July 2023, date of current version 11 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3291999

RESEARCH ARTICLE

Prediction of COVID-19 Data Using Improved ARIMA-LSTM Hybrid Forecast Models

YONG-CHAO JIN¹, QIAN CAO¹, KE-NAN WANG¹, YUAN ZHOU²,
YAN-PENG CAO¹, AND XI-YIN WANG^{1,3}

¹College of Science, North China University of Science and Technology, Tangshan 063210, China

²College of Artificial Intelligence, North China University of Science and Technology, Tangshan 063210, China

³Hebei Key Laboratory of Data Science and Application, Tangshan 063210, China

Corresponding authors: Xi-Yin Wang (wangxiyin@vip.sina.com) and Qian Cao (caoqian@ncst.edu.cn)


The work of Xi-Yin Wang was supported by the Hebei Key Laboratory of Data Science and Application and the Tangshan Municipal Funding for Talented Research under Grant 16013601.

ABSTRACT COVID-19 has developed into a global public health emergency and has led to restrictions in numerous nations. Thousands of deaths have resulted from the infection of millions of individuals globally. Additionally, COVID-19 has had a significant impact on social and economic activity around the world. The elderly and those with existing medical issues, however, are particularly vulnerable to the effects of COVID-19. Pneumonia, acute respiratory distress syndrome, organ failure, death, etc. are all possible outcomes in severe cases... Traditional prediction approaches like the ARIMA model and multiple linear regression model to handle the linear prediction problem because the new crown virus is in the process of continual mutation. Deep learning models that can take into account nonlinear elements include BP neural network prediction and LSTM neural network prediction. To combine the benefits of traditional and deep learning predictive models and create superior predictive models, we can blend traditional and deep learning predictive models. When the MSE, RMSE, and MAE of these three combined models, PSO-LSTM-ARIMA, MLR-LSTM-ARIMA, and BPNN-LSTM-ARIMA, are compared. We discovered that the third model, which included MSE, RMSE, and MAE, had the best prediction accuracy. The LSTM model and the ARIMA model were selected for this investigation. To begin, it employed a single model to forecast pandemic data in Germany. The BP neural network, particle swarm method, and multiple linear regression were then utilized to merge it. To corroborate this finding, we re-predicted the epidemic data from Japan and retrieved the MSE, RMSE, and MAE values of the BPNN-LSTM-ARIMA model, which were 6141895.956, 2478.285 and 1249.832. The most accurate model is still this integrated model. The BP neural network coupled LSTM model and ARIMA model offers the highest accurate prediction effect, according to our research. Combinatorial models anticipate outbreak data through our study, which can aid governments and public health authorities in improving their responses and educating the public about pandemic trends and potential future directions. As a result, industries and enterprises may make better risk-management decisions to protect the health and safety of their operations and personnel. It also helps healthcare facilities better prepare and deploy medical resources to better meet the demands from the pandemic.

INDEX TERMS COVID-19 prediction, ARIMA, LSTM, BPNN, MLR, PSO.

I. INTRODUCTION

The global epidemic, which has more than 762 million confirmed cases worldwide, is currently coming to an end.

The associate editor coordinating the review of this manuscript and approving it for publication was Jiajie Fan .

Reports state that the latest coronavirus is now to blame for 6.89 million fatalities [1]. The virus's ongoing mutation prevents us from letting up on our vigilance, so the research presented in this paper on the accuracy of the epidemic trend prediction model can increase the precision and accuracy of COVID-19 prediction and provide a theoretical

framework for epidemic prevention and control in various nations.

Researchers from several countries have forecasted and looked more closely at the COVID-19 trend since the COVID-19 epidemic. The three primary research approaches for the transmission and forecasting of epidemics are the infectious disease dynamics model, time series model, and machine learning model [2].

However, there is a dearth of research on the accuracy and error of epidemic transmission and prediction models. Some researchers use multiplicative trend exponential smoothing and LSTM forecast time series to predict COVID-19 cases [3], while others use a single ARIMA model to predict epidemic data [4], but these models are difficult to fully account for the epidemic's linear and nonlinear factors. In response to the aforementioned issues, our team has implemented the following changes: first, we utilize the ARIMA model and the LSTM model to more thoroughly evaluate linear and nonlinear components, and then we use the model to combine the prediction results of the two to produce more accurate results. When it comes to complex epidemic transmission, the complete use of models is essential, and our team's research can offer a more precise model for containing the epidemic. Our data came from Hopkins University (<https://www.jhu.edu/>) and we chose two nations, Japan and Germany, to represent the two countries between April 1, 2020 and March 9, 2023. We employed tools to pre-process the data after getting it, and the data was legitimate and dependable. We first selected the LSTM deep learning model to forecast the pandemic data in Germany. A recurrent neural network (RNN) called LSTM (Long Short-Term Memory) is frequently used to analyze sequence data [5]. LSTMs have more powerful memory and long-term dependent processing capacity than regular RNNs [6]. We then used the ARIMA model to predict the outbreak data in Germany. Based on observation and examination of past time series data, the ARIMA model uses autoregressive and moving average (ARMA) methodologies to forecast future values. Based on observation and examination of past time series data, the ARIMA model uses autoregressive and moving average (ARMA) methodologies to forecast future values. When dealing with linear issues, it performs better [7], [8], [9], [10]. We use the following three models to combine the LSTM model and the ARIMA model. The first is the multiple linear regression model. A typical statistical technique for creating models that explain the relationship between two or more dependent variables (or predictors) is multiple linear regression. The least squares method is frequently used to fit the data in multiple linear regression models because there are now numerous predictors instead of just one. This approach is centered on reducing the gap between the model's actual predicted values and actual observed values in order to reduce model error [11], [12], [13], [14], [15]. The second is the particle swarm model. Particle Swarm Optimization (PSO) is an optimization technique based on swarm intelligence that

tracks the population's and each individual's best solutions in order to find the objective function's optimal solution [16]. Finally, there is the BP neural network model. The backpropagation neural network model is a neural network model that is based on this approach [17]. In order to achieve the goal of prediction, it connects input data to output data through the connections between input layer, hidden layer, and output layer connections [18]. The benefits of BP neural network models include their ability to handle nonlinear relationships, adapt to different data types [19], and have a high level of flexibility and robustness [20]. Next, we combined the results of particle swarm fitting multiple linear function, multiple linear regression, and BP neural networks to obtain three combination models. By comparing the difference between the predicted and actual epidemic data, we found that the combination model of BP neural networks and LSTM and ARIMA models is the best combination model. Then, to confirm our findings, we used three combined models to predict the Japanese epidemic data. By comparing the predicted data to the actual values, we found that the combined model of the BP neural network and the LSTM and ARIMA models still had the highest prediction accuracy, supporting our findings. This more accurate combined model accounts for both linear and nonlinear elements affecting the pandemic. Finally, we project Japan for 60 days following March 9, 2023, using the combinatorial model. We intend to help the pandemic end sooner by using our portfolio approach.

A. RESEARCH IDEAS

We first train the BP neural network on the obtained two prediction data, particle swarm fitting multiple linear function and multiple linear regression, obtain the predicted value of the three combined models, obtain the predicted value [21] of the predicted value with the real value, and draw a time series graph. Finally, we compare the predicted value with the real value and draw a time series graph. By contrasting a single model with MSE, and MAE of the true value, as well as the combined model with the MSE, RMSE, and MAE of the genuine value, we may assess the model's correctness. After finding the most correct model, we chose the Japanese data for confirmation and repeated the method above to find the MSE, RMSE, and MAE models. We did this to make sure the conclusions were consistent with those in Germany. Finally, we choose the best model to forecast the epidemic data in Japan for a period of 60 days in the past and create a map of the time series. The overall technical roadmap of the article is shown in figure 1:

II. MODEL SELECTION

A. LSTM MODEL

The long short memory neural network (LSTM) (RNN) is a special kind of recurrent neural network. The original RNN frequently experiences the gradient explosion or vanishing problem during training, which makes it challenging

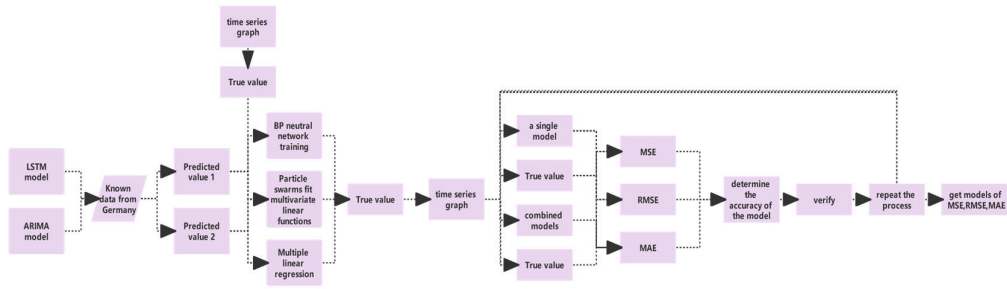


FIGURE 1. Technology roadmap.

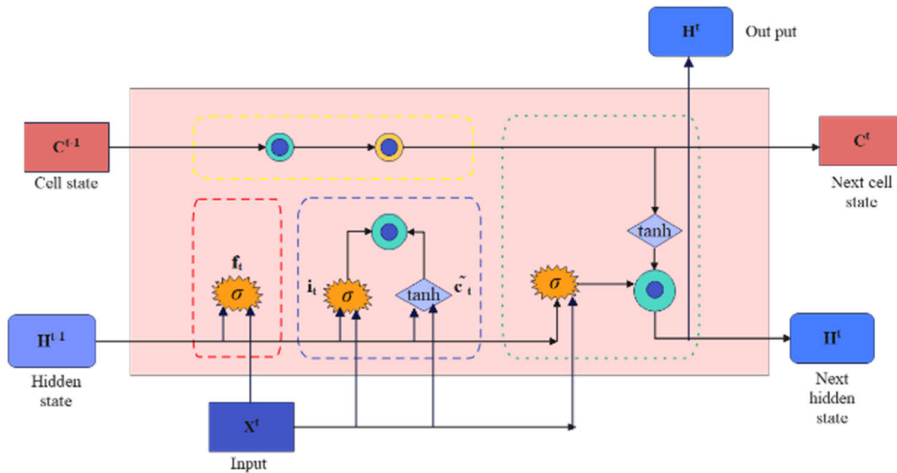


FIGURE 2. LSTM structural diagram.

to understand longer sequence data and, as a result, makes it impossible to extract information from long-distance data. The RNN is the LSTM’s ancestor. The LSTM [22] model addresses the short-term memory problem of RNN by incorporating gates on its foundation, allowing the cyclic neural network to effectively and fully employ long-distance time data. Three logic control units from LSTM, Input Gate, Output Gate, and Forget Gate [23], are now a part of the RNN infrastructure. Each of these units is tied to a multiplication element. You can regulate the input and output of the information flow as well as the state of the cell by modifying the weight value at the edge of the link between the neural network’s memory unit and other components.

The LSTM structural diagram is shown in figure 2:

Input Gate: The input gate, also known as, regulates the flow of information into the memory cell [24].

$$i_t = \text{sigmoid}(W_i \cdot [H^{t-1}, x_t] + b_i) \quad (1)$$

The Forget Gate, denoted by the symbol, controls whether data from the previous memory cell [25] is added to the

current memory cell.

$$f_t = \text{sigmoid}(W_f \cdot [H^{t-1}, x_t] + b_f) \quad (2)$$

$$\tilde{c}_t = \text{tanh}(W_c \cdot [H^{t-1}, x_t] + b_c) \quad (3)$$

Whether the data currently stored in the memory cell flows into the current hidden state is determined by the output gate, also known as O_t

$$o_t = \text{sigmoid}(W_o \cdot [H^{t-1}, x_t] + b_o) \quad (4)$$

Long-distance historical data can be stored, retrieved, reset, and updated by LSTM units thanks to the Cell: Memory unit, which simulates the memory of neuronal states. Short memory is indicated by and long memory is indicated by H^t

$$c_t = f_t \times c_{t-1} + i_t * \tilde{c}_t \quad (5)$$

$$h_t = o_t \times \text{tanh}(C^t) \quad (6)$$

The recursive connection weights of their associated thresholds are represented if is the input variable at time t, is the previous hidden state, is the future hidden state, and is the offset term. where the sigmoid and tanh activation functions are located.

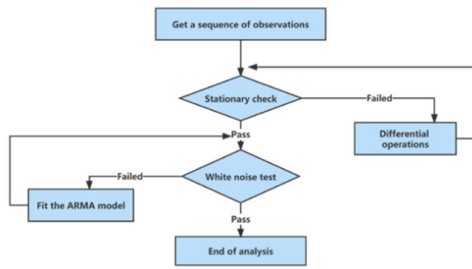


FIGURE 3. ARIMA operation flowchart.

Any integer between 0 and 1 can be converted to a binary classification using the sigmoid function. It is both differentiable and a smooth step function. The formula is as follows:

$$F(X) = \frac{1}{1 + e^{-x}} \tag{7}$$

The derivative value range for the tanh function is 0 to 1, and its output range is [-1, 1]. It is akin to an increasing amplitude sigmoid. The sigmoid's 0 to 1/4 range helps to slightly reduce the issue of gradient vanishing. The formula is as follows [26]:

$$H(X) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{8}$$

B. ARIMA(P,D,Q) MODEL

$$A(B)\nabla^d y(t) = C(B)e(t) \tag{9}$$

$$A(B) = 1 - a_1B - a_2B^2 - \dots - a_pB^p \tag{10}$$

$$C(B) = 1 - c_1B - c_2B^2 - \dots - c_pB^p \tag{11}$$

where {y(t)} and {e(t)} respectively represent the original sequence and the white noise sequence, and B is the backward operator, which satisfies the expression B^n y(t) = y(t-n), n = 1, 2, ..., ∇^d = (1 - B)^d is d difference, d = 1, carry out a differential processing, namely that, z_1(t) = ∇y(t) = y(t) - y(t - 1); d = 2, perform two differential processes, z_2(t) = ∇^2y(t) = ∇z_1(t) = z_1(t) - z_1(t - 1), and so on [27], [28] The ARIMA operation flowchart is shown in figure 3:

C. MULTIPLE LINEAR REGRESSION MODELS

Explanatory variable and the other multiple explanatory factors are provided by a multiple linear regression model, which is a linear regression model with multiple explanatory variables. Modeling it mathematically is

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_px_p + \varepsilon \tag{12}$$

It is clear that there are p-explanatory variables in the aforementioned equation, which indicates a p-element linear regression model. The change in the explanatory variable y may have two components: the first component is the linear change in y brought on by the change in the p explanatory

variables x:

$$\beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_px_p \tag{13}$$

The second part explains the portion of the change in y that is due to the random variable. This portion can be replaced by a portion that can be referred to as random error. The parameters in the formula are all the unknowns in the equation and can be expressed as partial regression constants and regression constants. So, a multiple linear regression model's regression equation is:

$$E(y) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_px_p \tag{14}$$

D. PARTICLE SWARM MODEL

An evolutionary computing method is particle swarm optimization, derived from research on the predation behavior of bird flocks. The fundamental goal of particle swarm optimization algorithms is to discover the best solution through group cooperation and information exchange. Only two characteristics of particles exist: position and speed, where position denotes the direction of motion and velocity the speed of the movement. Each particle performs a separate search for the ideal solution in the search space, records it as the current individual extreme, shares the value of the individual extreme with all of the other particles in the particle swarm, and discovers the ideal individual extreme value as the current global optimal solution of the entire particle swarm. and every particle in the particle swarm modifies its speed and position in accordance with the most recent global optimal solution discovered by the entire particle swarm. The algorithm uses six crucial parameters.

If there are N particles, each of which represents a solution, in the D-dimensional [29] search space, then:

The ith particle's location is:

$$X_{id} = (x_{i1}, x_{i2}, \dots, x_{iD}) \tag{15}$$

The velocity of the ith particle (the distance and direction in which the particle moves) is:

$$V_{id} = (v_{i1}, v_{i2}, \dots, v_{iD}) \tag{16}$$

The optimal position searched for by the ith particle is:

$$P_{id,pbest} = (p_{i1}, p_{i2}, \dots, p_{iD}) \tag{17}$$

The optimal position (group optimal solution) for group search is:

$$P_{d,gbest} = (p_{1,gbest}, p_{2,gbest}, \dots, p_{D,gbest}) \tag{18}$$

The adaptation value of the optimal position searched for by the ith particle is: f_p

The adaptation values for the optimal position searched by the population are: f_g

Speed update formula [30]:

The expression is called velocity, which is actually the distance and direction of the next iteration of the particle, that

is, a position vector:

$$v_{id}^{k+1} = \omega v_{id}^k + c_1 r_1 (p_{id}^k, \text{pbest} - x_{id}^k) + c_2 r_2 (p_{id}^k, \text{gbest} - x_{id}^k) \quad (19)$$

Thereinto:

N: particle swarm size, i: particle number, $i = 1, 2, 3, \dots, N$

D: particle dimension, d: particle dimension serial number, $i = 1, 2, 3, \dots, N$

K: Number of iterations: weight of inertia

$r_1 r_2$: A random number in the interval [0,1] to increase the randomness of the search

v_{id}^k : The velocity vector of particle i in dimension d, in the kth iteration

x_{id}^k : The position vector of particle i in the d-dimension in the kth iteration

p_{id}^k, pbest : The historical optimal position of particle i in the d-dimensional in the kth iteration, that is, the optimal solution obtained by the ith particle (individual) search after the kth iteration

$p_{d, \text{gbest}}^k$: The historical optimal position of the population in the d-dimension in the kth iteration, that is, the optimal solution in the entire particle population after the kth iteration.

Particle swarm algorithm flowchart is shown in figure 4:

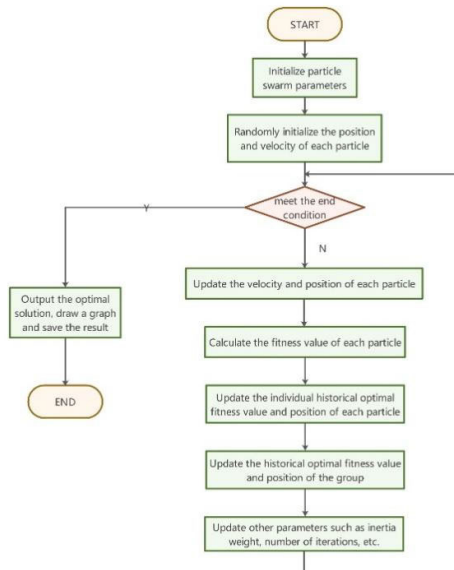


FIGURE 4. Particle swarm algorithm flowchart.

E. BP NEURAL NETWORK

The BP neural network [18] is a type of multi-layer feedforward network that was trained using error backpropagation (also known as error back transmission). Its algorithm is referred to as the BP algorithm, and its fundamental idea is to use gradient descent and gradient search technology to

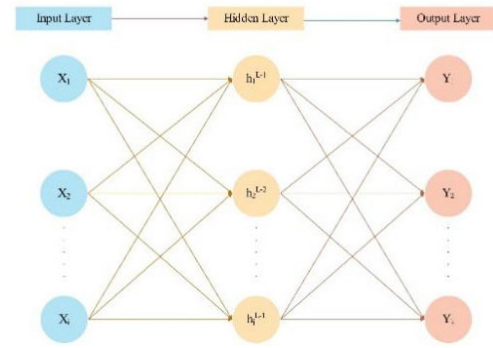


FIGURE 5. Diagram of a multi-layer neural network.

minimize the mean square error between the network’s actual and expected output values.

A forward calculation process and an inverse calculation process make up P neural network’s calculation process. A neuron’s state only influences the state of the layer of neurons after it in the forward propagation process, which processes the input mode layer by layer from the input layer through the hidden layer and onto the output layer. The error signal is returned along the original connection path in reverse propagation if the desired output cannot be produced in the output layer. The error signal is minimized by adjusting the weight of each neuron. Diagram of a multi-layer neural network is shown in figure 5.

Input layer input vector

$$X = (x_1, x_2, \dots, x_i, \dots, x_m) \quad (20)$$

L-layer implied layer vector

$$H^l = (h_1^l, h_2^l, \dots, h_j^l) \times (l = 2, 3, \dots, L - 1, j = 1, 2, \dots, j_l); \quad (21)$$

Output layer output vector

$$Y = (y_1, y_2, \dots, y_k, \dots, y_n) \quad (22)$$

Set the bias of the jth neuron in layer L and the connection weight between the ith and jth neurons in layer L-1 to obtain:

$$h_j^l = f(\text{net}_j^l) \quad (23)$$

$$\text{net}_j^l = \sum_{j=1}^{s_{j-1}} w_{ij}^l + b_j^l \quad (24)$$

where is the activation function, which is the same as the activation function of the LSTM model, of the jth neuron input of the L layer.

III. MODEL EVALUATION METRICS

To assess the model’s accuracy, it is crucial to employ a variety of performance evaluation criteria because the model may perform well for one metric but poorly for another. To assess our model, we used the three measures shown below.

1) Average absolute error (MAE): The mean defines the variation in the data collection, and the MAE of the dataset is the average distance between each data value. The excellent accuracy of the model is seen if the MAE is close to zero.

$$MAE = \frac{1}{m} \sum_{i=1}^m |y_i - f(x_i)| \quad (25)$$

2) The MSE, or mean squared error, is the square of the true value and the predicted value interpolated values, which are then added together and averaged.

$$MSE = \frac{1}{m} \sum_{i=1}^m (y_i - f(x_i))^2 \quad (26)$$

3) RMSE root mean square error: The root mean square error is the ratio of the number of observations n to the square root of the square of the deviation of the predicted value from the true value.

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^n (Y_i - f(x_i))^2} \quad (27)$$

IV. DATASET SELECTION

The data were obtained from the Hopkins University website (<https://www.jhu.edu/>), and a total of infected data from Japan and Germany from April 1, 2020 to March 9, 2023 were chosen since these two nations had more severe epidemics, were geographically far from one another, and were typical. The boxplot, outliers are removed, missing values are linearly imputed, and the data set is preprocessed by SPSS software before it is given to the model for training. This produces pretty smooth data.

V. MODEL APPLICATION

A. LSTM MODEL

The LSTM model is developed in this work using data available in Germany from April 1, 2020, to March 9, 2023, and the following parameters are chosen in Table 1:

TABLE 1. LSTM model parameter.

parameter	value
Hidden Units	250
Max Epochs	300
Rate of Learning	0.01
Optimization Approach	Adam
Learn Rate Drop Period	300
Learn Rate Drop Factor	0.38

We created the LSTM model using the software, chose 70% of the data for the training set and 30% for the test set, got the predicted value, and will compare the predicted value to the actual value to create a time series diagram in figure 6.

The value of the evaluation index is determined by comparison with the true value, as indicated in Table 2:

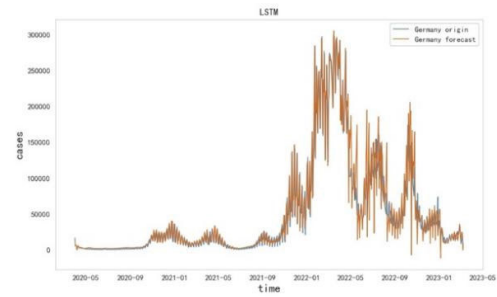


FIGURE 6. LSTM predictive effect in Germany.

TABLE 2. LSTM model evaluation index.

Evaluation indicators	numeric value
MSE	51566389.024
RMSE	36375.064
MAE	17312.186

B. ARIMA TIME SERIES MODEL

The ARIMA model is developed for the model in this study using data from April 1, 2020 to March 9, 2023 in Germany. The final model is ARIMA (0,1,9), the R side of the model reaches 0.95, and the anticipated time series diagram is obtained in figure 7:

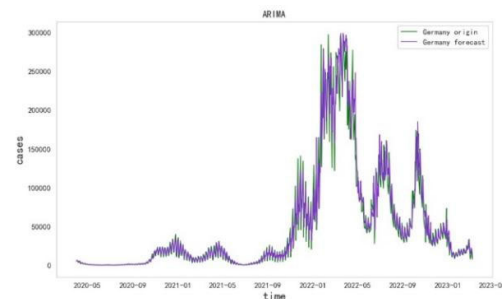


FIGURE 7. ARIMA predictive effect in Germany.

By comparing with the real value, we derive the evaluation index of the ARIMA model in Table 3:

TABLE 3. ARIMA evaluation indicators.

Evaluation indicators	numeric value
MSE	219950944.182
RMSE	14830.743
MAE	7383.430

C. COMBINED MODELS

1) PARTICLE SWARM FITTING MULTIVARIATE LINEAR FUNCTIO [31]

In this study, we use the predicted values of ARIMA and LSTM to fit the particle swarm algorithm, allowing our

combinatorial model to take into account both linear and nonlinear factors. We then use software to implement the particle swarm algorithm, setting the objective function as:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^n (Y_i - f(x_i))^2} \quad (28)$$

Our particle swarm technique, which sets $nvars = 2$, and two variables, K_1, K_2 , as the coefficients in front of the LSTM model and the ARIMA model, aims to minimize the RMSE of the model. We ultimately arrive at $K_1 = 0.2518$ and $K_2 = 0.7127$ through calculations.

Thus, the linear equation that is fitted is:

$$\hat{y}_1 = 0.2518 x_{LSTM} + 0.7127 x_{ARIMA} \quad (29)$$

2) THE TIME SERIES PLOT IS MADE IN FIGURE 8

Following the computation of the evaluation index of the PSO-LSTM-ARIMA combined model, we can already see from the time series diagram that the combined model's prediction accuracy has increased in comparison to the single model in Table 4.

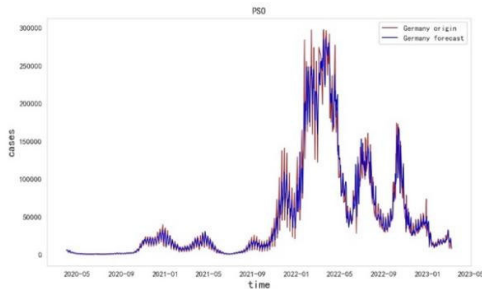


FIGURE 8. Time series plots of combined models.

TABLE 4. PSO-LSTM-ARIMA evaluation index.

Evaluation indicators	numeric value
MSE	174276596.420
RMSE	13201.386
MAE	6685.248

D. BP NEURAL NETWORK FITTING

We train the model using 70% of the training data, 15% of the verification data, and 15% of the test data in order to fit the predicted values of the LSTM and ARIMA mdels [32], [33]. The fitting effect is then obtained in figure 9:

The value of R2 reaches 0.97254, and the model's fitting effect is better, as can be seen from the above image. The BPNN-LSTM-ARIMA combination mode's predicted value is obtained by combining the prediction data from the LSTM with the prediction data from the ARIMA model. The time series diagram is then drawn in figure 10:

We calculate the evaluation index of the BPNN-LSTM-ARIMA combination model in Table 5:

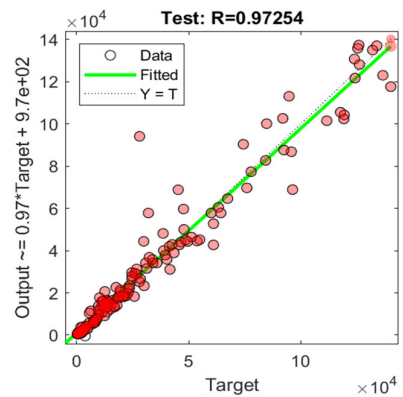


FIGURE 9. The BPNN model tests the fitting effect of the set.

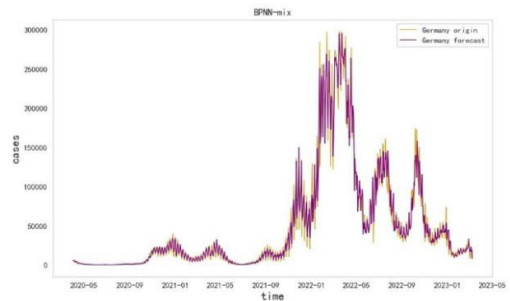


FIGURE 10. Time series plot of BPNN.

TABLE 5. BPNN-LSTM-ARIMA model evaluation index.

Evaluation indicators	numeric value
MSE	118411334.292
RMSE	10881.697
MAE	5687.514

E. MULTIPLE LINEAR REGRESSION COMBINATORIAL MODEL

We choose the third combinatorial model as the multiple linear regression-LSTM-ARIMA model, and let the model equation be as follows:

$$\hat{y}_2 = M_1 x_{LSTM} + M_2 x_{ARIMA} \quad (30)$$

We use software to perform multiple linear regression on the prediction data, and obtain the parameters of the model in Table 6:

The R² of the model reaches 0.96, as seen in the above table, and the P value is near to 0, indicating that our model fits the data better and that the regression coefficients are significant. The following is the regression equation that results:

$$\hat{y}_2 = 0.25069 x_{LSTM} + 0.71196 x_{ARIMA} + 267.62 \quad (31)$$

We put the predicted data into the regression equation to get the time series diagram in figure 11

TABLE 6. Multiple linear regression parameters.

parameter	numeric value
Regression coefficient M_1	0.251
Regression coefficient M_2	0.712
Constant terms	267.620
R^2	0.960
P-value	0.000

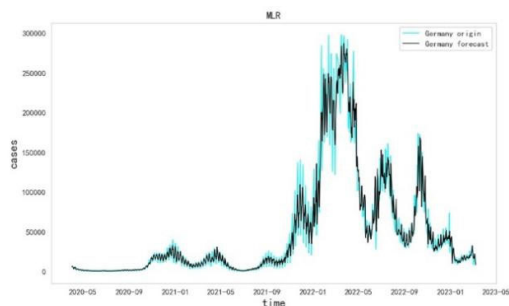


FIGURE 11. Time series plot for multiple linear regression combinatorial models.

We calculate the evaluation index of the combinatorial model multiple linear regression-LSTM-ARIMA model in Table 7:

TABLE 7. Multiple linear regression-LSTM-ARIMA model evaluation index.

evaluating indicator	numeric value
MSE	174230163.633
RMSE	13199.627
MAE	6727.384,

VI. COMPARISON OF COMBINED MODELS

The BPNN-LSTM-ARIMA model has the best prediction accuracy, followed by the MLR-LSTM-ARMA model and PSO-LSTM-ARIMA model [34]. In the previous article, we calculated the prediction accuracy of the three combined models, and then we compared the accuracy of the three models. In order to confirm our conclusion, we will use the aforementioned method to make a prediction on the epidemic data in Japan.

VII. MODEL VALIDATION

After determining that the ARIMA model is ARIMA (0,19), the model, we set the LSTM model parameters to match those used to predict Germany, and we then drew the time series diagram as follows:

First, we use the LSTM model and the ARIMA model to draw the sequence diagram in figure and figure 12:

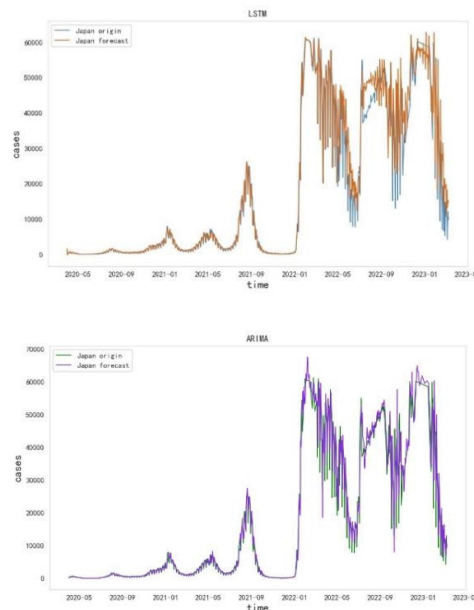


FIGURE 12. Time series plot of the LSTM and ARIMA model.

Next, we calculate the particle swarm fit multivariate linear function, and the multiple linear regression equation is as follows:

Particle swarm multivariate linear fitting

$$\hat{y}_3 = 0.3506x_{LSTM} + 0.6218x_{ARIMA} \tag{32}$$

Multiple linear regression equation:

$$\hat{y}_4 = 0.35433x_{LSTM} + 0.62192x_{ARIMA} - 168.9 \tag{33}$$

R^2 is 0.982 in the multiple linear regression equation. Our regression coefficient is significant and the model fits well because the p-value is zero.

A. BPNN-LSTM-ARIMA MODEL VALIDATION

The number of hidden layers is set to 10 in the training parameters of the BP neural network in the data used to predict Japan, and the other parameters are compatible with the prior values. The following is how the model training effect is obtained in figure 13:

The number of hidden layers is set to 10 in the training parameters of the BP neural network in the data used to predict Japan, and the other parameters are compatible with the prior values. The following is how the model training effect is obtained in figure 14:

VIII. MODEL COMPARISON

We calculate the MSE, RMSE, MAE of the five models, and get the results in Table 8:

We visualize the comparison results as shown in figure 15:

The BPGN-LSTM-ARIMA model's indicators, as seen in the aforementioned image, are lower than those of other models, supporting the earlier conclusion that it is the most accurate prediction model.

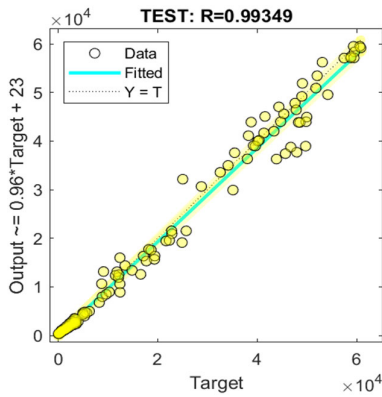


FIGURE 13. Fitting of the BPNN test set.

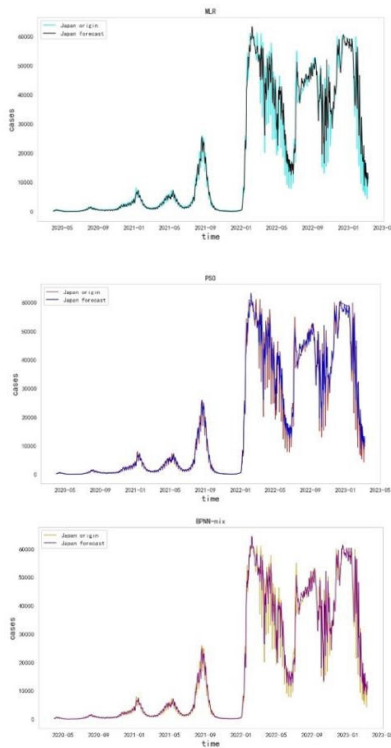


FIGURE 14. Line chart of three models that make predictions on Japanese data.

IX. ANALYSIS OF SEASONAL FACTORS FOR THE MODEL

In order to show that our model is more thorough, we looked at seasonal factors. We chose 92 days of data from Japan’s June 1 to August 31 as summer data, and then we chose the epidemic data from the following year’s November 1 to January 31 as winter data. We first used the ARIMA model to predict the data, and it does a good job of accounting for seasonal factors. Then, we are making predictions with our best forecasting model, LSTM-ARIMA-BPNN, and comparing MSE, RMSE, and MAE. Through the soft armor implementation, we obtain the comparative data as shown in Table 9, 10 below:

TABLE 8. Five model evaluation indicators.

model	Evaluation indicators	numeric value
LSTM model	MSE	20582526.517
	RMSE	4536.797
	MAE	2412.680
ARIMA model	MSE	11849988.738
	RMSE	3442.381
	MAE	1752.846
PSO-LSTM-ARIMA model	MSE	7844826.219
	RMSE	2800.862
	MAE	1414.308
MLR-LSTM-ARIMA model	MSE	7828502.594
	RMSE	2797.946
	MAE	1459.894
BPNN-LSTM-ARIMA model	MSE	6141895.956
	RMSE	2478.285
	MAE	1249.832

TABLE 9. Summer data model comparison.

model	Evaluation indicators	numeric value
ARIMA model	MSE	16393.798
	RMSE	128.038
	MAE	87.2641
BPNN-LSTM-ARIMA model	MSE	7880.645
	RMSE	88.773
	MAE	72.224

TABLE 10. Winter data model comparison.

model	Evaluation indicators	numeric value
ARIMA model	MSE	437362.808
	RMSE	661.334
	MAE	456.708
BPNN-LSTM-ARIMA model	MSE	206108.476
	RMSE	453.991
	MAE	300.710

According to our experimental data, the best model generated by us has greatly better prediction accuracy compared to the ARIMA model in the two usual seasons of winter and summer, and our model also predicts seasonal infectious illnesses well.

X. SUDDEN DATA CHANGE FORECASTS

We discovered through the time series chart that the epidemic data in Japan changed abruptly from November 2021 to March 2022, and we used the LSTM model and the best

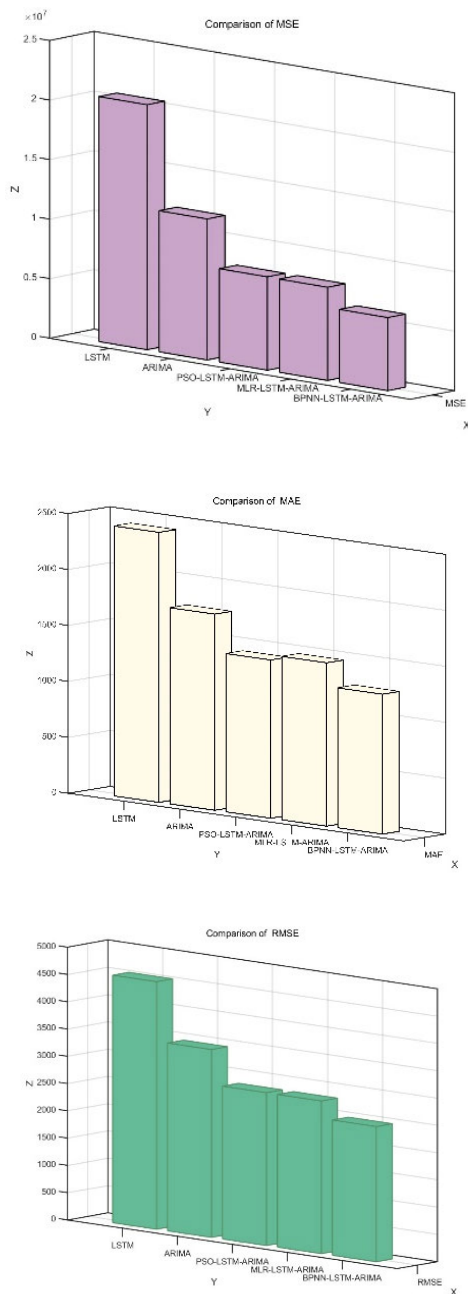


FIGURE 15. Evaluation indicators of three combined models.

prediction model obtained by our experimental research to predict this part of the data, comparing the MSE, RMSE, and MAE of the two. The resulting comparative data are shown in Table 11 below:

According to our experimental results, the model LSTM-ARIMA-BBNN model still works effectively in the face of abrupt data changes, and its accuracy is much higher than that of the single model LSTM.

XI. EPIDEMIC FORECAST IN JAPAN

To anticipate the epidemic data in Japan 60 days in advance, we choose the most precise prediction model, BPGN-LSTM-ARIMA, and obtain the prediction time series in figure 16:

TABLE 11. Sudden change data forecast-indicator comparison.

model	Evaluation indicators	numeric value
LSTM model	MSE	21106839.398
	RMSE	4594.218
	MAE	2161.410
BPNN-LSTM-ARIMA model	MSE	7621395.099
	RMSE	2760.6874
	MAE	1400.641

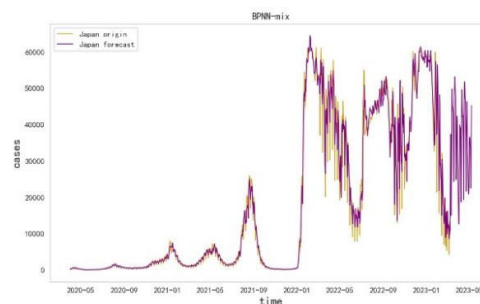


FIGURE 16. Time series plots that forecast Japanese data.

The prediction model developed in this paper can, as can be seen from the aforementioned figure, achieve a good prediction effect on the number of new cases in the future. When compared to the actual value, it is discovered that it essentially matches the real trend of the number of confirmed cases and can, to a certain extent, predict the development of the number of epidemic cases in a given area.

XII. CONCLUSION

A. RESULTS AND DISCUSSIONS

We created three combination models MLR-LSTM-ARIMA, PSO-LSTM-ARIMA, and BPNN-LSTM-ARIMA—using data from the COVID-19 outbreak that affected Germany and Japan. The MSE, RMSE, and MAE values for the prediction of epidemic data in Japan are 6141895.956, 2478.28, and 1249.83, respectively. These values significantly increase the prediction accuracy. The value of MAE is 118411334.292, 10881.697, and 5687.514. In the combined BPNN-LSTM-ARIMA model, the LSTM model and the ARIMA model can both effectively take into account the linear and nonlinear factors of the epidemic data. After receiving the results of the two, we then use the BP neural network to fit, in order to more thoroughly synthesize the benefits of the two single models. Then, using the BPNN-LSTM-ARIMA combination model to forecast the number of confirmed cases in Japan during the course of the following 60 days (up until May 8, 2023), we found that the pace of the epidemic has slowed down.

B. CONCLUSION AND SUGGESTIONS

The accuracy of epidemic prediction can be greatly increased by using combinatorial models, which combine the benefits

of deep learning LSTM model and conventional prediction model ARIMA model. We cannot only consider linear factors and non-linear factors when predicting an epidemic; the best solution should be to combine the two for comprehensive analysis. According to our research, various factors frequently contribute to disease outbreaks, so in order to anticipate COVID-19, we must be more adaptable in how we employ models.

Governments from all across the world must collaborate in order to create sensible anti-epidemic laws and maximize protection for people's lives if we are to successfully combat the epidemic.

REFERENCES

- [1] D. Haritha, N. Swaroop, and M. Mounika, "Prediction of COVID-19 cases using CNN with X-rays," in *Proc. 5th Int. Conf. Comput., Commun. Secur. (ICCCS)*, Patna, India, Oct. 2020, pp. 1–6, doi: [10.1109/ICCCS49678.2020.9276753](https://doi.org/10.1109/ICCCS49678.2020.9276753).
- [2] M. A. Achterberg, B. Prasse, L. Ma, S. Trajanovski, M. Kitsak, and P. Van Mieghem, "Comparing the accuracy of several network-based COVID-19 prediction algorithms," *Int. J. Forecasting*, vol. 38, no. 2, pp. 489–504, Apr. 2022, doi: [10.1016/j.ijforecast.2020.10.001](https://doi.org/10.1016/j.ijforecast.2020.10.001).
- [3] M. A. M. Arceda, P. C. L. Laura, and V. E. M. Arceda, "Forecasting time series with multiplicative trend exponential smoothing and LSTM: COVID-19 case study," in *Proc. Future Technol. Conf. (FTC)*, 2021, p. 1289.
- [4] A. K. Sahai, N. Rath, V. Sood, and M. P. Singh, "ARIMA modelling & forecasting of COVID-19 in top five affected countries," *Diabetes Metabolic Syndrome, Clin. Res. Rev.*, vol. 14, no. 5, pp. 1419–1427, Sep. 2020, doi: [10.1016/j.dsx.2020.07.042](https://doi.org/10.1016/j.dsx.2020.07.042).
- [5] S. Bouktif, A. Fiaz, A. Ouni, and M. A. Serhani, "Multi-sequence LSTM-RNN deep learning and metaheuristics for electric load forecasting," *Energies*, vol. 13, no. 2, p. 391, Jan. 2020, doi: [10.3390/en13020391](https://doi.org/10.3390/en13020391).
- [6] R. Pal, A. A. Sekh, S. Kar, and D. K. Prasad, "Neural network based country wise risk prediction of COVID-19," *Appl. Sci.*, vol. 10, no. 18, p. 6448, Sep. 2020, doi: [10.3390/app10186448](https://doi.org/10.3390/app10186448).
- [7] Y. Jin, R. Wang, X. Zhuang, K. Wang, H. Wang, C. Wang, and X. Wang, "Prediction of COVID-19 data using an ARIMA-LSTM hybrid forecast model," *Mathematics*, vol. 10, no. 21, p. 4001, Oct. 2022.
- [8] Y. Ning, H. Kazemi, and P. Tahmasebi, "A comparative machine learning study for time series oil production forecasting: ARIMA, LSTM, and prophet," *Comput. Geosci.*, vol. 164, Jul. 2022, Art. no. 105126.
- [9] T. Kriebchaumer, A. Angus, D. Parsons, and M. R. Casado, "An improved wavelet-ARIMA approach for forecasting metal prices," *Resour. Policy*, vol. 39, pp. 32–41, Mar. 2014.
- [10] S. Roy, G. S. Bhunia, and P. K. Shit, "Spatial prediction of COVID-19 epidemic using ARIMA techniques in India," *Model. Earth Syst. Environ.*, vol. 7, no. 2, pp. 1385–1391, Jun. 2021, doi: [10.1007/s40808-020-00890-y](https://doi.org/10.1007/s40808-020-00890-y).
- [11] H. Shu, C. Zou, J. Chen, and S. Wang, "Research on micro/nano surface flatness evaluation method based on improved particle swarm optimization algorithm," *Frontiers Bioeng. Biotechnol.*, vol. 9, pp. 1–10, Dec. 2021.
- [12] D. Wang, D. Tan, and L. Liu, "Particle swarm optimization algorithm: An overview," *Soft Comput.*, vol. 22, no. 2, pp. 387–408, Jan. 2018.
- [13] M. Kohler, M. M. B. R. Vellasco, and R. Tanscheit, "PSO+: A new particle swarm optimization algorithm for constrained problems," *Appl. Soft Comput.*, vol. 85, Dec. 2019, Art. no. 105865.
- [14] S. Rath, A. Tripathy, and A. R. Tripathy, "Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model," *Diabetes Metabolic Syndrome, Clin. Res. Rev.*, vol. 14, no. 5, pp. 1467–1474, Sep. 2020, doi: [10.1016/j.dsx.2020.07.045](https://doi.org/10.1016/j.dsx.2020.07.045).
- [15] X. Huang, H. Wang, W. Luo, S. Xue, F. Hayat, and Z. Gao, "Prediction of loquat soluble solids and titratable acid content using fruit mineral elements by artificial neural network and multiple linear regression," *Scientia Horticulturae*, vol. 278, Feb. 2021, Art. no. 109873, doi: [10.1016/j.scienta.2020.109873](https://doi.org/10.1016/j.scienta.2020.109873).
- [16] Y. Xiang and L. Jiang, "Water quality prediction using LS-SVM and particle swarm optimization," in *Proc. 2nd Int. Workshop Knowl. Discovery Data Mining*, 2009, pp. 900–904, doi: [10.1109/WKDD.2009.217](https://doi.org/10.1109/WKDD.2009.217).
- [17] L. Wang, Y. Zeng, and T. Chen, "Back propagation neural network with adaptive differential evolution algorithm for time series forecasting," *Expert Syst. Appl.*, vol. 42, no. 2, pp. 855–863, Feb. 2015.
- [18] H. Chen, Y. Wang, M. Zuo, C. Zhang, N. Jia, X. Liu, and S. Yang, "A new prediction model of CO₂ diffusion coefficient in crude oil under reservoir conditions based on BP neural network," *Energy*, vol. 239, Jan. 2022, Art. no. 122286.
- [19] Z. Wang, F. Wang, and S. Su, "Solar irradiance short-term prediction model based on BP neural network," *Energy Proc.*, vol. 12, pp. 488–494, Jan. 2011, doi: [10.1016/j.egypro.2011.10.065](https://doi.org/10.1016/j.egypro.2011.10.065).
- [20] L. Zhang, F. Wang, B. Xu, W. Chi, Q. Wang, and T. Sun, "Prediction of stock prices based on LM-BP neural network and the estimation of overfitting point by RDCI," *Neural Comput. Appl.*, vol. 30, no. 5, pp. 1425–1444, Sep. 2018, doi: [10.1007/s00521-017-3296-x](https://doi.org/10.1007/s00521-017-3296-x).
- [21] L. Mao, Y. Huang, X. Zhang, S. Li, and X. Huang, "ARIMA model forecasting analysis of the prices of multiple vegetables under the impact of the COVID-19," *PLoS ONE*, vol. 17, no. 7, Jul. 2022, Art. no. e0271594.
- [22] Y. K. Guo, Y. Y. Li, and Y. Xu, "Study on the application of LSTM-LightGBM model in stock rise and fall prediction," in *Proc. MATEC Web Conf.*, vol. 336, 2021, p. 05011, doi: [10.1051/mateconf/202133605011](https://doi.org/10.1051/mateconf/202133605011).
- [23] C. Zhang, D. Wang, J. Jia, L. Wang, K. Chen, L. Guan, Z. Liu, Z. Zhang, X. Chen, and M. Zhang, "Potential failure cause identification for optical networks using deep learning with an attention mechanism," *J. Opt. Commun. Netw.*, vol. 14, no. 2, pp. A122–A133, Feb. 2022.
- [24] X. Yin, S. Zheng, and Q. Wang, "Fine-grained Chinese named entity recognition based on RoBERTa-WWM-BiLSTM-CRF model," in *Proc. 6th Int. Conf. Image, Vis. Comput. (ICIVC)*, Jul. 2021, pp. 1–15.
- [25] N. Sakinah, M. Tahir, T. Badriyah, and I. Syarif, "LSTM with Adam optimization-powered high accuracy preeclampsia classification," in *Proc. Int. Electron. Symp. (IES)*, Sep. 2019, pp. 314–319.
- [26] D. Shao, Q. An, K. Huang, Y. Xiang, L. Ma, J. Guo, and R. Yin, "Aspect-level sentiment analysis for based on joint aspect and position hierarchy attention mechanism network," *J. Intell. Fuzzy Syst.*, vol. 42, no. 3, pp. 2207–2218, Feb. 2022.
- [27] Z. Jing and C. Jun, "Based on time sequence of ARIMA and BP neural network combination forecast model," *J. Statist. Decis.*, 2016.
- [28] H. Chao and S. Su, "Real-time adaptive prediction of short-term traffic flow based on ARIMA model," *J. Syst. Simul.*, 2004.
- [29] X. Wang and W. Yao, "Research on transmission task static allocation based on intelligence algorithm," *Appl. Sci.*, vol. 13, no. 6, p. 4058, Mar. 2023.
- [30] Q. Xiao and H. Wang, "Prediction of WEEE recycling in China based on an improved grey prediction model," *Sustainability*, vol. 14, no. 11, p. 6789, Jun. 2022.
- [31] Z. Y. Song, J. Xu, B. Y. Li, and X. Li, "Short-term passenger flow prediction of subway station based on PSO-LSTM," in *Proc. Int. Conf. Elect. Inf. Technol. Rail Transp.* Cham, Switzerland: Springer, 2022, pp. 355–363.
- [32] H. Abbasianjahromi and S. Shojaekhah, "Structural reliability assessment of steel four-bolt unstiffened extended end-plate connections using Monte Carlo simulation and artificial neural networks," *Iranian J. Sci. Technol., Trans. Civil Eng.*, vol. 45, no. 1, pp. 111–123, Mar. 2021.
- [33] K. Yang, M. Bi, Y. Liu, and Y. Zhang, "LSTM-based deep learning model for civil aircraft position and attitude prediction approach," in *Proc. Chin. Control Conf. (CCC)*, Jul. 2019, pp. 8689–8694.
- [34] H. Dai, G. Huang, J. Wang, H. Zeng, and F. Zhou, "Regional VOCs gathering situation intelligent sensing method based on spatial-temporal feature selection," *Atmosphere*, vol. 13, no. 3, p. 483, Mar. 2022.



YONG-CHAO JIN received the master's degree from the Department of Mathematics, North China University of Science and Technology, in 2015. Since 2018, he has been a Teacher with the College of Science, North China University of Science and Technology. His research interests include applied mathematical statistics and big data technology application.



QIAN CAO received the master's degree from the Department of Mathematics, North China University of Science and Technology, in 2022. Since 2023, she has been a Teacher with the Department of Science, North China University of Science and Technology. Her research interests include artificial intelligence algorithm application and big data technology application.



YAN-PENG CAO is currently pursuing the bachelor's degree in intelligent science and technology with the North China University of Science and Technology.



KE-NAN WANG is currently pursuing the bachelor's degree in data science and big data technology with the North China University of Science and Technology.



YUAN ZHOU is currently pursuing the bachelor's degree in computer science and technology with the North China University of Science and Technology.



XI-YIN WANG received the Ph.D. degree in biology from Peking University. He is currently a Professor, a Ph.D. Supervisor, and the Dean of the College of Science, North China University of Science and Technology. He has presided more than a number of national projects, mainly engaged in bioinformatics and genomics research. He was the chief scientist or the research group leader of bioinformatics and comparative genomics analysis in a number of international cooperation projects of plant genome sequencing. He is a member of the National Key Special Committee, the International Cotton Genome Committee, and the Bioinformatics Committee of the Chinese Society of Bioengineering, and the Director of the Center for Genomics and Computational Biology.

...