DOI: https://doi.org/10.26565/2304-6201-2024-64-06

УДК (UDC) 004.891.2

Kachanov Stanislav *PhD student, Department of Theoretical and Applied Computer Science V.N. Karazin Kharkiv National University, Svobody Sq 4, Kharkiv, Ukraine,61022*
e-mail: staskachanov2000@gmail.com
https://orcid.org/0009-0002-6938-6717

Chen Guoxin *Master's student, Department of Theoretical and Applied Computer Science V.N. Karazin Kharkiv National University, Svobody Sq 4, Kharkiv, Ukraine,61022*
e-mail: guoxin.chen@student.karazin.ua;
https://orcid.org/0009-0004-0502-3735

Morozova Anastasiia *PhD, Associate Professor, Department of Theoretical and Applied Computer Science V.N. Karazin Kharkiv National University, Svobody Sq 4, Kharkiv, Ukraine,61022*
e-mail: a.morozova@karazin.ua;
https://orcid.org/0000-0003-2143-7992

Rukkas Kyrylo *DSc, Associate Professor, Full Professor, Department of Theoretical and Applied Computer Science V.N. Karazin Kharkiv National University, Svobody Sq 4, Kharkiv, Ukraine,61022*
e-mail: rukkas@karazin.ua;
https://orcid.org/0000-0002-7614-0793

Prediction of the dynamics COVID19 epidemic process of using the Lasso regression model

This study integrates machine learning and deep learning methods to predict the COVID-19 pandemic.
Relevance. The global outbreak of the COVID-19 pandemic has had a profound impact on public health systems and socio-economic structures worldwide, highlighting the urgent need for effective forecasting tools to aid decision-making. The work is devoted to the development of multi-model framework for epidemic forecasting by integrating advanced methods of mathematical modeling and forecasting theory.
Goal. The purpose of the work was to analyze methods and algorithms for cumulative prediction of COVID-19 cases in order to provide scientific support for public health decision-making.
Research methods. The research methods are based on modern theories of mathematical modeling, artificial intelligence, epidemiological diagnostics, and forecasting theory, namely: Lasso regression, Long Short-Term Memory (LSTM) networks, and LSTM-Attention models. The research details the processes of data preprocessing, model training, evaluation, and visualization to maintain generalization and adaptability in the dynamic pandemic scenario.
The results. The application of Lasso regression model Long Short-Term Memory (LSTM) network for cumulative prediction of COVID-19 cases was investigated to provide scientific support for decision-making. The research details the processes of data preprocessing, model training, evaluation, and visualization to maintain generalization and adaptability in the dynamic pandemic scenario. Additionally, a Multi-Scale LSTM-Attention (MSLA) model was proposed to extract multi-period features from input sequences. These features are critical for addressing data non-stationary.
Conclusions. The task of developing a comprehensive multi-model COVID-19 prediction system by integrating machine learning and deep learning techniques was solved. The system combines Lasso regression, Long Short-Term Memory (LSTM) networks, and a novel Multi-Scale Cumulative Infection Prediction model based on an attention mechanism (MSLA), significantly enhancing prediction accuracy and reliability.

Keywords: *prediction, neural network, Lasso Regression Model, LSTM, COVID-19 prediction, Machine Learning, Deep Learning, Multi-scale model, database.*

Як цитувати: Kachanov S., Chen G., Morozova A., Rukkas K. Prediction of the dynamics COVID19 epidemic process of using the Lasso regression model. *Вісник Харківського національного університету імені В. Н. Каразіна, серія Математичне моделювання. Інформаційні технології. Автоматизовані системи управління*. 2024. вип. 64. С.54-65. https://doi.org/10.26565/2304-6201-2024-64-06

How to quote: S. Kachanov, G. Chen, A. Morozova, K. Rukkas, "Prediction of the dynamics COVID19 epidemic process of using the Lasso regression model", *Bulletin of V. N. Karazin Kharkiv National University, series Mathematical modelling. Information technology. Automated control systems*, vol. 64, pp. 54-65, 2024. https://doi.org/10.26565/2304-6201-2024-64-06

1 Introduction

The global outbreak of the COVID-19 pandemic has had a profound impact on public health systems and socio-economic structures worldwide, highlighting the urgent need for effective forecasting tools to aid decision-making [1]. As powerful data analysis tools, machine learning and deep learning have demonstrated significant potential in epidemic forecasting [2]. This study aims to construct a multi-model ensemble framework for epidemic forecasting by integrating methods such as Lasso regression, Long Short-Term Memory (LSTM) networks, and LSTM-Attention models. The framework is designed to capture complex patterns and temporal dependencies in epidemic data, providing more accurate prediction results to support public health decision-making.

Based on epidemic data from China, this research first preprocesses and engineers a large volume of data to ensure data quality and extract meaningful information. Subsequently, it employs Lasso regression as a machine learning method, alongside LSTM and other deep learning models, to predict epidemic trends. By comparing the performance of different models, this study adopts ensemble techniques to improve prediction accuracy and robustness. The paper details the processes of model training, evaluation, and result visualization and explores optimizing predictive performance by adjusting model parameters and architectures. Through this study, we aim to provide a robust forecasting tool for the public health sector, enabling a scientific and precise response to future public health challenges.

2 Current Research on Prediction Methods

2.1. Traditional Prediction Methods

Traditional forecasting techniques primarily include the moving average method, exponential smoothing, and the Autoregressive Integrated Moving Average (ARIMA) model. The moving average method [3] predicts future infection counts by calculating the arithmetic mean of historical data. While straightforward, it has limitations in capturing dynamic epidemic trends. Exponential smoothing assigns higher weights to more recent data points, making it suitable for handling datasets with trends and seasonal variations. The ARIMA model [4] integrates autoregressive, differencing, and moving average techniques, making it particularly effective for analyzing non-stationary time series data. These methods are generally employed for short-term forecasting and are often used in combination to enhance prediction accuracy.

Despite advancements in technology and the increasing popularity of modern methods such as neural networks and machine learning, traditional methods remain valuable. These approaches continue to provide insights for predicting COVID-19 infection counts, especially in scenarios with limited data or low complexity requirements.

2.2. Machine Learning-Based Prediction Methods

Machine learning-based methods for predicting COVID-19 infection counts have gained increasing attention in recent years due to advancements in data science technologies. Machine learning methods exhibit significant advantages in handling complex and nonlinear data compared to traditional statistical methods. With their strong fitting capabilities [5], these methods can capture intricate patterns in the spread of epidemics. They allow for the integration of various input variables, such as meteorological data, holidays, and population mobility, to improve prediction accuracy. Furthermore, machine learning models can automatically adjust parameters based on new data, adapting to changes in infection trends, and generally provide higher prediction accuracy compared to traditional approaches.

Although machine learning algorithms can handle large datasets and automatically learn patterns, they are typically restricted to panel data, limiting their ability to extract trends, causal relationships, and long-term dependencies from time series data. Model accuracy requires further improvement.

2.3. Deep Learning-Based Prediction Methods

Deep learning methods have shown significant potential in predicting COVID-19 infection and mortality rates [6], particularly in handling large datasets and capturing complex patterns. As epidemic

data often take the form of time series, various time series neural network models have been widely explored.

Long Short-Term Memory (LSTM) networks, a special type of Recurrent Neural Network (RNN), effectively address the long-term dependency challenges in time series data. In COVID-19 infection prediction, LSTMs capture trends and periodic features in time series, achieving accurate forecasts [7–8].

Convolutional Neural Networks (CNNs), primarily used for image recognition, can also extract local features from time series data in infection prediction. Combining CNNs with LSTMs further enhances predictive capabilities [9].

These methods can be summarized as multivariate time series prediction problems with different feature extraction techniques. However, their performance on data with varying periodicities remains suboptimal.

Therefore, the task of prediction of the dynamics covid19 epidemic process of using the Lasso regression model is a relevant problem.

3 Relevant Technical Principles

This paper focuses on the prediction of COVID-19 data, discussing the research background and current state of the field. Different machine learning and deep learning prediction models are utilized, and various ablation experiments are conducted based on real-world production scenarios.

3.1. LSTM Principle

Long Short-Term Memory (LSTM) networks [10] are a special type of Recurrent Neural Network (RNN) designed to address the gradient vanishing or explosion problems encountered by traditional RNNs when processing long sequences of data. LSTMs introduce three key gating mechanisms—the Forget Gate, Input Gate, and Output Gate—whose calculation formulas are given by (3.1), (3.2), (3.3), and (3.4).

$$f_t = \sigma\left(W_f h_{t-1} + U_f x_t + b_f\right) \tag{3.1}$$

$$i_t = \sigma(W_i h_{t-1} + U_i x_t + b_i) \tag{3.2}$$

$$\tilde{C}_t = tanh(W_a h_{t-1} + U_a x_t + b_a) \tag{3.3}$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \tag{3.4}$$

These three gates effectively maintain and update the internal states and outputs of the network, allowing LSTMs to capture long-term dependencies in time series data. The detailed structure of the LSTM is shown in Figure 1.

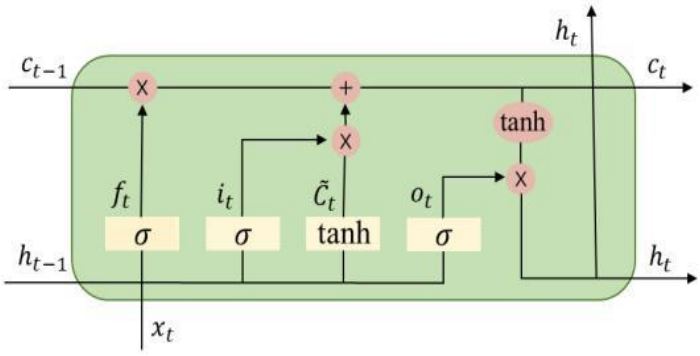

Figure 1. LSTM Structure Diagram

The Input Gate is responsible for deciding how much new information should be added to the cell state at the current time step. The Forget Gate determines what information should be discarded from the cell state at the current time step. The introduction of these two gates allows LSTM to selectively retain or forget information at different time points in the sequence.

3.2. Attention Mechanism

Figure 2 shows the architecture of the encoder part of a Transformer model [11]. The core of this architecture lies in its Multi-Head Attention mechanism, which allows the model to focus on different pieces of information when processing sequence data, thereby capturing richer contextual relationships.

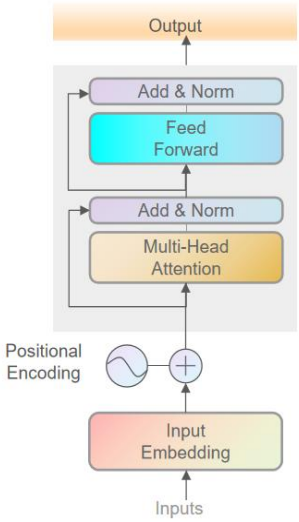

Figure 2. Attention Mechanism Diagram

The input data is first transformed into embedding vectors through the Input Embedding layer. This process maps the features in the sequence to a continuous vector space to capture semantic relationships between different time points. Subsequently, Positional Encoding is added to the embedding vectors to provide positional information for each element in the sequence, as the Transformer model itself lacks a sequence processing mechanism. Next, the data is processed through the Multi-Head Attention mechanism, where multiple attention heads work in parallel. Each head independently calculates attention weights, which are then combined to obtain a richer representation.

Applying the self-attention mechanism to the output of each time step in an LSTM (Long Short-Term Memory) network can further improve the model's performance, especially when handling long sequence data.

3.3. Lasso Regression

Lasso regression is a widely used regression method in machine learning [12]. It introduces an L1 regularization term to promote sparsity in the model coefficients, enabling automatic feature selection. The basic principle of Lasso regression is to add a regularization term to the least squares objective function, as shown in the minimization expression (3.5).

$$||Y - X\beta||^2 + \lambda||\beta||_1 \tag{3.5}$$

In this expression, Y is the vector of observed values, X is the feature matrix, β is the regression coefficient vector, and λ is the hyperparameter that controls the strength of regularization. The introduction of the L1 regularization term forces the coefficients of some features to shrink to zero, thereby achieving feature selection.

When performing Lasso regression, it is necessary to choose an appropriate regularization parameter λ. This can be done using cross-validation, for example, by using LassoCV or GridSearchCV for hyperparameter tuning. During the feature selection process, as λ increases, the coefficients of less important features gradually become zero, while the coefficients of important features remain non-zero. By observing the changes in coefficients for different α (equivalent to λ), it is possible to determine which features are important.

The advantage of Lasso regression in feature selection lies in its interpretability. Since it can set the coefficients of irrelevant or redundant features to zero, the resulting model is more compact. Additionally, Lasso regression is well-suited for high-dimensional datasets, as it can alleviate the curse of dimensionality through feature selection.

4 Data Preprocessing and Visualization

4.1. Data Preprocessing

Table 1 shows a portion of the data used in this study, which records the daily reported new and cumulative confirmed COVID-19 cases, as well as new deaths in China (country code: CN) within the World Health Organization's Western Pacific Region (WPR) from April 13, 2020, to April 21, 2020. Each record includes the report date (**Date reported**), country code, country name, WHO region, the number of new confirmed cases on that day (**New cases**), the total number of cumulative confirmed cases (**Cumulative cases**), and the number of new deaths on that day (**New deaths**), reflecting the development trend and statistical changes during this period.

Table 1. A Portion of the Raw Data

Date reported	Country code	Country	WHO region	New cases	Cumula-tive cases	New deaths	Cumula-tive deaths
2020/4/13	CN	China	WPR	115	83597	2	3351
2020/4/14	CN	China	WPR	99	83696	0	3351
2020/4/15	CN	China	WPR	49	83745	1	3352
2020/4/16	CN	China	WPR	52	83797	0	3352
2020/4/17	CN	China	WPR	27	83824	0	3352
2020/4/18	CN	China	WPR	356	84180	1290	4642
2020/4/19	CN	China	WPR	21	84201	0	4642
2020/4/20	CN	China	WPR	36	84237	0	4642
2020/4/21	CN	China	WPR	13	84250	0	4642
2020/4/22	CN	China	WPR	37	84287	0	4642
2020/4/23	CN	China	WPR	15	84302	0	4642

The dataset preprocessing mainly consists of the following steps:

Data Cleaning: Handling missing values is a crucial task during the data preprocessing phase, as it directly affects the accuracy of subsequent data analysis and modeling. Therefore, this study adopts interpolation methods, which are more suitable for estimating missing values. Linear interpolation is used to fill in the missing data, which is particularly effective for time-series data or data with obvious trends.

Outlier Detection: To ensure the quality of the data, outlier detection is performed to identify and handle potential extreme values. A common method used is the standard deviation-based detection approach. This method assumes that the data follows a normal distribution, and any data point that is more than a certain number of standard deviations away from the mean is considered an outlier. After the initial processing of the dataset, the mean and standard deviation of each column can be calculated to identify all data points that deviate more than a specified number of standard deviations from the mean. Once outliers are identified, they can be either deleted or replaced with other values (such as the median, mean, or boundary values) depending on the specific situation.

4.2. Data Presentation

Subsequently, this study performs descriptive statistics on the collected data, as shown in **Table 2**. The data covers 1,724 days of epidemic statistics, including daily new confirmed cases, cumulative cases, daily new deaths, and cumulative deaths.

The data shows that, on average, approximately 57,645 new confirmed cases were reported daily, with an average of about 37,676,665 cumulative cases. The average daily new deaths were around 70, while the average cumulative deaths were about 49,697.

In addition, the standard deviations for new cases and cumulative cases were 457,776 and 47,068,774, respectively, indicating significant daily fluctuations in case numbers. Similar trends were observed for new deaths and cumulative deaths, but the magnitude of the fluctuations was smaller.

Table 2. Descriptive Statistics of COVID-19 Data in China

	New cases	Cumulative cases	New deaths	Cumulative deaths
count	1724	1724	1724	1724
mean	57645	37676665	70	49697
std	457776	47068774	1095	53633
min	0	1	0	0
25%	23	102167	0	4848
50%	104	1716445	0	15630
75%	1464	99296718	19	121536
max	6966046	99380363	44047	122358

These statistics suggest that during the early stages of the pandemic, new cases and deaths were relatively low, but as the pandemic progressed, these numbers surged, particularly during the peak of the outbreak.

Moreover, the **maximum number of new cases** reached 6,966,046, and the **maximum cumulative cases** totaled 99,380,363, reflecting extreme circumstances on certain days during the pandemic's peak. The **maximum cumulative deaths** reached 122,358, highlighting the severe impact the pandemic had on global health.

5 Prediction Models and Results Analysis

5.1. Problem Definition

Time series forecasting refers to the prediction of the unknown system states over time. This paper primarily focuses on predicting the cumulative COVID-19 infection cases, which essentially involves building an appropriate model based on collected daily case data to predict the number of new cases in the future. Below is the definition of the cumulative infection case prediction problem.

Forecasting Problem Definition: For the multi-step prediction problem of COVID-19 cumulative cases, assume that we have historical time steps of data observations for the new COVID-19 cases $x = (X_1, X_2, \dots, X_p) \in R^{Q \times N}$.Where p represents the dimension of the features. Our goal is to predict the target values for the next Q time steps $y = (Y_{p+1}, Y_{p+2}, \dots, Y_{p+Q}) \in R^{Q \times N}$. The process of predicting the cumulative number of infections can be represented as equation (5.1)

$$Y_{p+1}, Y_{p+2}, \dots, Y_{p+Q} = f(G;\ X_1, X_2, \dots, X_p) \tag{5.1}$$

5.2. Lasso Regression-based Cumulative Infection Prediction

From Table 3, we can summarize the Mean Absolute Error (MAE) and Relative Error (RE) of the Lasso model for different prediction horizons in 2020. The data indicates that as the prediction horizon increases, both MAE and RE also increase, suggesting that the prediction error of the Lasso model grows as the forecast period extends. Specifically, the 3-day forecast has the smallest MAE and RE, while the 30-day forecast has the largest MAE and RE.

Table 3. Summary of Lasso From 2020

Horizon	Mean Absolute Error (MAE)	Relative Error (RE)
3 days	8575221.91	8.63%

5 days	8613817.86	8.67%
7 days	8651950.19	8.76%
10 days	8708268.21	8.84%
14 days	8781684.2	8.96%
21 days	8905424.0	8.96%
30 days	9055299.820	9.11%

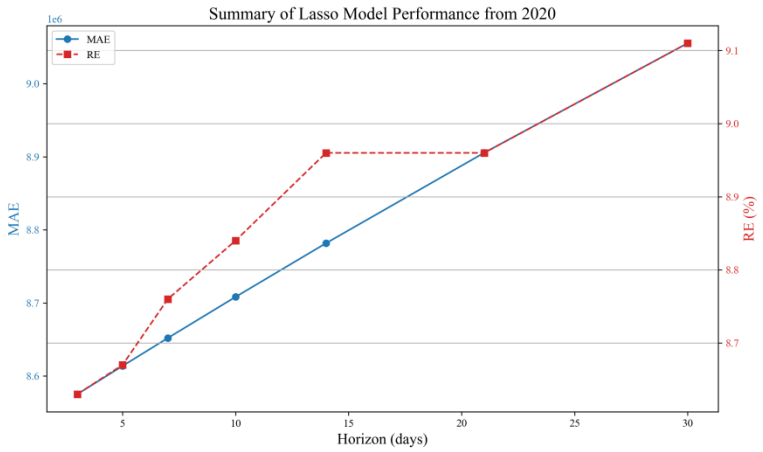

Figure 3. MAE and RE as a Function of Prediction Horizon.

This trend is expected because longer-term forecasts typically involve more uncertainty and variability, making accurate predictions more difficult. The increase in MAE indicates that the average absolute difference between the predicted and actual values is growing, while the rise in RE shows that the error percentage relative to the actual values is also increasing.

5.3. Prediction of Cumulative Infections Based on LSTM

Although traditional Lasso regression can also predict the future cumulative number of infections, the error gradually increases as the prediction horizon expands, leading to a decrease in performance. Meanwhile, both the MAE and RE of the Lasso model are relatively large, making it unsuitable for accurate applications.

Therefore, this study uses the deep learning LSTM model for predicting cumulative infections. Compared to the Lasso model, the LSTM model is capable of extracting trend features from the time series and automatically quantifying these features, which can then be used to predict the number of infections for a given future time horizon.

Table 4. LSTM Network Architecture Diagram

Network Layer Name	Network Layer Shape	Meaning
InputLayer	(None,LookBack,1)	Input as historical LookBack time steps
Dense	(None,LookBack,64)	Non-linear Mapping
LSTM	(None,32)	LSTM Features
Dense	(None,Horizon)	Prediction Layer

In Table 4, the input layer receives historical data, where "None" represents any batch size, LookBack indicates the number of historical time steps considered, and "1" means each time step has

one feature, i.e., the number of infections. Next is a fully connected layer that maps the input data into a 3D tensor with LookBack time steps and 64 neurons. This is followed by an LSTM layer, a long short-term memory network layer specifically designed to capture long-term dependencies in time series data, with an output feature size of 32. Finally, there is a Dense prediction layer that maps the output of the LSTM layer to the predicted results, where Horizon represents the number of future time points the model needs to forecast. The overall design of the model enables it to effectively process and predict time series data.

Table 5. Summary of LSTM From 2020

Horizon	Mean Absolute Error (MAE)	Relative Error (RE)
3 days	3307105.52	1.2%
5 days	3260368.28	3.2%
7 days	3140938.28	3.27%
10 days	2619542.4	2.7%
14 days	984467.71	0.9%
21 days	993624.71	0.11%
30 days	1586796.86	1.6%

In Table 5, the MAE gradually decreases, indicating that the LSTM model achieves higher prediction accuracy in the short term, with the error decreasing over time.

Regarding long-term predictive capabilities, the results for the 21-day and 30-day predictions show relatively low MAE and RE values, which may indicate that the LSTM model has a certain advantage in long-term forecasting, especially in the 30-day forecast where the RE is only 0.11%, demonstrating the model's high accuracy in long-term predictions.

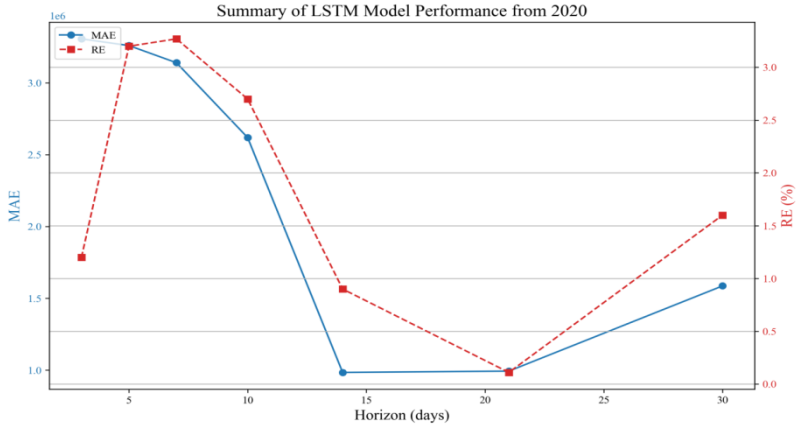

Figure 4. LSTM Model MAE and RE Trend with Varying Prediction Lengths.

Figure 4 shows the trend of MAE and RE for different forecast horizons. It can be observed that the values fluctuate significantly, which may be closely related to seasonality or periodic patterns.

6 Prediction of Cumulative Infections Based on MultiScale-LSTM-Attention

From the previous sections, it can be observed that LSTM performs better than machine learning methods. However, there are still fluctuations in LSTM predictions, which may be due to non-stationary factors at different periods. Existing models often only consider fixed-length historical inputs and extract overall time series trend features, ignoring the impact of different cycles within the data. These cyclical factors are crucial for addressing non-stationarity. Traditional LSTMs only consider the overall

30-day sequence, and some CNN-LSTM models can only consider a fixed number of days, failing to account for changing days.

Therefore, building on the previous work in the sections, this section introduces a multi-scale LSTM model with an attention mechanism to enhance the robustness of the model. Figure 5 shows the specific structure of the Multi-Scale Long Short-Term Memory (LSTM) network. It is designed to capture features at different time scales. The analysis of the figure is as follows:

Input Sequence: The input sequence at the top, denoted as $X_1 - X_t$, is processed by individual LSTM units.

Multi-Scale Processing: The figure shows three LSTM layers at different time scales, labeled as $T = 3$, $T = 5$, $T = 7$. These layers process input sequences with varying lengths of time windows. The $T = 3$ layer processes three consecutive inputs, $T = 5$ processes five consecutive inputs, and $T = 7$ processes seven consecutive inputs.

Feature Extraction: Each LSTM layer at each scale extracts features from its corresponding time window, such as patterns, trends, or other important information in the sequence.

Output: After the lowest layer LSTM, there is an output layer that generates the final prediction or classification result.

In summary, the advantage of the multi-scale LSTM model lies in its ability to simultaneously capture both short-term and long-term temporal dependencies. By combining features from these different scales, the model's performance is improved.

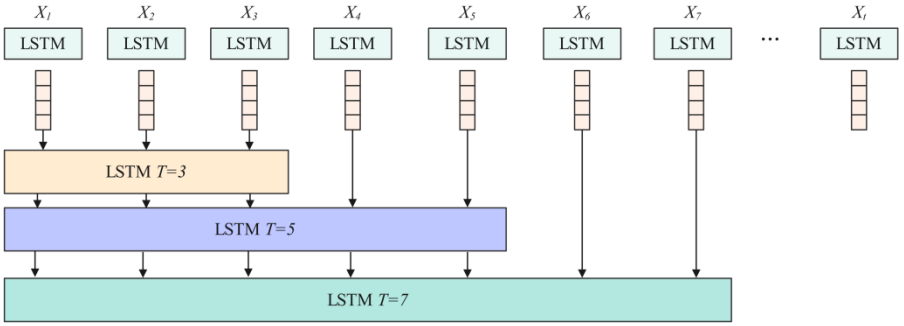

Figure 5. MultiScale-LSTM-Attention Model

7 Comparison of Results between LSTM and MSLA

LSTM demonstrated good performance in previous prediction results, and MLSA further improved upon LSTM by adding multi-scale and attention mechanisms. To validate the effectiveness of MLSA and the added modules, an ablation comparison was conducted.

Table 6. MAE Comparison between LSTM and MLSA

Horizon	LSTM	MLSA
30 days	927.66	786.4

In Table 6, for the 30-day forecast period, the prediction result from LSTM is 927.66, while the prediction result from MLSA is 786.4, indicating that MLSA performs significantly better than LSTM.

Table 7. The results of the ablation experiments for the MLSA model.

Module	MAE
+Attention	803.4
+MultiScale-3	850.32
+MultiScale-5	845.33

+MultiScale-7	820.34
+MultiScale-10	812.56
+MultiScale-14	810.73
+MultiScale-21	800.65

When evaluating the prediction performance of different modules, the Mean Absolute Error (MAE) is used as the metric. MAE represents the average absolute difference between the predicted values and the actual values; the lower the MAE, the higher the prediction accuracy of the model.

Figures 6 and Table 7 show significant differences in performance for each module over a 30-day prediction horizon. First, the "+Attention" module achieved the lowest MAE value of 803.4, indicating that the attention mechanism significantly improves prediction performance. Next, the MultiScale series of modules demonstrated a trend of decreasing MAE as the scale parameter increased. Specifically, the MAE decreased from 850.32 for MultiScale-3 to 800.65 for MultiScale-21, showing a consistent downward trend. This suggests that larger scale parameters help improve the model's prediction accuracy. However, even the optimal MultiScale-21 configuration did not outperform the +Attention and MLSA models in terms of MAE. This result indicates that, although the MultiScale model can be gradually optimized by adjusting the scale parameter, it does not provide higher prediction accuracy than the +Attention/MLSA configuration in the current setup.

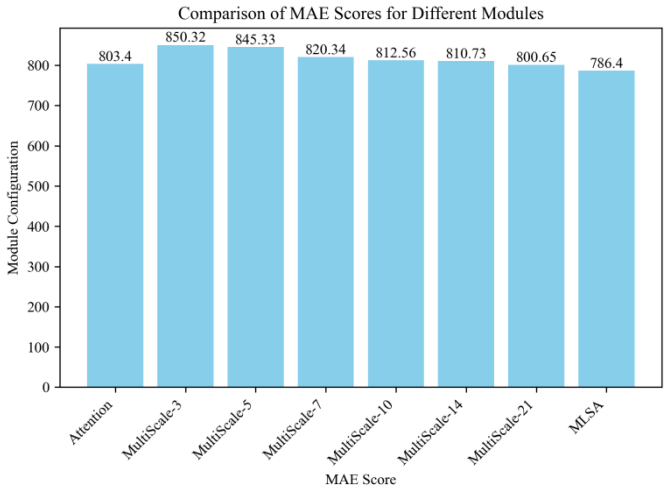

Figure 6. The ablation results of different modules

Based on the above analysis, if the goal is to select the most accurate prediction model, the +Attention or MLSA models should be prioritized based on the MAE metric. Both of these models not only offer the lowest prediction errors, but their performance is also entirely consistent, providing a reliable choice for practical applications. Additionally, incorporating features from different scales significantly improves prediction performance, especially when the scales align with non-stationary cycles. For example, when the non-stationary cycle is 7, using a scale of 7 significantly enhances the prediction performance.

8 Conclusions

This study successfully developed a comprehensive multi-model COVID-19 prediction system by integrating machine learning and deep learning techniques. The system combines Lasso regression, Long Short-Term Memory (LSTM) networks, and a novel Multi-Scale Cumulative Infection Prediction model based on an attention mechanism (MSLA), significantly enhancing prediction accuracy and reliability. Relying on COVID-19 data from China, the study carefully performed data preprocessing and feature engineering to ensure data accuracy and adequate information extraction. The training, evaluation, and visualization of the models were thoroughly documented, providing clear guidance for future model optimization.

The integration of the MSLA model allows this research to simultaneously account for multiple periodic factors, effectively addressing the non-stationary nature of epidemic data. Experimental results reveal that the MSLA model outperforms the traditional LSTM model in 30-day long-term predictions, reducing the Mean Absolute Error (MAE) by 15.22%, offering more accurate scientific support for public health decision-making.

Overall, this study not only provides an innovative and effective tool for COVID-19 prediction but also offers valuable experience and methodology for addressing similar pandemics in the public health field. As the pandemic continues to evolve, the methodology and framework presented in this study will provide strong support for global epidemic monitoring and response, helping to mitigate the impact of the epidemic on social economies and public health systems.

REFERENCES

1. Ciotti, M., Ciccozzi, M., Terrinoni, A., Jiang, W. C., Wang, C. B., & Bernardini, S. (2020). The COVID-19 pandemic. *Critical Reviews in Clinical Laboratory Sciences*, *57*(6), 365–388. https://doi.org/10.1080/10408363.2020.1783198
2. Santosh, K.C. COVID-19 Prediction Models and Unexploited Data. *J Med Syst* 44, 170 (2020). https://doi.org/10.1007/s10916-020-01645-z
3. Singh, R. K., Rani, M., Bhagavathula, A. S., Sah, R., Rodriguez-Morales, A. J., Kalita, H., ... & Kumar, P. (2020). Prediction of the COVID-19 pandemic for the top 15 affected countries: Advanced autoregressive integrated moving average (ARIMA) model. JMIR public health and surveillance, 6(2), e19115. https://doi.org/10.2196/19115
4. Alabdulrazzaq, H., Alenezi, M. N., Rawajfih, Y., Alghannam, B. A., Al-Hassan, A. A., & Al-Anzi, F. S. (2021). On the accuracy of ARIMA based prediction of COVID-19 spread. Results in Physics, 27, 104509. https://doi.org/10.1016/j.rinp.2021.104509
5. Heidari, A., Jafari Navimipour, N., Unal, M. *et al.* Machine learning applications for COVID-19 outbreak management. *Neural Comput & Applic* 34, 15313–15348 (2022). https://doi.org/10.1007/s00521-022-07424-w
6. Alakus, T. B., & Turkoglu, I. (2020). Comparison of deep learning approaches to predict COVID-19 infection. Chaos, Solitons & Fractals, 140, 110120. https://doi.org/10.1016/j.chaos.2020.110120
7. Shahid, F., Zameer, A., & Muneeb, M. (2020). Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM. Chaos, Solitons & Fractals, 140, 110212. https://doi.org/10.1016/j.chaos.2020.110212
8. Islam, M. Z., Islam, M. M., & Asraf, A. (2020). A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images. Informatics in medicine unlocked, 20, 100412. https://doi.org/10.1016/j.imu.2020.100412
9. Shah, P. M., Ullah, F., Shah, D., Gani, A., Maple, C., Wang, Y., & Islam, S. U. (2021). Deep GRU-CNN model for COVID-19 detection from chest X-rays data. Ieee Access, 10, 35094-35105. https://doi.org/10.1109/ACCESS.2021.3077592
10. Yu, Y., Si, X., Hu, C., & Zhang, J. (2019). A review of recurrent neural networks: LSTM cells and network architectures. Neural computation, 31(7), 1235-1270. https://doi.org/10.1162/neco_a_01199
11. Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., & Wang, Y. (2021). Transformer in transformer. Advances in neural information processing systems, 34, 15908-15919.
12. Ranstam, J., & Cook, J. A. (2018). LASSO regression. Journal of British Surgery, 105(10), 1348-1348. https://doi.org/10.1002/bjs.10895
13. Van Tinh, N. (2020). Forecasting of COVID-19 confirmed cases in Vietnam using fuzzy time series model combined with particle swarm optimization. Comput Res Prog Appl Sci Eng, 6(2), 114-120. https://crpase.com/archive/CRPASE-Vol-06-issue-02-20802699.pdf
14. Song, J., Xie, H., Gao, B., Zhong, Y., Gu, C., & Choi, K. S. (2021). Maximum likelihood-based extended Kalman filter for COVID-19 prediction. Chaos, Solitons & Fractals, 146, 110922. https://doi.org/10.1016/j.chaos.2021.110922
15. Chen Guoxin (2024) Prediction of the dynamics covid19 epidemic process of using the Lasso regression model (master diploma) V. N. Karazin Kharkiv National University.

Качанов Станіслав Андрійович
здобувач третього (освітньо-наукового) рівня вищої освіти за спеціальністю 122 Комп'ютерні науки кафедри теоретичної та прикладної інформатики
Харківський національний університет імені В .Н. Каразіна, майдан Свободи, 4, Харків, Україна, 61022
e-mail: staskachanov2000@gmail.com
https://orcid.org/0009-0002-6938-6717

Чень Госінь
здобувач другого (магістерського) рівня вищої освіти за спеціальністю 122 Комп'ютерні науки, освітньо-професійної програми "Інформатика" кафедри теоретичної та прикладної інформатики
Харківський національний університет імені В .Н. Каразіна, майдан Свободи, 4, Харків, Україна, 61022
e-mail: guoxin.chen@student.karazin.ua;
https://orcid.org/0009-0004-0502-3735

Морозова Анастасія Геннадіївна
к.т.н., доцент закладу вищої освіти кафедри теоретичної та прикладної інформатики
Харківський національний університет імені В .Н. Каразіна, майдан Свободи, 4, Харків, Україна, 61022
e-mail: a.morozova@karazin.ua
https://orcid.org/0000-0003-2143-7992

Руккас Кирило Маркович
д.т.н, доцент, професор закладу вищої освіти кафедри теоретичної та прикладної інформатики
Харківський національний університет імені В .Н. Каразіна, майдан Свободи, 4, Харків, Україна, 61022
e-mail: rukkas@karazin.ua
https://orcid.org/0000-0002-7614-0793

Прогнозування динаміки епідемічного процесу COVID19 з використання моделі Ласо регресії

У роботі запропоновано об’єднати методи машинного та глибокого навчання для прогнозування пандемії COVID-19.
Актуальність. Глобальний спалах пандемії COVID-19 мав великий вплив на системи здоров’я та соціально-економічні структури в усьому світі, підкреслюючи нагальну потребу в ефективних інструментах прогнозування для допомоги в прийнятті рішень. Робота присвячена побудові мультимодельної структури для прогнозування епідемії шляхом інтеграції передових методів математичного моделювання та теорії прогнозування.
Мета. Метою роботи було проаналізувати методи та алгоритми для кумулятивного прогнозування випадків COVID-19з метою надання наукової підтримки для прийняття рішень у сфері охорони здоров’я.
Методи дослідження. Методи дослідження базуються на сучасних теоріях математичного моделювання, штучного інтелекту, епідеміологічної діагностики, теорії прогнозування, а саме: регресія Лассо, мережі довгострокової короткочасної пам’яті (LSTM) і моделі LSTM-Attention.
Результати. Було досліджено застосування моделі Лассо регресії та мережі довгострокової короткочасної пам’яті (LSTM) для кумулятивного прогнозування випадків COVID-19з метою надання наукової підтримки для прийняття рішень. В роботі детально описано процеси попередньої обробки даних, навчання моделі, оцінювання та візуалізації для підтримки узагальнення та адаптивності в динамічному сценарії пандемії. Крім того, була запропонована модель Multi-Scale LSTM-Attention (MSLA) для вилучення багатоперіодних ознак із вхідних послідовностей. Ці функції є критично важливими для вирішення проблеми нестаціонарності даних.
Висновки. Вирішено задачу розробки системи прогнозування COVID-19 шляхом інтеграції методів машинного та глибокого навчання. Розроблена система поєднує в собі регресію Лассо, мережі довгострокової короткочасної пам’яті (LSTM) і нову багатомасштабну модель кумулятивного прогнозування зараження на основі механізму уваги (MSLA), що значно підвищує точність і надійність прогнозування.

Ключові слова: *прогнозування, нейронні мережі, Lasso Regression Model, LSTM, COVID-19, Машинне навчання, Глибоке навчання, Multi-scale model, бази даних.*

Надійшла у першій редакції 12.08.2024, в останній - 29.09.2024.
The first version has been received on 12.08.2024, the final version - on 29.09.2024.