IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 8 December 2023, accepted 23 December 2023, date of publication 25 December 2023,
date of current version 10 January 2024.

Digital Object Identifier 10.1109/ACCESS.2023.3347403

RESEARCH ARTICLE

Models for COVID-19 Data Prediction Based on Improved LSTM-ARIMA Algorithms

YONG-CHAO JIN[1], QIAN CAO[1], QIAN SUN[1], YE LIN[1], DONG-MEI LIU[1], SHAN-YU[1], CHEN-XI WANG[1], XIAO-LING WANG[1], AND XI-YIN WANG[1,2]

[1]College of Science, North China University of Science and Technology, Tangshan 063210, China
[2]Hebei Key Laboratory of Data Science and Application, Tangshan 063210, China

Corresponding authors: Xi-Yin Wang (wangxiyin@vip.sina.com), Ye Lin (linye315317@163.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 Tangshan Municipal Funding for Talented Researcher under Grant 16013601.

ABSTRACT The global repercussions of the COVID-19 pandemic on economies and public health worldwide have been profound. This study aims to examine the developmental trends of the COVID-19 pandemic, establish predictive models, and provide insights for effective control measures against potential future disease outbreaks. Considering the coexistence of both linear and nonlinear factors in COVID-19 data, conventional single-machine learning and traditional forecasting models encounter challenges in accurately predicting pandemic trends. To enhance the precision of COVID-19 pandemic predictions by integrating linear and nonlinear factors, this study proposes three combined forecasting models: CNN-LSTM-ARIMA, TCN-LSTM-ARIMA, and SSA-LSTM-ARIMA. These models leverage the strengths of deep learning in capturing nonlinear factors and the capabilities of the traditional ARIMA model in handling linear factors. Initially, LSTM and ARIMA models are used to model and predict the COVID-19 pandemic in Quebec, Canada. Subsequently, CNN models, TCN models, and the Sparrow Search Algorithm are employed to integrate predictions from the LSTM and ARIMA models. Comparative analyses of the three combined models, it was found that the CNN-LSTM-ARIMA model exhibits the highest predictive accuracy, with an MSE of 7048.26, RMSE of 83.95, MAE of 61.18, MAPE of 0.16, and R^2 of 0.95. To validate the applicability and stability of the CNN-LSTM-ARIMA model in predicting COVID-19 pandemics, Italian COVID-19 pandemic data was employed. The three combined forecasting models are established and evaluated using model evaluation metrics. The results affirm that the CNN-LSTM-ARIMA model remains the optimal choice, underscoring its high stability and suitability for COVID-19 pandemic forecasting endeavors.

INDEX TERMS COVID-19, LSTM, ARIMA, CNN-LSTM-ARIMA, TCN-LSTM-ARIMA, SSA-LSTM-ARIMA.

I. INTRODUCTION

On December 8, 2019, the Chinese government reported the first fatality attributed to what would later be identified as COVID-19, originating from Wuhan pneumonia. Concurrently, a novel coronavirus strain rapidly disseminated globally, with its epicenter in Wuhan, Hubei province, China [1], [2]. The COVID-19 pandemic has wrought considerable damage to the global economy and the health of citizens across various nations. The virus's highly mutable nature posed significant challenges for effective control during its initial outbreak.

Since the onset of the COVID-19 pandemic, predicting its trajectory has become a focal point for scholars worldwide. In the early stages, Zhu Renjie and colleagues achieved noteworthy predictive results by employing the classical infectious disease model, SIR, to forecast the COVID-19 pandemic in seven countries: Italy, South Korea, the United Kingdom, the United States, France, Spain, and Germany [3]. Leonid Kalachev and others utilized the SIQR model to

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

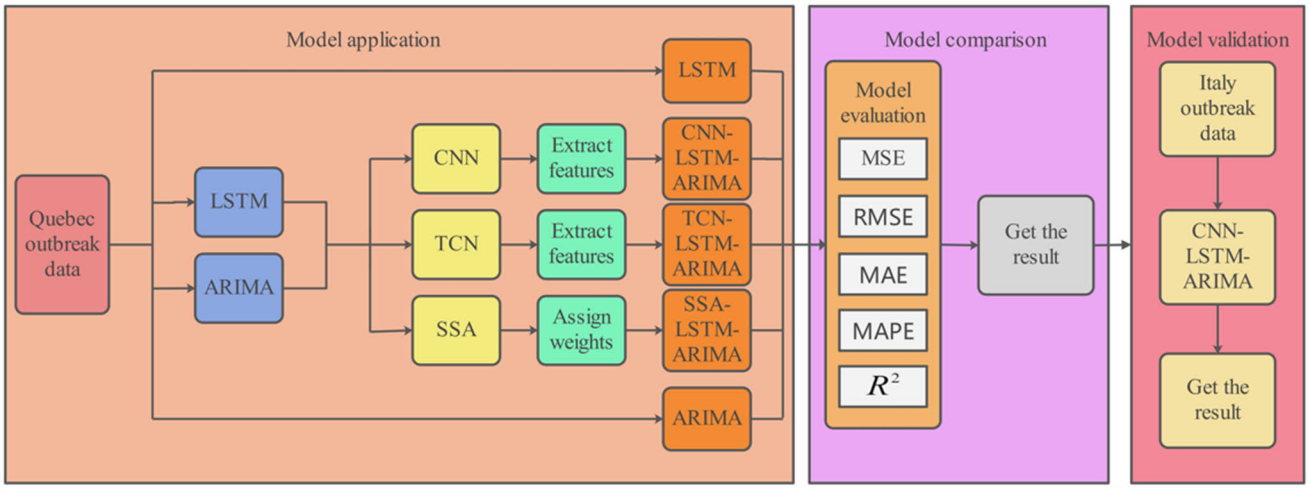

FIGURE 1. Technology roadmap.

forecast the autumn 2020 COVID-19 pandemic in Missoula County, Montana, USA, demonstrating the model's superior predictive accuracy compared to the SIR model [4].

With the ongoing development of deep learning, Jin Weiqiu and colleagues proposed a prediction model known as TCN-GRU-DBN-q-SVM. In this model, TCN, GRU, and DBN are integrated as elements of a hybrid model, with SVM employed to estimate the error of the TCN-GRU-DBN-q predictive sequence. This adaptive model allows adjustments based on data characteristics, ensuring robustness and generalization [5]. Zhang Qing derived optimal parameters for the TCN model by continually adjusting the window size and convolutional kernel size. The TCN epidemic prediction model, constructed using these optimal parameters, achieves highly accurate daily forecasts of the number of cases in the United States [6].

Abdelkader Dairi established a CNN-LSTM hybrid model for predicting the pandemic in seven countries. Comparing its performance with single LSTM and CNN models, the results indicated a significant improvement in the CNN-LSTM model's performance [7]. Chen Honglin introduced an attention mechanism into the CNN-LSTM model to better explore long time-series data features. This model outperformed the CNN-LSTM model when applied to predict the COVID-19 situation in China [8]. Yogesh Gautam trained an LSTM model using early COVID-19 data from Italy and the United States. This model was used to predict the pandemic in Germany, France, Brazil, India, and Nepal, successfully demonstrating the effectiveness of the LSTM model [9]. Shahid et al. employed the Bi-LSTM model to forecast the pandemic in China, and the results showed that the Bi-LSTM model had lower MAE and RMSE compared to the LSTM model, indicating an enhancement in prediction accuracy [10].

To improve the long-term forecasting accuracy of the LSTM model, Wang Peipei and colleagues embedded a rolling update mechanism into the LSTM. This enhanced model's robustness and effectiveness when forecasting pandemics in Russia, Peru, and Iran [11]. Yuan Meng and collaborators optimized the hyperparameters of the LSTM model using Bayesian optimization. They validated the model's ability to enhance the efficiency of LSTM model learning and improve prediction accuracy using COVID-19 data from China [12]. Sarbhan Singh applied the traditional ARIMA model to predict the pandemic in Malaysia. The results showed good predictive performance of the ARIMA model in the test dataset [13]. Machine learning prediction models have also been used for pandemic forecasting. Li Shaoting and colleagues developed the CEEMDAN-XGB&WSD model, where CEEMDAN effectively removes local noise, and WSD supplements historical information. They demonstrated the model's robustness and generalization ability when predicting the pandemic in China using this combined model [14]. Hu Haiwen used a regression decision tree model to predict the pandemic in the United States. Comparing it with linear regression, XGBOOST, SVR, LSTM, and CNN-LSTM models, they confirmed the regression decision tree model's superior performance in pandemic prediction [15]. Jin Yongchao used an ARIMA-LSTM model to predict the pandemic. They found that the predictive accuracy of the ARIMA-LSTM model was significantly higher than that of the ARIMA model and the LSTM model [16].

Although the aforementioned forecasting methods have shown promise in COVID-19 pandemic prediction, they often lack consideration for the linear and nonlinear factors present in pandemic data. Even though some studies have applied the ARIMA-LSTM model, the common approach to combining these two models is still using the LSTM model to correct the residuals of the ARIMA model or applying linear regression to the predictions of both models to obtain weights.

In this study, to better capture the linear and nonlinear factors present in pandemic data, we first modeled and predicted

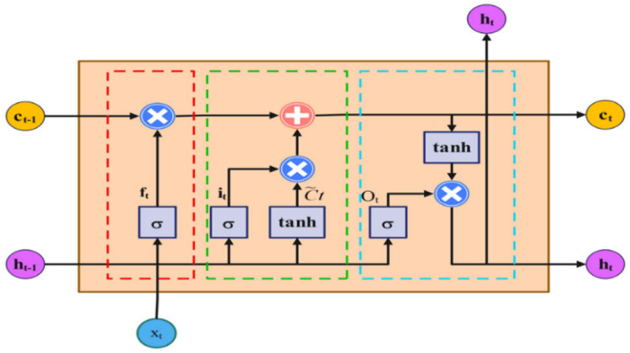

FIGURE 2. LSTM model structure diagram.

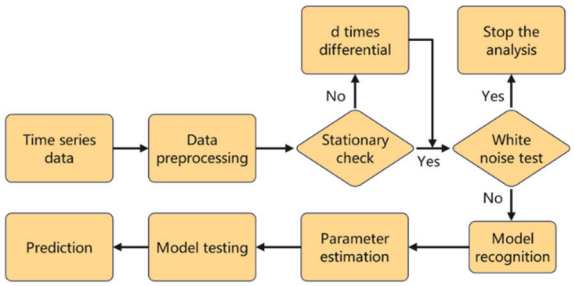

FIGURE 3. ARIMA model modeling forecast flowchart.

the pandemic data using a single deep learning model, LSTM, and the traditional forecasting model, ARIMA. Subsequently, we combined the forecasts from LSTM and ARIMA, effectively utilizing the linear and nonlinear factors within the data. To integrate the predictions from LSTM and ARIMA, we employed one-dimensional convolutional neural networks, temporal convolutional networks, and the Sparrow Search Algorithm. The overall technical roadmap of the article is shown in figure 1.

II. MODEL SELECTION

A. LSTM MODEL

LSTM (Long Short-Term Memory) stands out as a distinctive type of recurrent neural network. Although traditional recurrent neural networks excel in processing data with sequential features, they often encounter challenges such as vanishing gradients and exploding gradients, limiting their effectiveness in capturing extended dependencies within sequential data. The LSTM model improves upon conventional recurrent neural networks by incorporating a state structure and three gate structures: the cell state, forget gate, input gate, and output gate [16]. These enhancements facilitate the dynamic adjustment of self-recurrent weights, effectively mitigating issues related to vanishing and exploding gradients while providing both long-term and short-term memory capabilities. The structure of the LSTM model is depicted in Figure 2.

The following section provides an introduction to the three gate structures of the LSTM model.

1) FORGET GATE

The forget gate reviews the current time step's input information, denoted as x_t, and the previous time step's output information, denoted as h_{t-1}. When $f_t = 0$, the gate discards the read information. Conversely, when $f_t = 1$, it retains the read information. The calculation formula for f_t is as follows [17] and [18]:

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{1}$$

Here, σ signifies the sigmoid activation function, W_f is the weight matrix for the forget gate, and b_f is the bias coefficient.

2) INPUT GATE

This gate determines which new input information to store in the neuron. It initiates by creating a candidate cell state $\tilde{C}_t$, and then, the input gate i_t updates the candidate cell state. The new information is subsequently added to the cell state. The specific formula is as follows [19]:

$$\tilde{C}_t = tanh\left(W_c\left[h_{t-1}, x_t\right] + b_c\right) \tag{2}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{3}$$

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t \tag{4}$$

In the above formula, W_c is the weight matrix for the cell state, b_c is the bias coefficient for the cell state, W_i is the weight matrix for the input gate, and b_i is the bias coefficient for the input gate.

3) OUTPUT GATE

The output gate determines the final output h_t using the cell state. It starts by processing the current input information x_t and the previous output information h_{t-1}. Then, it multiplies these values by the cell state processed by the tanh layer to obtain the final output h_t. The specific formula is as follows [20]:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{5}$$

$$h_t = o_t \times \tanh(C_t) \tag{6}$$

In this formula, W_o is the weight matrix for the output gate, and b_o is the bias coefficient for the output gate.

B. ARIMA MODEL

The Auto Regressive Integrated Moving Average model, commonly known as ARIMA(p, d, q), involves parameters p, d, and q, representing the number of Auto Regressive (AR) terms, the number of Moving Average (MA) terms, and the order of differencing needed to achieve stationarity in the time series. The general form is expressed as follows [21] and [22]:

$$Y_t = c + \sum_{i=1}^{p} r_i y_{t-i} + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \tag{7}$$

In the above formula, c represents the constant term, r_i represents the autocorrelation coefficients, ε_t is the error term. The modeling and forecasting process of the ARIMA model is illustrated in Figure 3.

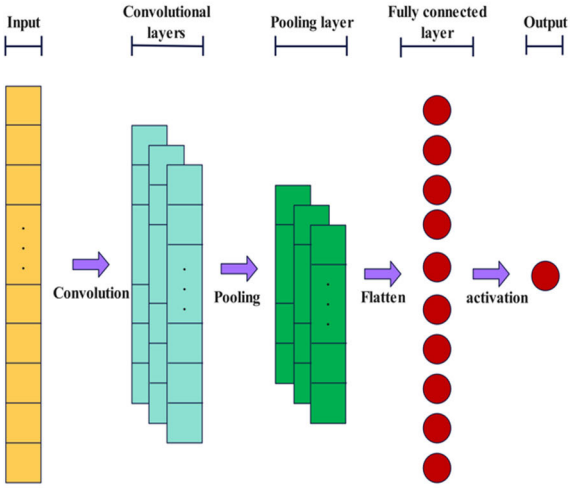

FIGURE 4. 1D-CNN model structure diagram.

C. 1D-CNN MODEL

The 1D-CNN model, which stands for one-dimensional Convolutional Neural Network, is a special type of convolutional neural network. It is named ''one-dimensional'' because the convolutional kernels of 1D-CNN operate only along the sequence of time steps. One-dimensional CNNs excel in processing sequential data and primarily consist of convolutional layers, pooling layers, and fully connected layers [23]. The model structure of the 1D-CNN is illustrated in Figure 4. We will delve into the architecture of the 1D-CNN model, elucidating its convolutional layers, pooling layers, and fully connected layers.

1) CONVOLUTIONAL LAYERS

These layers function by traversing a fixed-size convolutional kernel across the input data, executing convolution operations to discern features within the input. The convolutional kernel essentially constitutes a weight matrix. The computational formula for the convolutional layer is outlined below [24], [25]:

$$x_i^l = \sigma\left(\sum_{j \in M_i} x_j^{l-1} \cdot W_{ij}^l + b_i^l\right) \tag{8}$$

In the above formula, x_i^l is the result of the convolution operation, σ is the activation function, M_i is the input operation, x_j^{l-1} is the region to be convolved, W_{ij}^l is the convolutional kernel, $(\cdot)$ is the convolutional operation, and b_i^l is the bias coefficient of the corresponding convolutional kernel.

2) POOLING LAYER

Pooling layer, also known as sampling layer, usually follows the convolutional layer immediately. Its function is to reduce the dimensionality of input data to reduce computational costs. The commonly used pooling operation is the maximum pooling operation, as shown in Figure 5 [25].

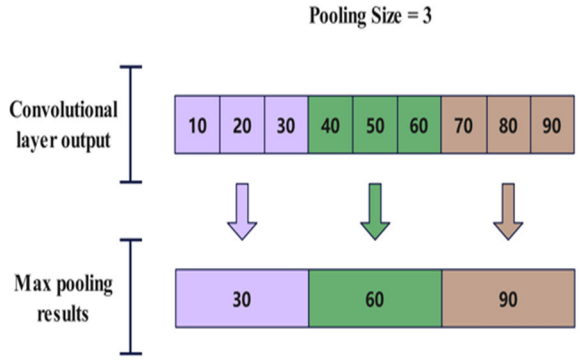

FIGURE 5. Schematic diagram of the max pooling operation.

3) FULLY CONNECTED LAYERS

These layers play a pivotal role in converting the feature maps discerned by the convolutional and pooling layers into the ultimate output. The computational formula for a fully connected layer is articulated below [26]:

$$y = \sigma(W \cdot x + b) \tag{9}$$

In the given formula, σ signifies the activation function, W stands for the weight matrix associated with the fully connected layer, and b denotes the bias term pertinent to the fully connected layer.

D. TCN MODEL

The Temporal Convolutional Network (TCN) model is a special type of convolutional neural network designed for parallel processing of sequential data. This feature reduces the time required for modeling and forecasting. Furthermore, TCN models can work with sequences of arbitrary lengths, and they can directly map input sequences to output sequences of the same length.

Within TCN models, dilated causal convolutions and residual connections are integrated into the network structure. Let's delve into the elements of the TCN model, focusing on dilated causal convolutions and residual connections.

1) DILATED CAUSAL CONVOLUTIONS

Dilated causal convolutions, a key component of TCN models, apply a dilation factor to causal convolution, extending the receptive field of the convolutional network [27]. Incorporating distinct dilation factors across various convolutional layers enables the model to grasp dependencies spanning diverse time scales. The configuration of dilated causal convolutions is depicted in Figure 6.

Residual connection: Residual connection is used to solve the potential gradient vanishing problem in complex neural networks, thereby improving the training efficiency and accuracy of the model [28]. The structure of the residual block includes two layers of convolution, nonlinear mapping, Dropout, and WeightNorm, where Dropout and WeightNorm are used to regularize the neural network. The structure of the residual block is shown in Figure 7.

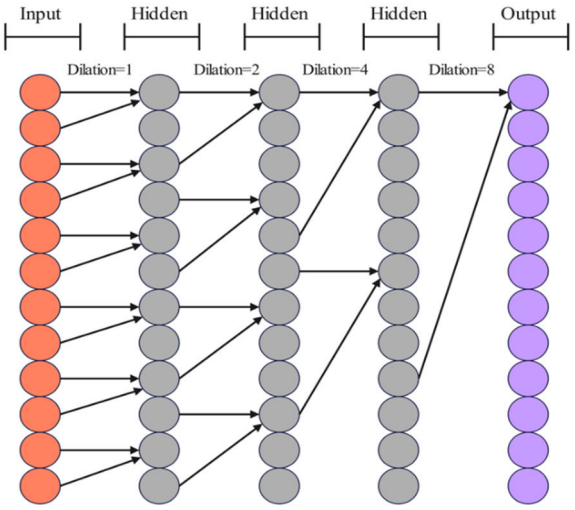

FIGURE 6. Diagram of the expansive causal convolutional structure.

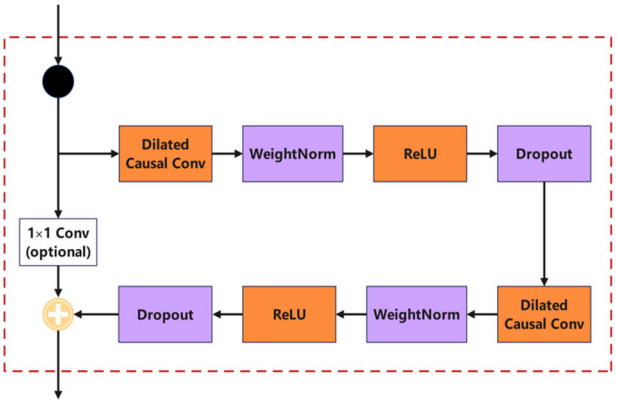

FIGURE 7. Residual block structure diagram.

E. SSA MODEL

The Sparrow Search Algorithm (SSA) is a heuristic algorithm inspired by sparrows' foraging behavior, exhibits exceptional global search capabilities. This algorithm classifies sparrows into three roles: finders, followers, and sentinels [27]. Finders bear the responsibility of foraging for food and supplying information about foraging areas to the population. Consequently, finders boast the broadest search range within the entire population. The formula governing the update of finders' positions is outlined below [28], [29]:

$$X_{ij}^{t+1} = \begin{cases} X_{ij}^{t} \cdot \exp\left(\dfrac{-i}{\alpha \cdot iter_{max}}\right) & R_2 < ST \\ X_{ij}^{t} + Q \cdot L & R_2 \geq ST \end{cases} \tag{10}$$

In the formula above, the variables and terms are defined as follows: X_{ij}^{t} represents the position of the i-th sparrow in the j-th dimension at the t-th iteration. α is a random number in the range (0, 1], $iter_{max}$ is the maximum number of iterations, Q is a random number following a normal distribution, L is a constant matrix of size 1×d, where d is the dimensionality, R_2 is a warning value, and it lies within the range [0,1], ST is a safety threshold, and it falls within the range [0.5,1]. When $R_2 < ST$, it indicates that there are no predators within the current foraging area, allowing the finders to expand their search range. When $R_2 \geq ST$, it implies the presence of predators within the current foraging area, and all sparrows must immediately move to a safe zone for foraging.

Joiners compete for more food resources by continuously monitoring the behavior of discoverers, in order to improve their predation rate. The formula for updating the position of the enrollee is as follows [30]:

$$X_{ij}^{t+1} = \begin{cases} Q \cdot \exp\left(\dfrac{X_{worst}^{t} - X_{ij}^{t}}{i^2}\right) & i > n/2 \\ X_{p}^{t+1} + \left|X_{ij}^{t} - X_{p}^{t+1}\right| \cdot A^{+} \cdot L & otherwise \end{cases} \tag{11}$$

In the above formula, n is the number of sparrows, and the parameter Q is the same as formula (10), X_{worst}^{t} is the position with the worst global fitness during the t-th iteration, X_{p}^{t+1} is the location where the fitness of the discoverer is optimal during the t+1st iteration, $A^{+} = A^{T}(AA^{T})^{-1}$, where A is a matrix of size 1 × d, elements are random values of 1 or -1, and parameter L is the same as formula (10).

Guardians are randomly selected from the sparrow population, typically selecting a 10% to 20% proportion of sparrows as guarders. They always remain vigilant to the surrounding environment during the foraging process. Once a predator is discovered, the sparrow population will immediately engage in anti predatory behavior. The formula for updating the position of the vigilant is as follows [31]:

$$X_{ij}^{t+1} = \begin{cases} X_{best}^{t} + \beta \cdot \left|X_{ij}^{t} - X_{best}^{t}\right| & f_i > f_g \\ X_{ij}^{t} + k \cdot \left(\dfrac{|X_{ij}^{t} - X_{worst}^{t}|}{(f_i - f_w) + \varepsilon}\right) & f_i = f_g \end{cases} \tag{12}$$

In the formula above, X_{best}^{t} denotes the globally optimal position during the t-th iteration, β represents a random number following a standard normal distribution, k is a random number within the range [-1,1], f_i is the fitness value of the current sparrow, f_w is the current global best fitness value, f_g denotes the current global worst-case fitness value, and ε is a minimal constant.

F. EVALUATION INDICATIONS

In order to compare the predictive effects of different prediction models more intuitively, this study used Mean Squared Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error(MAPE), and Coefficient of determination(R^2). The calculation formulas for each evaluation index are as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{13}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{14}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{15}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \quad (16)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \quad (17)$$

In the above formulas (13) - (17), n is the total sample size, $\hat{y}_i$ is the predicted value of the model, y_i is the true value, $\bar{y}$ is the average of the true value [32].

III. MODEL APPLICATION

A. SELECTION OF DATASETS

This study uses the daily newly diagnosed number of COVID-19 from March 1, 2020 to October 21, 2021 in Quebec, Canada, and the daily newly diagnosed number of COVID-19 from February 21, 2020 to October 12, 2021 in Italy for modeling, prediction and analysis. The data is sourced from the Johns Hopkins University website. Upon acquiring the data, the outliers were checked by drawing box plots and time series diagrams, and the mean substitution treatment was performed on the outliers.

B. MODEL CONSTRUCTION SOFTWARE AND PACKAGES

The programming language used for implementing deep learning models in this study is Python 3.11.6. TensorFlow is employed to build deep learning models, while NumPy is utilized for mathematical computations and array operations. The Pandas package is used for data processing and analysis. For plotting and ARIMA model implementation in this study, the programming language is R 4.3.1. The ggplot2 package is used for data visualization, t-series for time series stationarity testing, and forecast for the automatic order determination of the ARIMA model.

C. LSTM MODEL

The LSTM model was constructed by using the daily newly diagnosed number of COVID-19 patients from March 1, 2020 to October 21, 2021 in Quebec, Canada. Both the input and output of the model are the daily new confirmed COVID-19 cases. The first two thirds of the data were used as training sets for the training model. The last one-third of the data is used as the test set to evaluate the model's generalization ability. The parameter settings of the LSTM model are shown in Table 1.

Based on the above parameters and data, an LSTM model is constructed, and the time series diagram comparing the predicted and true values of the model is shown in Figure 8. The evaluation indicators for the predictive performance of the LSTM model on the test set are shown in Table 2.

D. ARIMA MODEL

In this study, the ARIMA model was constructed by using the daily newly diagnosed number of COVID-19 patients from March 1, 2020 to October 21, 2021 in Quebec, Canada. Firstly, the original sequence is subjected to differential processing. After first order differential processing, the sequence becomes stationary and non white noise. Determine the final model as ARIMA (2, 1, 0) using the BIC minimum criterion, and the residual of the model is white noise. The comparison of ARIMA model fitting values and true values in time series is shown in Figure 9. The evaluation indicators for the prediction effect of the ARIMA model are shown in Table 3.

TABLE 1. LSTM model parameters.

parameter	value
Hidden Units of the first layer	128
Hidden Units of the second layer	64
Hidden Units of the third layer	32
Epochs	200
Dropout Rate	0.1
Batch Size	32
Optimizer	Adam

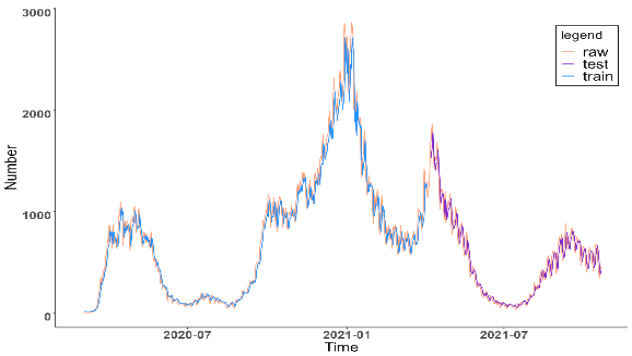

FIGURE 8. LSTM prediction results in Quebec.

TABLE 2. Evaluation index of the LSTM model test set.

Evaluation indicators	Value
MSE	8947.83
RMSE	94.59
MAE	70.89
MAPE	0.19
R^2	0.94

E. COMBINATION MODEL

1) CNN-LSTM-ARIMA MODEL

After obtaining the predicted values of LSTM and ARIMA models, in order to integrate linear (ARIMA) and nonlinear (LSTM) factors, the predicted values of LSTM, ARIMA and the daily number of newly diagnosed COVID-19 patients in Quebec, Canada, from March 1, 2020 to October 21, 2021 were used as variables to build the CNN-LSTM-ARIMA model. Among them, the first 2/3 of the data is used as the training set, and the last 1/3 of the data is used as the test set. The parameter settings of the CNN model are shown in Table 4.

After configuring the parameters, use the training set to train the CNN model, and then use the model to make predictions in the test set. The comparison time series diagram

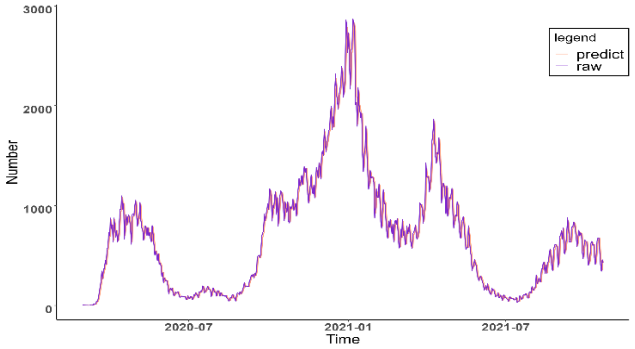

FIGURE 9. ARIMA prediction results in Quebec.

TABLE 3. ARIMA model evaluation index.

Evaluation indicators	Value
MSE	9140.14
RMSE	95.60
MAE	68.03
MAPE	0.17
R^2	0.94

TABLE 4. CNN model parameters.

parameter	value
kernel Size	3
Number of Filters	64
Pool Size	1
Epochs	200
Batch Size	32
Optimizer	Adam

between the predicted values and the actual values of the CNN-LSTM-ARIMA model is shown in Figure 10.

The evaluation indicators for the prediction performance of the CNN LSTM ARIMA model on the test set are shown in Table 5.

Upon comparing the performance metrics of the CNN-LSTM-ARIMA model in Table 5 with those of the LSTM and ARIMA models, it is evident that the predictive accuracy of the combined CNN-LSTM-ARIMA model surpasses that of both the LSTM and ARIMA models significantly.

2) TCN-LSTM-ARIMA MODEL

TCN-LSTM-ARIMA model and CNN-LSTM-ARIMA model have the same principle. They also use the predicted value of LSTM, the predicted value of ARIMA and the daily newly diagnosed number of COVID-19 in Quebec, Canada, from March 1, 2020 to October 21, 2021 as variables to build the TCN-LSTM-ARIMA model. The partitioning of the training and testing sets is the same as the CNN LSTM ARIMA model. The parameter settings of the TCN model are shown in Table 6.

The time series diagram of the TCN-LSTM-ARIMA model's predicted and true values is shown in Figure 11.

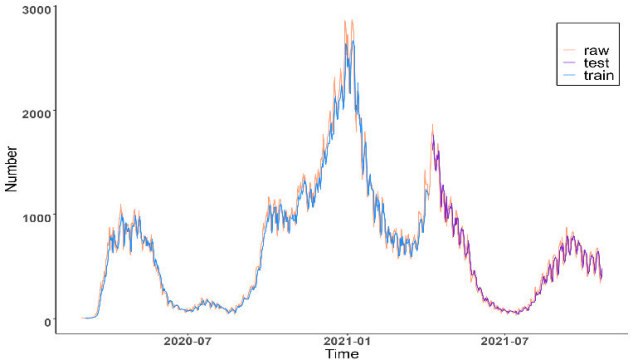

FIGURE 10. CNN-LSTM-ARIMA prediction results in Quebec.

TABLE 5. Evaluation index of the CNN-LSTM-ARIMA model test set.

Evaluation indicators	Value
MSE	7048.26
RMSE	83.95
MAE	61.18
MAPE	0.16
R^2	0.95

TABLE 6. TCN model parameters.

parameter	value
kernel Size	3
Filters of the first layer	128
Filters of the second layer	64
Filters of the third layer	32
Learning Rate	0.001
Dropout Rate	0.2
Epochs	300
Batch Size	32
Optimizer	Adam

The evaluation indicators for the predictive performance of the TCN-LSTM-ARIMA model on the test set are shown in Table 7.

3) SSA-LSTM-ARIMA MODEL

In the SSA-LSTM-ARIMA model, use the sparrow search algorithm to search for the weight W_1 of the LSTM predicted value in formula (18) and weight W_2 of the ARIMA predicted values. Integrate linear and nonlinear factors into the predicted value $\hat{y}$ the RMSE of is the smallest. About to $\hat{y}$ the RMSE is used as the fitness value for the sparrow search algorithm.

$$\hat{y} = W_1 x_{LSTM} + W_2 x_{ARIMA} \tag{18}$$

The parameter settings of the sparrow search algorithm are shown in Table 8.

After setting the parameters of the sparrow search algorithm, use the sparrow search algorithm to search for the weight W_1 and W_2 between the intervals [0,1]. Finally, the

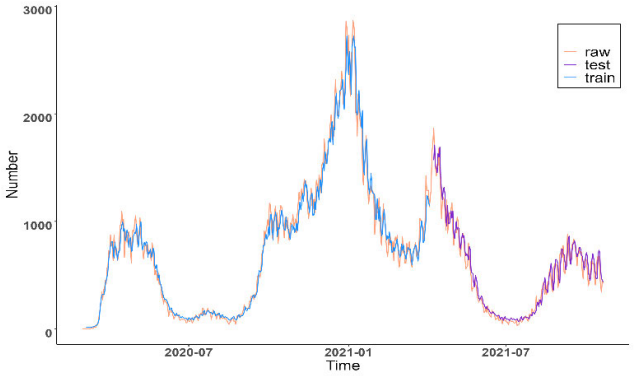

FIGURE 11. TCN-LSTM-ARIMA prediction results in Quebec.

TABLE 7. Evaluation index of the TCN-LSTM-ARIMA model test set.

Evaluation indicators	Value
MSE	8203.44
RMSE	90.57
MAE	66.81
MAPE	0.20
R^2	0.94

TABLE 8. SSA parameters.

parameter	value
Proportion of Producer	0.7
Proportion of Vigilantes	0.2
Security thresholds	0.8
Number of sparrows	100
Maximum number of iterations	500

optimal result is obtained as shown in formula (19).

$$\hat{y} = 0.601x_{LSTM} + 0.412x_{ARIMA} \tag{19}$$

The prediction results of the LSTM and ARIMA models are incorporated into formula (16) to obtain a time series diagram of the comparison between the predicted and true values of the SSA-LSTM-ARIMA model, as shown in Figure 12.

The evaluation indicators for the prediction performance of the SSA-LSTM-ARIMA model on the test set are shown in Table 9.

F. COMPARISON OF COMBINED MODELS

Based on the three evaluation indicators selected in this study, the prediction performance of the three combination models was compared. Among them, the CNN-LSTM-ARIMA model had the best prediction performance, followed by the TCN-LSTN-ARIMA model, and the SSA-LSTM-ARIMA model had the worst prediction performance. However, the prediction performance of these three combination models was better than that of the LSTM model and ARIMA model.

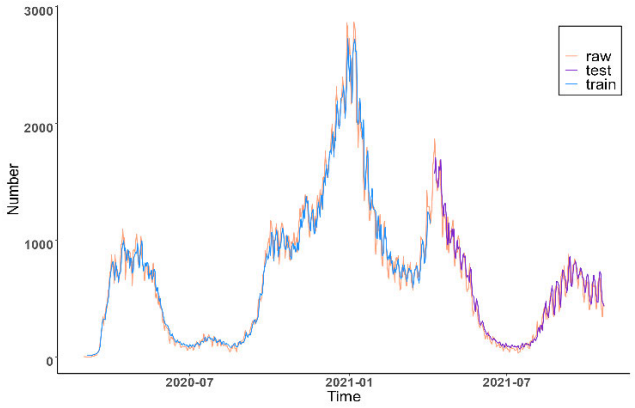

FIGURE 12. SAA-LSTM-ARIMA prediction results in Quebec.

TABLE 9. Evaluation index of the SSA-LSTM-ARIMA model test set.

Evaluation indicators	Value
MSE	8815.50
RMSE	93.89
MAE	68.50
MAPE	0.18
R^2	0.95

IV. MODEL VALIDATION

In order to verify the applicability and stability of CNN-LSTM-ARIMA model in the prediction of COVID-19, this study uses the daily newly confirmed number of COVID-19 patients in Italy from February 21, 2020 to October 12, 2021 to build the above five models.

The relevant parameters of the LSTM model and the division method of the training and testing sets are the same as above. The ARIMA model is determined as ARIMA (2, 1, 0) using the BIC minimum criterion. The time series diagrams comparing the fitted and true values of the LSTM and ARIMA models are plotted as shown in Figures 13 and 14.

After the prediction results of LSTM and ARIMA models were obtained, CNN-LSTM-ARIMA and TCN-LSTM-ARIMA models were constructed by using TCN and CNN models respectively, taking the prediction values of LSTM, ARIMA and the daily newly diagnosed number of COVID-19 patients in Italy from February 21, 2020 to October 12, 2021 as variables. Draw a time series diagram comparing the predicted values of the model with the actual values, as shown in Figures 15 and 16.

Finally, using the sparrow search algorithm, search for the optimal weights of LSTM and ARIMA models, and obtain the optimal results as shown in formula (20). Among them, the sparrow search algorithm parameter settings are the same as above.

$$\hat{y} = 0.752x_{LSTM} + 0.232x_{ARIMA} \tag{20}$$

Draw a time series diagram comparing the predicted values and actual values of the SSA-LSTM-ARIMA model, as shown in Figure 17.

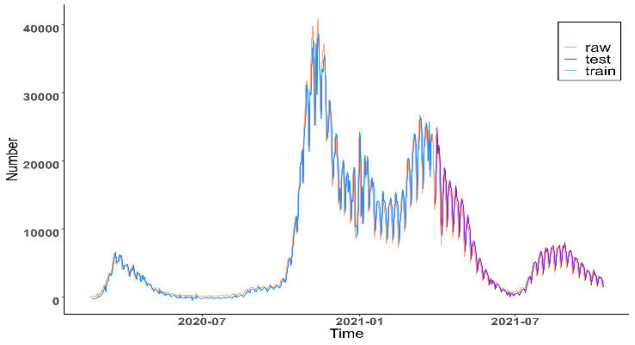

FIGURE 13. LSTM prediction results in Italy.

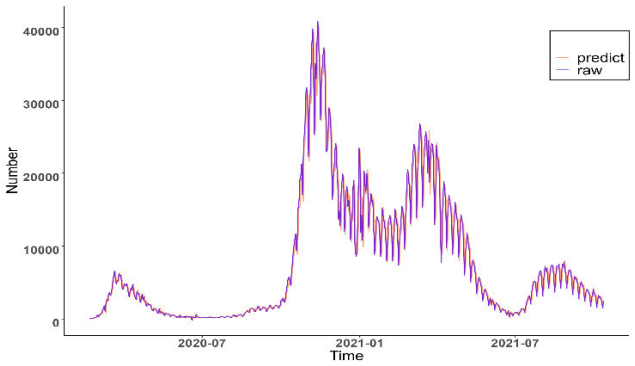

FIGURE 14. ARIMA prediction results in Italy.

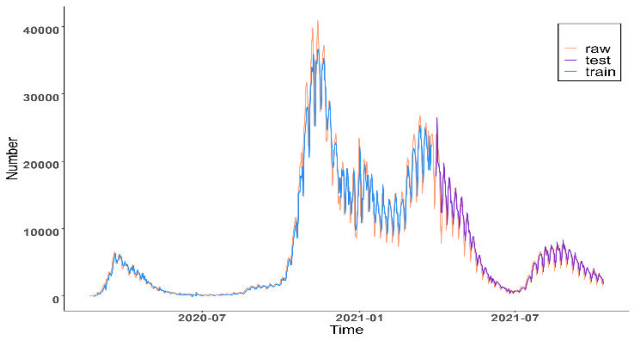

FIGURE 15. CNN-LSTM-ARIMA prediction results in Italy.

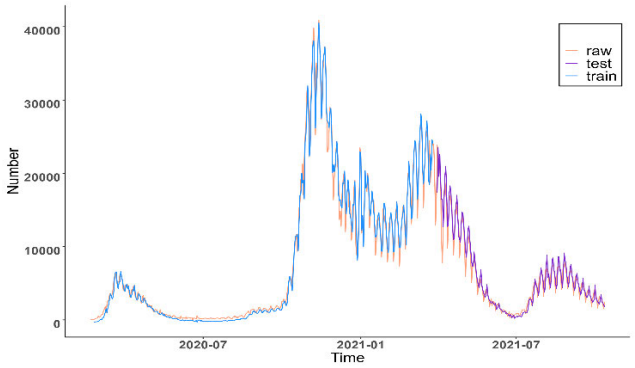

FIGURE 16. TCN-LSTM-ARIMA prediction results in Italy.

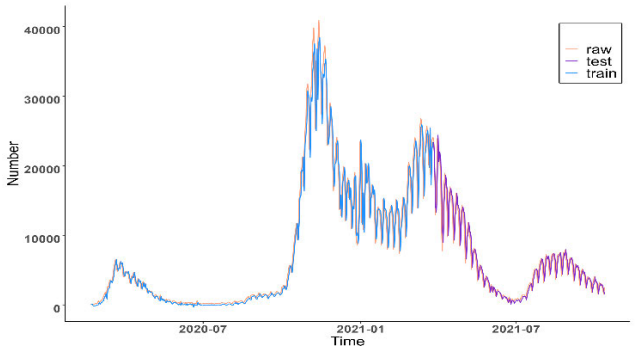

FIGURE 17. SSA-LSTM-ARIMA prediction results in Italy.

TABLE 10. Model evaluation index.

Model	MSE	RMSE	MAE	MAPE	R^2
ARIMA	2490569	1578.15	1022.84	0.23	0.90
LSTM	2497026	1580.20	1069.28	0.25	0.89
CNN-LSTM-ARIMA	1766342	1329.04	828.51	0.21	0.92
TCN-LSTM-ARIMA	1792313	1338.77	876.43	0.21	0.92
SSA-LSTM-ARIMA	2320461	1523.31	1032.01	0.25	0.91

After completing the model modeling and prediction, calculate the evaluation indicators of the above five models in the prediction of COVID-19 in Italy, as shown in Table 10.

Table 10 shows that the prediction accuracy of CNN-LSTM-ARIMA model is the best among the five models, which further verifies the applicability and stability of CNN-LSTM-ARIMA model in the prediction of COVID-19.

V. RESULTS AND DISCUSSION

In this study, three combined models were developed: CNN-LSTM-ARIMA, TCN-LSTM-ARIMA, and SSA-LSTM-ARIMA, aimed at enhancing the precision of LSTM and ARIMA models in predicting the COVID-19 pandemic. Initially, utilizing COVID-19 data from Quebec, Canada, we modeled and predicted the pandemic using ARIMA and LSTM models. The calculated metrics for the ARIMA model were MSE=9140.14, RMSE=95.60, MAE=68.03, MAPE=0.17, R^2 =0.94, and for the LSTM model were MSE=8947.83, RMSE=94.59, MAE=70.89, MAPE=0.19, R^2=0.94.

To enhance predictive accuracy, CNN, TCN, and SSA algorithms were employed to refine the predictions of ARIMA and LSTM models, integrating both linear and nonlinear factors present in the pandemic data. The calculated metrics for the combined models were as follows: CNN-LSTM-ARIMA model's

MSE=7048.26, RMSE=83.95, MAE=61.18, MAPE=0.16, R^2 =0.95; TCN-LSTM-ARIMA model's MSE=8203.44, RMSE=90.57, MAE=66.81, MAPE=0.20, R^2 =0.94; SSA-LSTM-ARIMA model's MSE=8815.50, RMSE=93.89, MAE=68.50, MAPE=0.18, R^2=0.95.

Subsequently, utilizing COVID-19 data from Italy for modeling and prediction, the CNN-LSTM-ARIMA model remained the optimal choice, affirming its applicability and stability in forecasting COVID-19 pandemics.

VI. CONCLUSION AND SUGGESTIONS

The COVID-19 pandemic data often exhibits the complexity of having both linear and nonlinear trends, posing significant challenges for predictive work. Addressing these challenges, we employed the CNN-LSTM-ARIMA model to integrate both nonlinear and linear factors within the epidemic data, thereby enhancing the accuracy of COVID-19 pandemic predictions.

Through the collective efforts of people worldwide, we have overcome the COVID-19 pandemic. However, the outbreak of the pandemic has undoubtedly sounded an alarm for the world to prioritize public health security. The CNN-LSTM-ARIMA predictive model established in this paper can serve as a reference for governments in similar infectious disease prevention and control efforts, aiming to minimize the losses caused by pandemic outbreaks. Additionally, governments should guide the public to strengthen awareness of epidemic prevention and control, emphasizing the importance of disinfection in public spaces.

REFERENCES

[1] F. Ahouz and A. Golabpour, "Predicting the incidence of COVID-19 using data mining," *BMC Public Health*, vol. 21, no. 1, p. 1087, Dec. 2021.
[2] M. E. H. Chowdhury, T. Rahman, A. Khandakar, S. Al-Madeed, S. M. Zughaier, S. A. R. Doi, H. Hassen, and M. T. Islam, "An early warning tool for predicting mortality risk of COVID-19 patients using machine learning," *Cognit. Comput.*, pp. 1–16, Apr. 2021.
[3] R. Zhu, "Prediction of the novel coronavirus pneumonia epidemic based on an improved SIR model and the impact of prevention and control on the development of the epidemic," *J. Shaanxi Norm. Univ.*, vol. 48, no. 3, pp. 33–38, 2020.
[4] L. Kalachev, E. L. Landguth, and J. Graham, "Revisiting classical SIR modelling in light of the COVID-19 pandemic," *Infectious Disease Model.*, vol. 8, no. 1, pp. 72–83, Mar. 2023.
[5] W. Jin, S. Dong, C. Yu, and Q. Luo, "A data-driven hybrid ensemble AI model for COVID-19 infection forecast using multiple neural networks and reinforced learning," *Comput. Biol. Med.*, vol. 146, Jul. 2022, Art. no. 105560.
[6] Q. Zhang, "Research and prediction on the new crown pneumonia epidemic," Shandong Univ., Tech. Rep., 2022.
[7] A. Dairi, F. Harrou, A. Zeroual, M. M. Hittawe, and Y. Sun, "Comparative study of machine learning methods for COVID-19 transmission forecasting," *J. Biomed. Informat.*, vol. 118, Jun. 2021, Art. no. 103791.
[8] H. Chen, "Research on new crown epidemic prediction based on neural network multi-algorithm combination model," North Univ. Nationalities, Tech. Rep., 2023.
[9] Y. Gautam, "Transfer learning for COVID-19 cases and deaths forecast using LSTM network," *ISA Trans.*, vol. 124, pp. 41–56, May 2022.
[10] F. Shahid, A. Zameer, and M. Muneeb, "Predictions for COVID-19 with deep learning models of LSTM, GRU and bi-LSTM," *Chaos, Solitons Fractals*, vol. 140, Nov. 2020, Art. no. 110212.
[11] P. Wang, X. Zheng, G. Ai, D. Liu, and B. Zhu, "Time series prediction for the epidemic trends of COVID-19 using the improved LSTM deep learning method: Case studies in Russia, Peru and Iran," *Chaos, Solitons Fractals*, vol. 140, Nov. 2020, Art. no. 110214.
[12] M. Yuan, "Multi-sequence prediction of COVID-19 based on Bayes-weighted LSTM network," *J. China Jiliang Univ.*, vol. 33, no. 3, pp. 379–387, 2022.
[13] S. Singh, B. M. Sundram, K. Rajendran, K. B. Law, T. Aris, H. Ibrahim, S. C. Dass, and B. Singh Gill, "Forecasting daily confirmed COVID-19 cases in Malaysia using ARIMA models," *J. Infection Developing Countries*, vol. 14, no. 9, pp. 971–976, Sep. 2020.
[14] S. Li and X. Wang, "Research and application of XGBoost model in the prediction of new crown epidemic," *J. Chin. Comput. Syst.*, vol. 42, no. 12, pp. 2465–2472, 2021.
[15] H. Hu, "Influencing factors and prediction analysis of novel coronavirus pneumonia epidemic based on machine learning," Lanzhou Jiaotong Univ., Tech. Rep., 2023.
[16] 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.
[17] X. Chen, X. Ding, X. Wang, Y. Zhao, C. Liu, H. Liu, and K. Chen, "Multi-task data imputation for time-series forecasting in turbomachinery health prognostics," *Machines*, vol. 11, no. 1, p. 18, Dec. 2022.
[18] D. Pan, Z. Song, L. Nie, and B. Wang, "Satellite telemetry data anomaly detection using bi-LSTM prediction based model," in *Proc. IEEE Int. Instrum. Meas. Technol. Conf.*, May 2020, pp. 1–6.
[19] P. Huang and Z. Chen, "Deep learning for nonlinear seismic responses prediction of subway station," *Eng. Struct.*, vol. 244, Oct. 2021, Art. no. 112735.
[20] S. Liu, X. Liu, Q. Lyu, and F. Li, "Comprehensive system based on a DNN and LSTM for predicting sinter composition," *Appl. Soft Comput.*, vol. 95, Oct. 2020, Art. no. 106574.
[21] L. Zhai, K. Yang, B. Hu, and S. Li, "Study on the oil particle contamination forecasting using LSTM network," in *Proc. Prognostics Syst. Health Manag. Conf. (PHM-Qingdao)*, Oct. 2019, pp. 1–6.
[22] Y.-C. Jin, Q. Cao, K.-N. Wang, Y. Zhou, Y.-P. Cao, and X.-Y. Wang, "Prediction of COVID-19 data using improved ARIMA-LSTM hybrid forecast models," *IEEE Access*, vol. 11, pp. 67956–67967, 2023.
[23] F. Hu, M. Zhou, P. Yan, D. Li, W. Lai, K. Bian, and R. Dai, "Identification of mine water inrush using laser-induced fluorescence spectroscopy combined with one-dimensional convolutional neural network," *RSC Adv.*, vol. 9, no. 14, pp. 7673–7679, 2019.
[24] L. Shang, Y. Bao, J. Tang, D. Ma, J. Fu, Y. Zhao, X. Wang, and J. Yin, "A novel polynomial reconstruction algorithm-based 1D convolutional neural network used for transfer learning in Raman spectroscopy application," *J. Raman Spectrosc.*, vol. 53, no. 2, pp. 237–246, Feb. 2022.
[25] R. K. Vankadara, M. Mosses, M. I. H. Siddiqui, K. Ansari, and S. K. Panda, "Ionospheric total electron content forecasting at a low-latitude Indian location using a bi-long short-term memory deep learning approach," *IEEE Trans. Plasma Sci.*, vol. 51, no. 11, pp. 3373–3383, Nov. 2023.
[26] Z. Ma, J. Wang, S. Ye, R. Wang, F. Dong, and Y. Feng, "Real-time indoor thermal comfort prediction in campus buildings driven by deep learning algorithms," *J. Building Eng.*, vol. 78, Nov. 2023, Art. no. 107603.
[27] Y. Ma, C. Xu, H. Wang, R. Wang, S. Liu, and X. Gu, "Model NO_x, SO_2 emissions concentration and thermal efficiency of CFBB based on a hyper-parameter self-optimized broad learning system," *Energies*, vol. 15, no. 20, p. 7700, Oct. 2022.
[28] F. Zhang, S. Deng, H. Zhao, and X. Liu, "A new hybrid method based on sparrow search algorithm optimized extreme learning machine for brittleness evaluation," *J. Appl. Geophys.*, vol. 207, Dec. 2022, Art. no. 104845.
[29] F. S. Gharehchopogh, M. Namazi, L. Ebrahimi, and B. Abdollahzadeh, "Advances in sparrow search algorithm: A comprehensive survey," *Arch. Comput. Methods Eng.*, vol. 30, no. 1, pp. 427–455, Jan. 2023.
[30] M. A. Awadallah, M. A. Al-Betar, I. A. Doush, S. N. Makhadmeh, and G. Al-Naymat, "Recent versions and applications of sparrow search algorithm," *Arch. Comput. Methods Eng.*, pp. 1–28, Feb. 2023.
[31] A. G. Gad, K. M. Sallam, R. K. Chakrabortty, M. J. Ryan, and A. A. Abohany, "An improved binary sparrow search algorithm for feature selection in data classification," *Neural Comput. Appl.*, vol. 34, no. 18, pp. 15705–15752, Sep. 2022.
[32] X. Qiang, M. Aamir, M. Naeem, S. Ali, A. Aslam, and Z. Shao, "Analysis and forecasting COVID-19 outbreak in Pakistan using decomposition and ensemble model," *Comput., Mater. Continua*, vol. 68, no. 1, pp. 841–856, 2021.

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 School 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.

QIAN SUN is currently pursuing the bachelor's degree in data science and big data technology with the North China University of Science and Technology.

YE LIN is currently pursuing the degree in data science and artificial intelligence with the School of Science, North China University of Science and Technology, Tangshan, China. He is also an Engineer with the North China University of Science and Technology. His research interests include data analytics machine learning and deep learning models.

DONG-MEI LIU was born in 1986. She received the Ph.D. degree. She is currently an Associate Professor. Her research interests include the interfacial properties of composite materials, dynamics of public opinion dissemination in complex networks, network pharmacology, virtual screening, and related fields.

SHAN-YU received the master's degree from the Department of Mathematics, North China University Science and Technology, in 2021. Since 2021, she has been a Teacher with the School of Science, North China University of Science and Technology. Her research interest includes applied mathematical statistics.

CHEN-XI WANG received the Bachelor of Science degree, in 2018. She is currently pursuing the degree in mathematics with the College of Sciences, North China University of Science and Technology. Her research interest includes the prediction and prevention strategy of infectious diseases.

XIAO-LING WANG received the Bachelor of Science degree in mathematics and applied mathematics from Tangshan Normal University, in 2021. She is currently pursuing the Master of Science degree in mathematics with a concentration in mathematical statistics 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 Doctoral Supervisor, the Director of the Center for Genomics and Computational Biology, and the Dean of the School of Science, North China University of Science and Technology. He has presided over 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.