On the Evolution of COVID-19 Virus Based on the Prediction Model of Deep Learning and Emotion Analysis

Yucheng Fan, Dong Qiu

***Abstract*—In January 2020, COVID-19 broke out in Wuhan, China. In just a few months, the virus spread around the world. In the past three years, the virus has undergone continuous mutations, with the latest variant being the Omicron variant. Whenever a new variant emerges, there are significant changes in the transmission rate, mortality rate, and other essential disease characteristics. These characteristics of the virus have posed considerable challenges to countries and health departments. Time series models used in this study incorporate text sentiment. To achieve this, we utilized snscrape to retrieve tweets and applied different keywords to filter the tweets. Subsequently, we used clean text as input for a pre-trained model to conduct the sentiment analysis. Finally, the sentiment analysis results and other epidemic features were combined as inputs for time series models to generate predictions. We can observe changes in the virulence of different variants through the models' outcomes. Research has shown that incorporating sentiment analysis results can effectively improve the model's predictive performance. When a model is trained using historical data, it cannot accurately predict viruses that will emerge in the future. The same holds in reverse. The result indicates that the nature of the virus has changed during different stages, suggesting the emergence of new variants. The study will help local health departments improve control measures, enabling adjustments to be made specifically for different variants.**

***Index Terms*—COVID-19, mutant strains, deep learning, sentiment analysis, time series prediction.**

I. Introduction

THE COVID-19 pandemic, also known as the coronavirus pandemic, is a new respiratory disease [36]. In December 2019, an ophthalmologist from Wuhan discovered the first case of COVID-19 in China. On 30 January 2020 following the recommendations of the Emergency Committee, the WHO Director General declared that the outbreak constitutes a Public Health Emergency of International Concern. In March 2020, experts classified it as a pandemic and determined that the virus could spread among humans through respiratory droplets. COVID-19 has significantly impacted our lives [51], [52], [53], [54]. We need to formulate corresponding policies according to the changes in the characteristics of the virus so that our lives can return to normal as soon as possible.

Manuscript received April 18, 2023; revised December 19, 2023.

This work was supported by The National Natural Science Foundations of China (Grant no.12171065 and 11671001).

Y. Fan is an undergraduate student of School of Computer Science and Technology, Chongqing University of Posts and Telecommunications Nanan, Chongqing, 400065, P.R. China. (e-mail: fyc5860@gmail.com).

D. Qiu is a professor of the College of Science, Chongqing University of Posts and Telecommunications Nanan, Chongqing, 400065, P.R. China (corresponding author to provide phone: 15123126186; fax: 023-62641001; e-mail: dongqiumath@163.com).

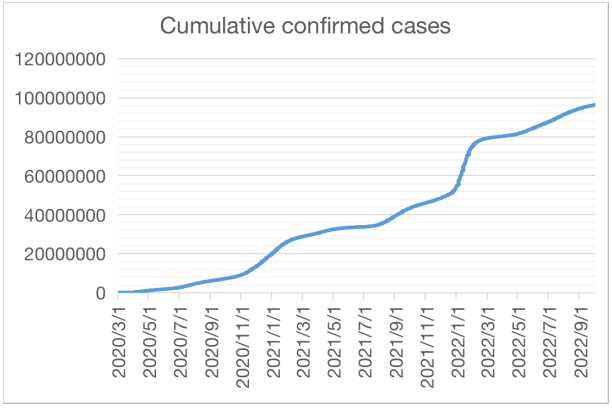


Fig. 1. The cumulative number of confirmed cases in America

Taking the United States as an example, by the end of 2022, it had experienced about five waves of significant outbreaks. Due to limited interventions by the US government and local health departments, the population's daily lives were not significantly affected. That is also why the pandemic has spread faster now in the United States. During the early to mid-2020 period, the original variant was prevalent in society. The original variant has a high virulence and mortality rate, marking the first wave of a significant outbreak in the United States. Subsequently, local health departments began to require people to stay at home and wear masks when going out. From July to October 2020, the premature reopening of certain states without strictly adhering to new preventive measures led to the second wave of the pandemic.

As some individuals had already developed antibodies in mid-2020, both infection and mortality rates decreased. In October 2020, the Alpha variant was detected for the first time among confirmed cases [55]. The emergence of the Alpha variant led to an increase in both diagnosis and mortality rates, marking the beginning of the third wave of the pandemic. Towards the end of 2020, the US government began to implement immunization programs for the population. In 2021, the Delta and Omicron variants emerged [56]. Each occurrence of a new variant typically will lead to a recent major outbreak. Fig. 1 shows the cumulative number of confirmed COVID-19 cases in the United States.

Due to the constantly evolving nature of RNA viruses, the prevention and control of the epidemic have been significantly affected. It requires us to promptly identify changes in viral characteristics and design more effective diagnostic tests. At the same time, it can also help local health departments adjust their preventive measures. Biolo-

gists have analyzed viral mutation and evolutionary trends from the perspectives of genes and proteins. Scientists have proposed different detection methods to track the mutations more effectively. The most effective and widely used way is RT-PCR testing [57]. This technique combines reverse transcription of RNA and polymerase chain reaction (PCR) amplification of cDNA to detect gene expression levels in cellular tissues. Since different variants may be circulating in other regions, some countries conduct wastewater testing on incoming flights to prevent imported cases. However, this method has certain limitations as most passengers do not use the restroom on the plane. In hospitals, biosensors and CT scans of the lungs are also standard diagnostic methods [50], [58]. Akhtar et al. [49] employed machine learning algorithms using full blood count to predict the prevalence of COVID-19 in the future. Fig. 2 shows these diagnostic methods.

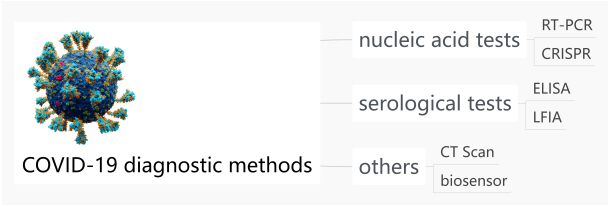


Fig. 2. Common testing methods for COVID-19

In addition to measures such as detecting incoming cases and conducting relevant biological analyses of strains, we can also use deep learning models and mathematical models [7], [8], [25], [27], [28], [34], [42], [43] to identify changes in the various characteristics of the virus. It can provide local health departments with a rough assessment of mutations, allowing for the dynamic adjustment of defense measures [39], [41]. It also provides the health authorities with a guide for handling mutations [29], [30], [31], [32], [33].

II. Related research works

To detect the rapid mutations of the viruses, Wang et al. [6] investigated the characteristics of the virus through various channels, including its structure, epidemiology, and different therapeutic approaches. They determined the sequence variations of nucleotides and amino acids by investigating the virus's genomic sequences and protein structures. After evaluating the complete genome sequences of 11 variants, they inferred that COVID-19 might produce new mutations, specifically in glycoproteins. They calculated the mutation rates to validate the effectiveness of the currently used diagnostic reagents. Their study contributed to the development of effective diagnostic methods and therapeutic drugs.

In addition to conducting biological analyses of the virus, we can indirectly reflect the mutations of the virus through machine learning. Before this study, some individuals also devoted themselves to researching fields related to the pandemic. These studies [9] - [24] analyzed various datasets, such as Weibo, Twitter, and other international platforms. Furthermore, there are many types of research in this field with different focuses, including research on public opinions regarding mask-wearing and analyzing public attitudes towards different vaccine brands.

Nanning et al. proposed a hybrid AI model for predicting the COVID-19 epidemiological trends in China [1]. The model combined natural language processing (NLP) with LSTM networks. The dataset used in their study includes text data from reports and news related to the pandemic. The text data serves as input to the NLP model, which then outputs the text features. Traditional methods of forecasting usually assume that the infectivity of patients remains the same. However, in reality, the infectivity of confirmed cases varies in the early and middle stages. The ISI model in the article aims to address this issue. However, the data calculated by the ISI model does not consider the impact of implemented measures. Therefore, the research ultimately combines the NLP model with LSTM to help correct the results obtained from the ISI model. Sumit et al. proposed an improved model based on Prophet for predicting cumulative confirmed cases in India [2]. Their study used natural language processing libraries (TextBlob, VADER, Stanza) for sentiment analysis. The results indicate that the improved model outperforms the commonly used ARIMA in long-term forecasting capability, per previous studies. In India's second COVID-19 wave, Delta presented enhanced infectivity levels compared with others. Additionally, they provided a comparative comparison of natural language processing libraries for text analysis and visualized text data using word clouds.

In the waves of COVID-19, national governments and local health departments implemented various measures. For example, implementing social distancing policies and closing public venues are common restrictions. Vaccination is also a crucial step that plays an important role in reducing the harm caused by the virus. Sattar [3] analyzed the perception of the U.S. population towards vaccines produced by different companies. The study utilized a tweets dataset. High-frequency words in the text included "vaccine willingness" and "vaccine side effects". Although most tweets exhibited neutral sentiments, the proportion of tweets expressing negative sentiment did not exceed that of tweets expressing positive sentiment. Therefore, the majority of people are not opposed to vaccine administration. The model results indicated that after July 2021, the percentage of the population who received at least one dose of the vaccine would reach 62.44% of the U.S. population. As the vaccination coverage continued to rise, most people showed a greater willingness to reopen. High-frequency words in tweets during this period were mainly about "gatherings", "reopening", and "travel", with roughly equal proportions of tweets expressing positive and negative sentiment. Before implementing the vaccination program, high-frequency words in the text were primarily about "thrombosis" and "fever". The public's attitude towards vaccines evolved from anxiety to acceptance. Their study would assist local health departments and policymakers in understanding the psychological state of people and determining what services and resources should be provided to the public during such a period.

LSTM (Long Short-Term Memory) is a type of artificial neural network [1], [5], [61]. Unlike feedforward neural networks, LSTM has feedback connections. It can process not only individual data points but also data sequences. LSTM has been the most cited neural network in the mid-20th century. In this study, we also use LSTM as a representative deep learning model. Compared to the recurrent neural

network (RNN), LSTM can better capture the COVID-19 epidemiological trends. LSTM has several benefits that make it a powerful tool for modeling sequential data.

In our research, the first step is to collect relevant tweets. After preprocessing the text, it will be used as input to a pre-trained model. The output provides the sentiment analysis results. Finally, the sentiment analysis results and the virus's characteristic data will be combined as inputs to the model to predict the COVID-19 epidemiological trends.

III. Research methods

A. Dataset

In this study, Twitter will be used as a social platform to obtain public attitudes towards COVID-19 and new variants. The snscrape library is a text-scraping tool that helps us retrieve tweets [35]. To collect tweets from the United States, the tweets will be filtered based on the geographical coordinates. In order to ensure the relevance of the collected tweets to the epidemic, the tweets will be filtered again using several keywords related to the epidemic, such as "COVID-19", "coronavirus", "Omicron", and so on. After that, the timeline is divided into multiple stages to compare the characteristics of different variants. During different periods, different keywords related to the variants will be used for further filtering during the tweet retrieval process.

Once we have the text dataset, the time series models require relevant statistical data, such as the infection rate, mortality rate, control measures, and so on. We can preliminarily identify specific variants' transmission patterns and characteristics with these statistical data. For example, as the vaccination rate rises, there is going to be a declining trend in the mortality rate among the population. After discovering new variants, the infection rate initially rises and then declines, indicating the formation of herd immunity. Additionally, as the number of input variables in the model increases, its performance improves compared to the models training with a single variable, resulting in higher accuracy. Table I shows the characteristics of the virus.

TABLE I
Primary Input Variables of the Model

Id	Feature	Description
a.	Cumulative confirmed cases	The number of cumulative confirmed cases
b.	New confirmed cases	The number of new confirmed cases on that day
c.	Cumulative deaths	The number of cumulative deaths
d.	New deaths	The number of new deaths on that day
e.	ICU patients	The number of patients in ICU
f.	Hospitalized cases	The number of hospitalized cases
g.	Vaccination	The number of individuals who have received at least one dose and those who are fully vaccinated
h.	Not fully vaccinated	The number of people who have received only one dose of the vaccine
i.	Fully vaccinated	The number of people who have been fully vaccinated with two or three doses
j.	New vaccinations	The number of newly vaccinated individuals with at least one dose
k.	Booster dose	The number of people who received an additional dose
l.	Blocking measures	Quantified restriction level: the larger the number, the more restrictive measures

B. Data Cleaning

First of all, the text data must be preprocessed to optimize the model's performance. This step could avoid noise affecting the experimental results. The NLTK (Natural Language Toolkit) is used to clean the noise in the text. Also, it can be used for tasks such as stemming or lemmatization [1]. Since the collected text is in English, it is necessary to tokenize all the words and convert them to lowercase. Unlike Chinese, English sentences have words separated by spaces, making tokenization relatively straightforward. Next, most tweets contain web links, which could not be more helpful for sentiment analysis. Therefore, they need to be removed from the sentences. Then, the stop words in the text should also be removed. After completing the above steps, we perform stemming and lemmatization on the remaining vocabulary. In text processing, the original form of words can be obtained through stemming and lemmatization. However, stemming is more aggressive and may produce incorrect results when searching for word stems [59]. For example, "leaves" could be incorrectly called "leav". Lemmatization, however, is more conservative and typically handles only words that can be correctly transformed. This study uses the WordNet provided by NLTK. WordNet aims to preserve the original meaning of the words instead of overly simplifying them. Fig. 3 illustrates this process. Finally, the clean words must be put into sentences since only sentences can serve as input for the pre-trained model.

Fig. 3. Process of text preprocessing

IV. Models

A. Pre-Training Model

RoBERTa is a pre-trained model optimized based on the BERT (Bidirectional Encoder Representations from Trans-

formers) model, with some hyperparameter modifications [60]. RoBERTa is a transformers model trained on a large corpus of English data in a self-supervised fashion. This means it was trained on the raw texts only, with no humans labeling them in any way (which is why it can use lots of publicly available data) with an automatic process to generate inputs and labels from those texts. In this research, the pre-trained model categorizes the sentiment of sentences into three classes: negative, neutral, and positive. This process involves scoring each word in the sentence individually and deriving the final result. It also highlights the importance of text preprocessing, as we must ensure that each word in the sentences carries meaningful information for the model's output to be accurate.

B. Time Series Forecasting

1) Time series concepts: Time series data, also called time-stamped data, is a sequence of data points indexed in time order. Time series forecasting is a regression prediction method and falls under quantitative forecasting. Its fundamental principle is twofold: firstly, it recognizes the continuity of the development of phenomena and utilizes past time series data for statistical analysis to infer the development trends of the phenomena [62]. Secondly, it takes into account the randomness caused by incidental factors.

Time series prediction is an essential field in prediction research. Davidescu et al. [26] used ARIMA to predict social unemployment and studied uncertainty based on sector graphs. Gunay et al. [37] combined the epidemic forecast with the economic situation and predicted the COVID-19 epidemiological trends in China in the second quarter of 2020.

2) Deep learning models: Generally, time series forecasting involves two types of approaches. The first type is traditional forecasting methods, including ARIMA [15], [43], [65], mean regression, exponential smoothing, and so on [38], [63]. The second type is machine learning models [44], [45], [46], [47], [48]. In the field of time series forecasting, we can further classify machine learning models into two categories: tree-based models and neural network models. Tree-based models are widely used in machine learning, and they have many variants, including decision trees, random forests, XGBoost, and more [2], [12], [16], [33]. Tree-based models are easier to interpret than deep learning models. They are less prone to overfitting during training and require fewer computational resources. Deep learning has gained popularity recently as traditional methods struggle to capture the mixed information of long-term and short-term sequences. Deep learning models can capture non-linear interdependencies in the data, compensating for the limitations of traditional machine learning methods. It has led to the gradual superiority of deep learning models over conventional methods [40]. This research will focus on four deep learning models showing results in time series forecasting, including LSTM, GRU, CNN, and BiLSTM [45].

C. Deep Learning Models

1) LSTM: An ordinary LSTM unit comprises a cell, an input gate, an output gate, and a forget gate. The cell has the ability to retain values across arbitrary time intervals, while the three gates govern the inflow and outflow of information within the cell. Forget gates play a crucial role in determining which information from the previous state to discard; they assign a value between 0 and 1, where 1 signifies information retention and 0 implies discarding. Similarly, input gates decide which new information to incorporate into the current state using a comparable mechanism as forget gates. Output gates, considering both the prior and present states, dictate the information to be output by assigning values from 0 to 1. This selective outputting of pertinent information empowers the LSTM network to preserve valuable, long-term dependencies for making predictions across current and future time-steps. Fig. 4 shows the basic structure of the LSTM model.

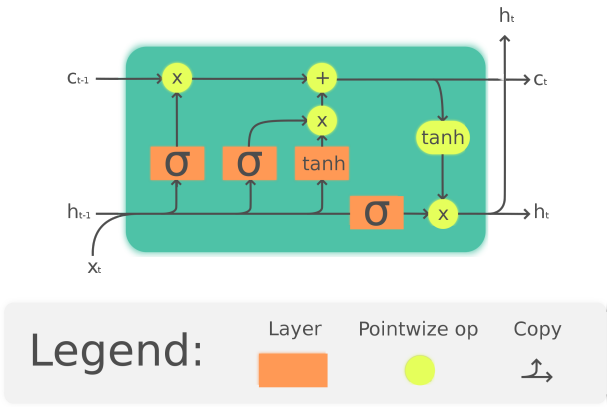


Fig. 4. The architecture of LSTM

$$I_t = \sigma(X_t W_{xi} + H_{t-1} W_{hi} + b_i) \tag{1}$$

$$F_t = \sigma(X_t W_{xf} + H_{t-1} W_{hf} + b_f) \tag{2}$$

$$O_t = \sigma(X_t W_{xo} + H_{t-1} W_{ho} + b_o) \tag{3}$$

$$\widetilde{C}_t = tanh(X_t W_{xc} + H_{t-1} W_{hc} + b_c) \tag{4}$$

$$C_t = F_t \odot C_{t-1} + I_t \odot \widetilde{C}_t \tag{5}$$

$$H_t = O_t \odot tanh(C_t) \tag{6}$$

2) BiLSTM: Bidirectional LSTMs (Long Short-Term Memory) represent a type of recurrent neural network (RNN) architecture that evaluates input data in both forward and backward directions. Unlike conventional LSTMs, which exclusively process information from past to future for predictions, bidirectional LSTMs consider both past and future contexts. Comprising two LSTM layers—one processing input in the forward direction and the other in the backward direction—this design allows simultaneous access to information from preceding and succeeding time steps. Consequently, bidirectional LSTMs prove highly advantageous for tasks demanding a thorough comprehension of input

sequences, including applications like sentiment analysis, machine translation, and named entity recognition in natural language processing. Fig. 5 shows the basic structure of the BiLSTM model.

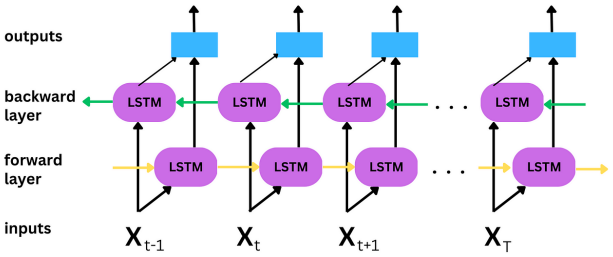


Fig. 5. The architecture of BiLSTM

3) CNN: A convolutional neural network operates as a feed-forward neural network, featuring artificial neurons capable of responding to adjacent units within a specific coverage area. Renowned for its exceptional performance in extensive image processing, the convolutional neural network includes one or more convolutional layers, culminating in a fully connected layer akin to a conventional neural network. These layers, along with associated weights and pooling layers, enable the network to effectively leverage the two-dimensional structure of input data. In comparison to alternative deep learning architectures, convolutional neural networks consistently yield superior results in tasks such as image and speech recognition. Notably, max pooling, a specific pooling operation, extracts the maximum element from the region covered by the filter in the feature map, resulting in an output feature map that encapsulates the most prominent features from the previous layer. Fig. 6 shows the pooling process of the CNN model.

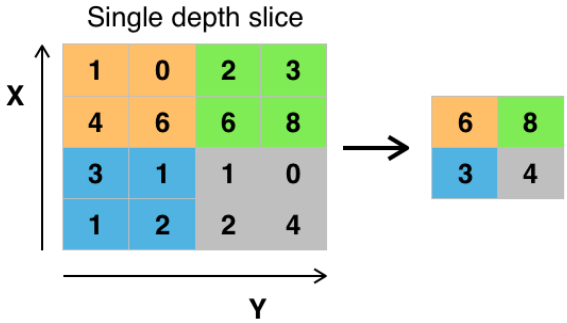


Fig. 6. Polling layer

4) GRU: To some extent, GRU is also an optimization of the structural complexity of LSTM. LSTM can solve the gradient disappearance and gradient explosion problems caused by long-term dependence of recurrent neural networks. However, LSTM has three different gates and many parameters, making it difficult to train. GRU only contains two gating structures, and when all hyperparameters are tuned, the performance of the two is equivalent, but the GRU structure is simpler, has fewer training samples, and is easy to implement. Fig. 7 shows the basic structure of the GRU model.

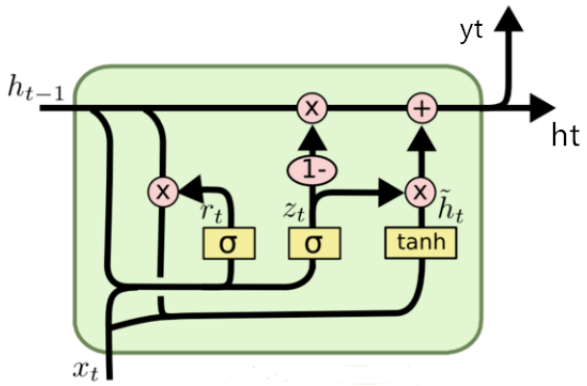


Fig. 7. The architecture of GRU

$$Z_t = \sigma(W_z[h_{t-1}, x_t]) \tag{7}$$

$$\widetilde{h} = tanh(W[r_t \odot h_{t-1}, x_t]) \tag{8}$$

$$R_t = \sigma(W_r[h_{t-1}, x_t]) \tag{9}$$

$$H_t = O_t \odot tanh(C_t) \tag{10}$$

CNNs typically excel when dealing with data featuring spatial relationships. While traditionally configured for two-dimensional input like a matrix, they can be adapted to a one-dimensional format, enabling them to capture the internal structure of sequences. This versatility allows CNNs to extend their effectiveness to various data types with spatial interdependencies. For instance, the ordered relationship between words in textual documents or the sequential nature of time steps in time series data.

In contrast, RNNs, initially designed for sequence prediction tasks, historically posed training challenges. The Long Short-Term Memory (LSTM) network has proven remarkably successful in addressing these issues, making it a widely applied solution across diverse applications. Notably, RNNs and LSTMs demonstrate significant efficacy when handling word and paragraph sequences, commonly associated with natural language processing.

V. Experiments

A. Sentiment Classification

Tweets are collected and filtered using epidemic-related keywords. As a result, the number of tweets expressing positive sentiment is less than that expressing negative sentiment. About half of all texts express neutral sentiment, while one-third express negative sentiment. It shows that people in the United States have a negative attitude towards the epidemic. Fig. 8 shows the average proportions of emotions in these tweets.

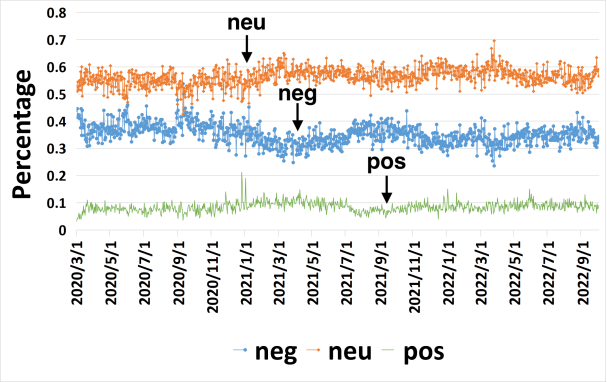


Fig. 8. Sentiment distribution of tweets (United States)

B. Long-term Forecasts and Short-term Forecasts

1) Epidemic transmission timeline: In March 2020, the epidemic broke out in the United States and spread rapidly. The state governments began to ban the export of medical supplies to maintain the medical system and declared a state of emergency. In early April, the state governments issued an immigration ban. After that, the epidemic in the United States ushered in the first peak. The first wave of the epidemic caused by the original virus has a high mortality rate. In the middle of 2020, most states across the country dropped COVID-related restrictions as cases dropped and gradually reopened. The premature opening up accelerated the arrival of the second wave of the epidemic. At the end of 2020, the United States found the Alpha variant in confirmed cases. According to the research, patients cough more frequently after being infected with the Alpha variant, which led to a higher infection rate of the Alpha variant compared with the original virus. At the same time, the state governments announced a vaccination program. At the beginning of 2021, the U.S. government issued several policies to slow the spread of the virus. In March 2021, the Delta variant appeared, and its transmission rate was about twice that of the original virus. However, as the population's vaccination rate gradually rose, the number of coronavirus infected people declined. At the end of 2021, the United States discovered the Omicron variant. Because of the reduction of its toxicity and the promotion of the booster dose, the mortality rate is lower than that of the Delta variant. Fig. 9 shows the epidemic timeline.

2) Predictions for different variants: For the same type of variant, they have the same characteristics. When the model is trained using the time series data of the current phase, it will better fit the recent variant growth rate. The study divides the timeline into three periods. The dominant variant in the first wave of the epidemic was the original virus. This phase started in March 2020 and continued until the end of 2020. The dominant variant in the second wave of the epidemic was the Delta virus. The Omicron variant caused the third wave of the epidemic. This phase started at the end of 2021 and continues to date. The models are trained using three different types of data. The experiment in this study predicts the short-term and long-term trend prediction of the COVID-19 pandemic. The experiment sets the window size to 7 data points. To make the experiment results more interpretable, the experiment scales the error results of the models to the range of [0, 1].

TABLE II
COMPARISON OF ERROR RESULTS FOR SHORT-TERM AND LONG-TERM FORECASTS

	Long-term prediction of original virus			
	LSTM	CNN	BiLSTM	GRU
mse	0.424	0.065	0.04	0.18
mae	0.484	0.227	0.185	0.388
mape	0.192	0.099	0.081	0.167
	Short-term prediction of original virus			
	LSTM	CNN	BiLSTM	GRU
mse	1.142	0.089	0.073	0.598
mae	1.055	0.289	0.267	0.769
mape	0.391	0.107	0.1	0.286

For the original virus, the model performs better in long-term trend prediction. In long-term forecasting, the BiLSTM model performs the best with the MSE result of 0.04. In short-term forecasting, the BiLSTM model also performs the best with the MSE result of 0.073. The error data of these models are shown in Table II, and the prediction results are displayed in Fig. 10 and Fig. 11.

TABLE III
COMPARISON OF ERROR RESULTS FOR SHORT-TERM AND LONG-TERM FORECASTS

	Short-term prediction of the Delta variant			
	LSTM	CNN	BiLSTM	GRU
mse	0.201	0.065	0.181	0.24
mae	0.447	0.252	0.423	0.49
mape	0.229	0.129	0.217	0.251
	Long-term prediction of the Delta variant			
	LSTM	CNN	BiLSTM	GRU
mse	0.095	0.029	0.073	0.164
mae	0.233	0.162	0.249	0.403
mape	0.125	0.091	0.14	0.224

For the Delta variant, the performance of the models on long-term prediction is better than that on short-term prediction. In long-term forecasting, the CNN model performs the best with the MSE result of 0.029. In short-term forecasting, the CNN model also performs the best with the MSE result of 0.065. The error data of these models are shown in Table III, and the prediction results are displayed in Fig. 12 and Fig. 13.

TABLE IV
COMPARISON OF ERROR RESULTS FOR SHORT-TERM AND LONG-TERM FORECASTS

	Short-term prediction of the Omicron variant			
	LSTM	CNN	BiLSTM	GRU
mse	0.042	0.02	0.092	0.044
mae	0.204	0.141	0.303	0.211
mape	0.165	0.114	0.245	0.17
	Long-term prediction of the Omicron variant			
	LSTM	CNN	BiLSTM	GRU
mse	0.012	0.157	0.024	0.024
mae	0.084	0.375	0.128	0.128
mape	0.069	0.313	0.106	0.106

For the Omicron variant, the performance of the models on long-term prediction is better than that on short-term prediction. In short-term forecasting, the CNN model performs the best with the MSE result of 0.02. In long-term forecasting, the LSTM model performs the best with the MSE result of 0.012. The error data of these models are shown in Table IV, and the prediction results are displayed in Fig. 14 and Fig. 15.

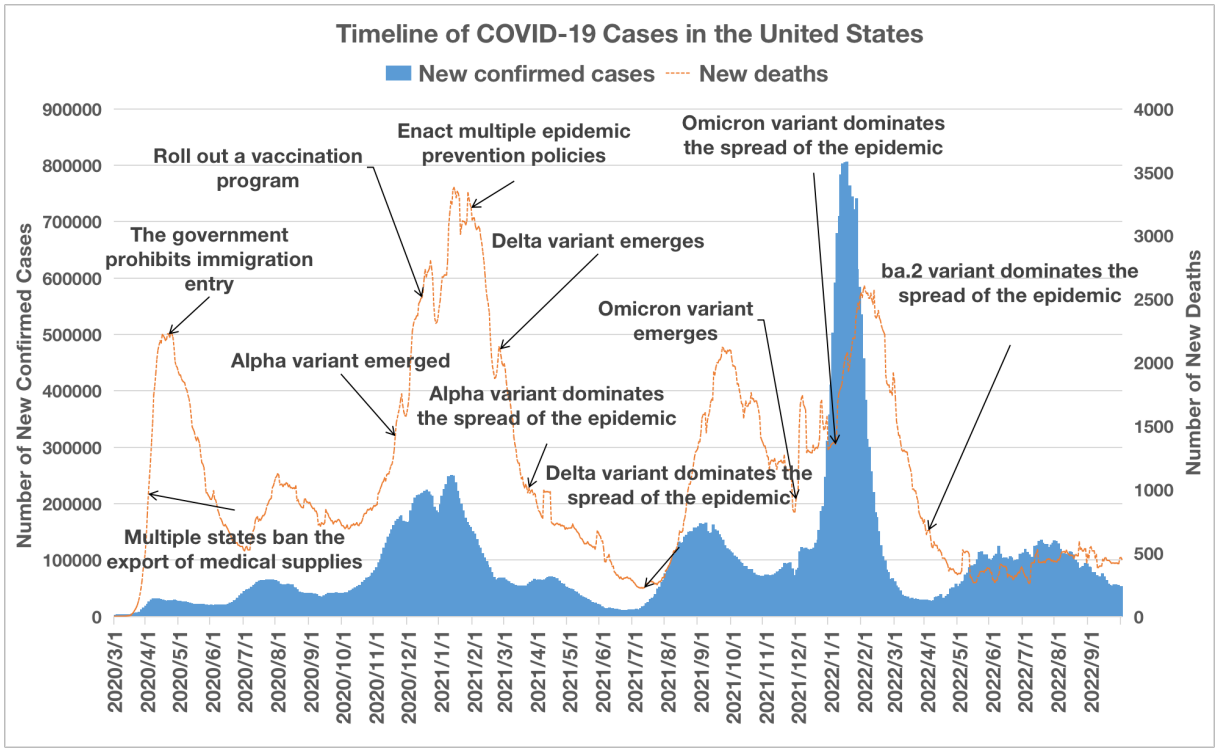


Fig. 9. Timeline of COVID-19 in the United States

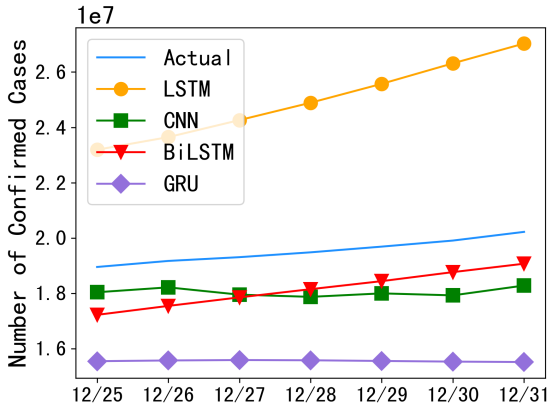


Fig. 10. Short-term prediction of the original virus

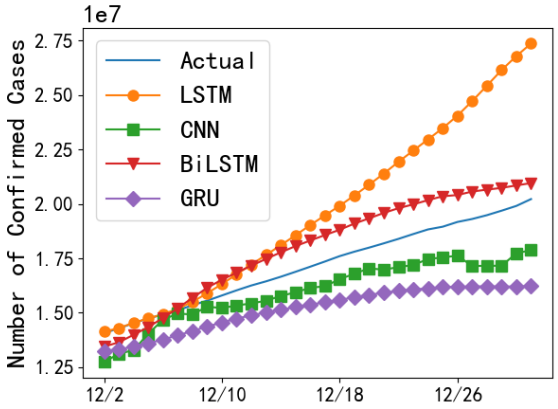


Fig. 11. Long-term prediction of the original virus

In this part of the study, the experiment does not add sentiment score to the model input. In these models, the accuracy of long-term forecasting is generally higher than that of short-term forecasting. Next, the study is going to verify that adding the results of emotion analysis to the model input could actually improve the performance of models.

3) Correction of forecasts by sentiment data: In the experiment described in section 5.2.1, we train the models using a dataset that does not include sentiment data. We then demonstrate the performance of models in predicting epidemics caused by different variants. The results indicate that the deep learning model can increase its accuracy after adding more data samples from different periods to the training data. However, predictive accuracy needs further improvement. It is found that adding more explanatory variables to the model input could help it correct the results.

Conversely, suppose the input of the models is variables with little or no correlation. In that case, it may reduce the predictive accuracy. After introducing a new mathematical model, Nanning et al. [1] also joined the LSTM model to correct the prediction results. The NLP module was also added to the model, indirectly improving predictive accuracy. The input of the NLP model contains a large amount of text information strongly related to the epidemic.

We collected approximately 500,000 tweets related to the pandemic to validate our idea. After completing the text preprocessing, we utilized the pre-trained model released by HuggingFace [4] to score sentiment on the text data. The model is trained on a dataset containing 54 million English tweets. It can classify the emotions of English tweets, as shown in Table V. We then use the classification results as input to the Granger causality test, which helps us verify whether the current variable is related to the variables we

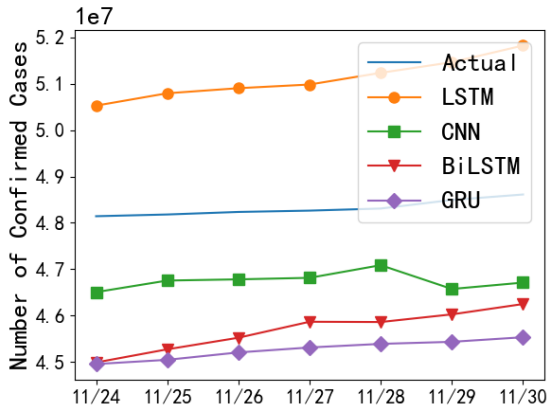


Fig. 12. Short-term prediction of the Delta variant

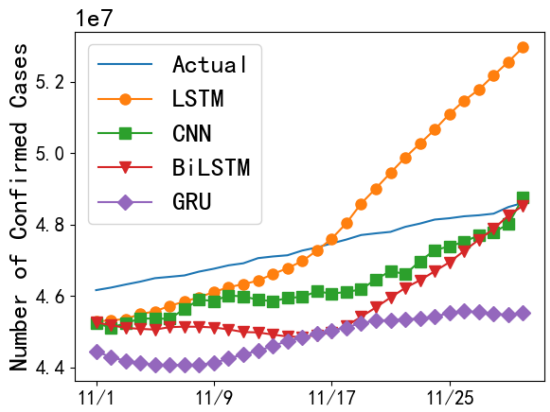


Fig. 13. Long-term prediction of the Delta variant

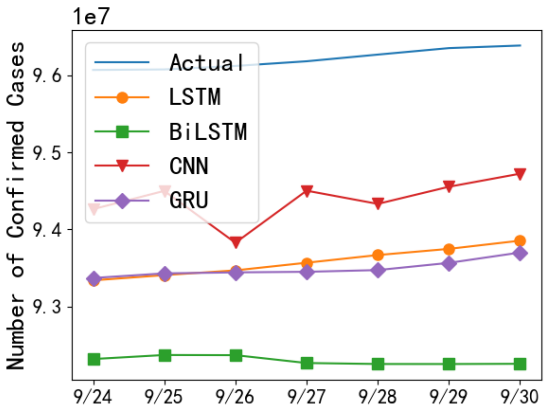


Fig. 14. Short-term prediction of the Omicron variant

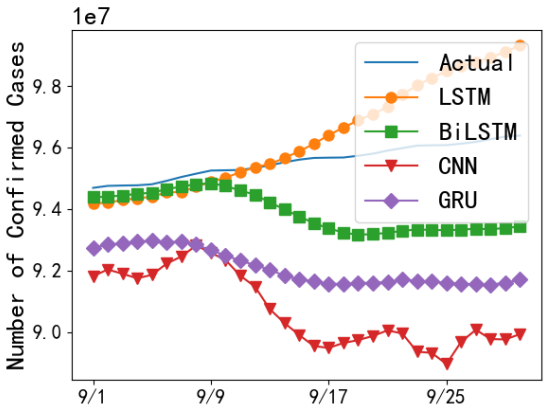


Fig. 15. Long-term prediction of the Omicron variant

want to forecast. Finally, strongly correlated variables are used as additional explanatory variables. We will combine these variables with the transmission features of the COVID-19 epidemic to improve the predictive accuracy of the time series models.

From the experiment in section 5.2.1, it is evident that these four models perform better in long-term forecasting than in short-term forecasting. The following experiment will focus on verifying the improvement in accuracy for long-term forecasting with the inclusion of sentiment data. In the context of coronavirus disease, long-term forecasting is more meaningful than short-term forecasting.

After training these four models, we observe that the inclusion of sentiment data improves the predictive accuracy of the models. The results are shown in Fig. 16 to 18 and the error results of the models are shown in Table VI to Table VIII.

For the original virus, the predictive accuracy of long-term prediction is generally improved after incorporating sentiment data. Regarding short-term forecasting, the CNN model performs the best with an MSE result of 0.022. The CNN model also achieves the best long-term forecasting with an MSE result of 0.003.

In short-term forecasting, the BiLSTM model performs the best for the Delta variant with an MSE result of 0.002. The LSTM model performs the best in long-term forecasting with an MSE result of 0.01.

For the Omicron variant, when the sentiment data is added to the model input, LSTM and BiLSTM perform better in long-term forecasting. At the same time, CNN and GRU excelled in short-term forecasting. Regarding long-term forecasting, the LSTM model performs the best with an MSE result of 0.003. In short-term prediction, the CNN model performed the best with an MSE result of 0.006. Overall, incorporating more strongly correlated variables into the model's input variables led to improved prediction performance. The sentiment classification results were instrumental in enhancing the accuracy of the predictions made by the models.

C. Evolution of Strain Properties

In the experiment conducted in section 5.2, we demonstrated the predictive accuracy of models on different variants under short-term and long-term forecasting. Also, adding classification results to the model input could improve the predictions. In the upcoming experiments, we will use these enhanced models to forecast the COVID-19 epidemiological trends during different periods.

We trained the models using data from three periods to predict epidemiological trends during the current stage and different stages. Sometimes, we could not obtain all the variables listed in Table I, which requires us to adjust

the input of the models promptly. We will compare the experimental results in this chapter with the results from Chapter 5. The results indicate that when the model was trained with data from December 2021 to October 2022 and then used these models to predict the epidemiological trends from March 2021 to November 2021 and from March 2020 to November 2020, the performance was poor, and vice versa. The models were only effective in predicting infection trends for the same stage. For example, when we used a model trained with 2020 pandemic data to predict the infection trends in 2021, we found that the model's prediction performance significantly deteriorated. It suggests a change in the characteristics of the virus in 2021, indicating the emergence of a new variant. According to the findings of detection agencies, the Delta variant emerged in 2021. Therefore, time series prediction is sensitive to time. When there is a significant deviation in forecasting, it implies a change in the characteristics of the virus. The models trained on previous data are no longer applicable for predicting the epidemiological trends of new variants. The error results of the models in this experiment are presented in Table IX.

VI. Experiment Summary

The experiment includes two parts: The first part is the experiment in section 5.2. In this part, we used four deep learning models to predict the epidemiological trends of different variants in long-term and short-term forecasting. It was found that the performance of the models improved when the model input was augmented with sentiment data. The other part of the experiment is in section 5.3. Since the time series data used in our experiment does not exhibit periodicity, the models cannot be reused for long-term forecasting. When the predictive accuracy of the models declines, it indicates a change in the characteristics of the virus.

VII. Conclusion

This study employed four deep learning models. We incorporated sentiment data into the model input and used the improved models to predict epidemiological trends. From the error data of the models, sentiment data helps refine the predictions. Furthermore, the time series models exhibit time sensitivity. In addition to studying the characteristics of variants using biotechnology in epidemic prevention and control, we could also reflect the changes in viral properties through forecasting. The experimental results of this study provide feasible approaches for studying the characteristics of coronavirus. They can assist relevant institutions in formulating more rational policies.

Appendix

The authors declare that they have no competing interests. The first author proposed the idea, wrote the manuscript and did all the related computational study with the direction given by the second author. The second author reviewed the manuscript and supervised it. All authors read and approved the final manuscript.

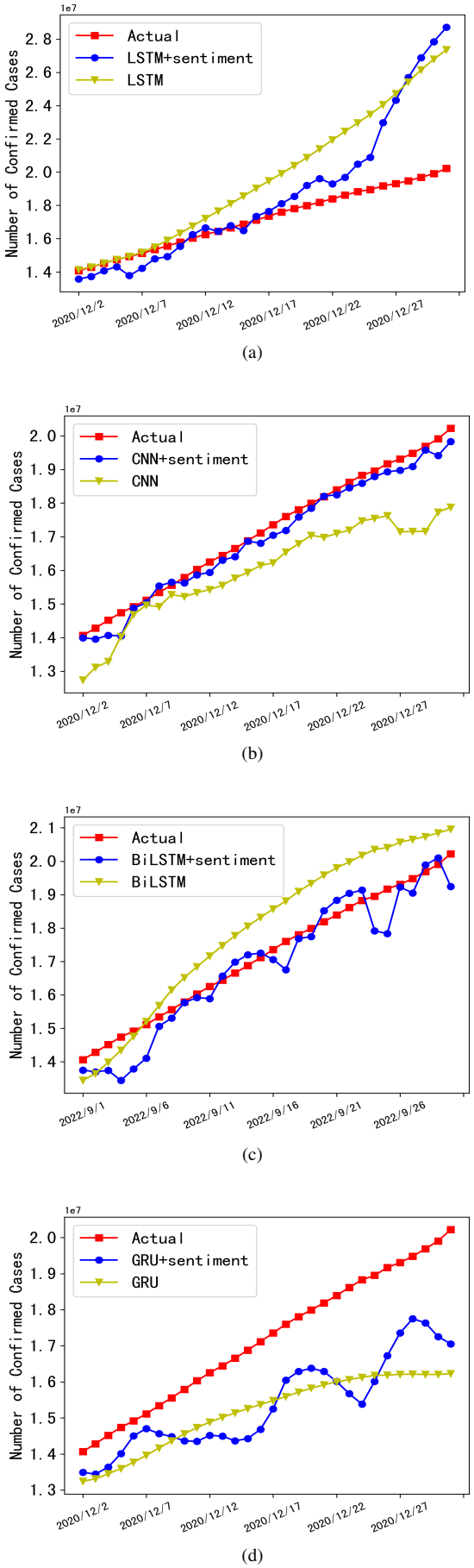
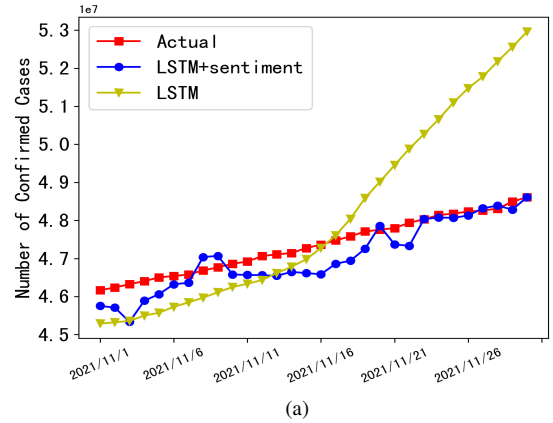
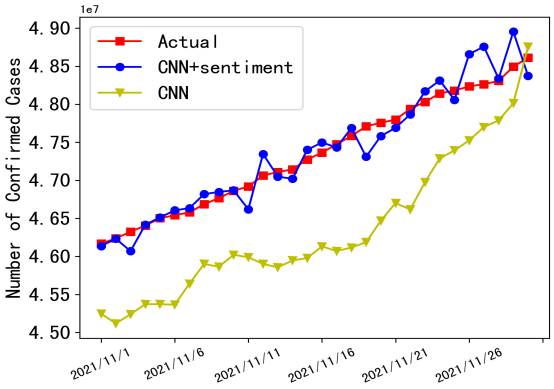


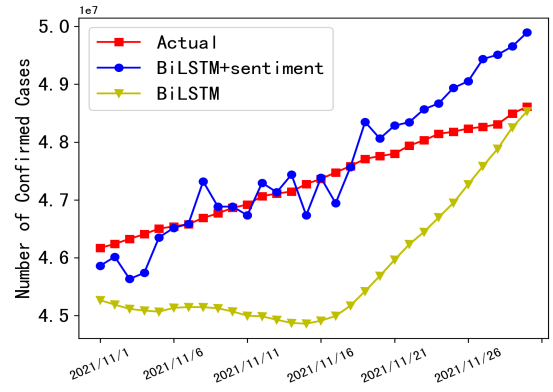
Fig. 16. Long-term prediction of the original virus



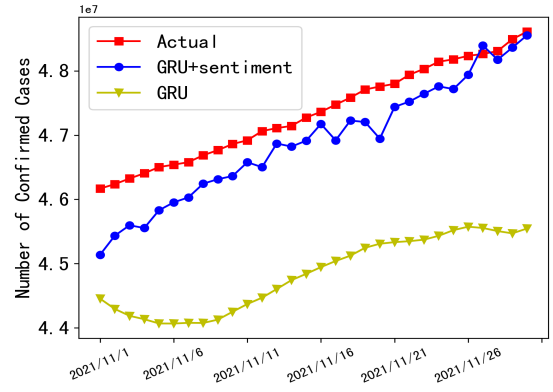
(a)



(b)

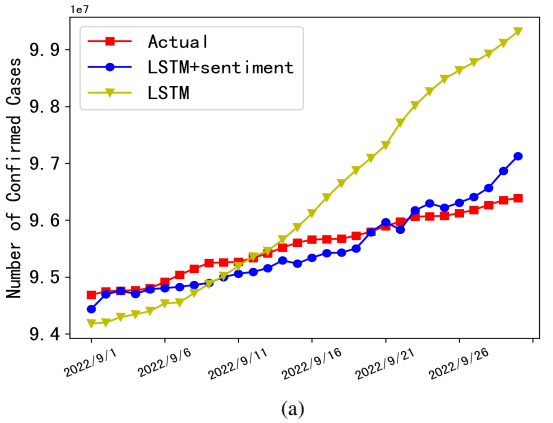


(c)

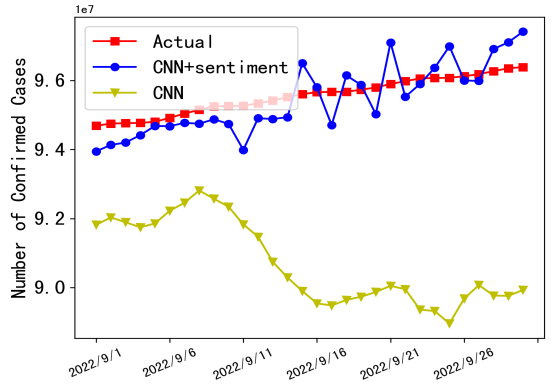


(d)

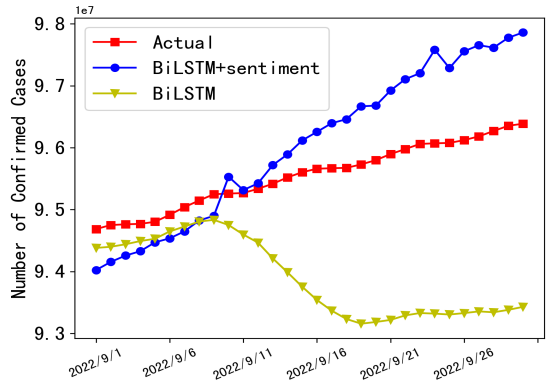
Fig. 17. Long-term prediction of the Delta variant



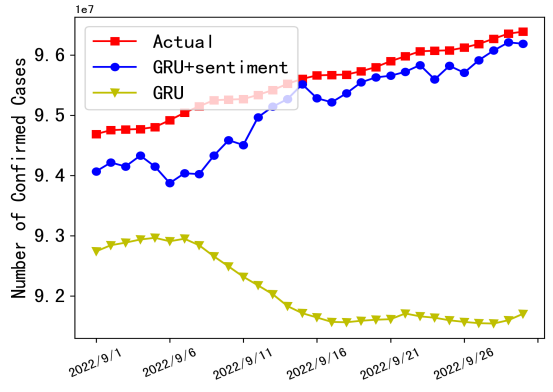
(a)



(b)

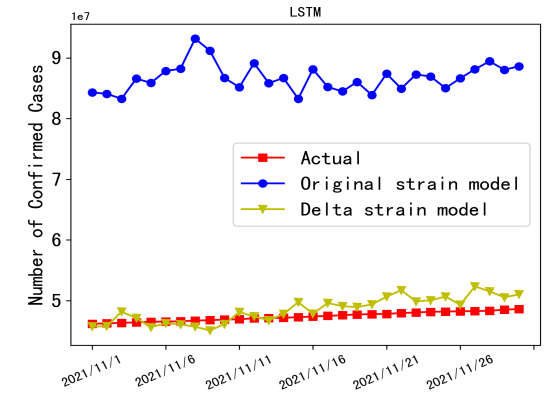


(c)

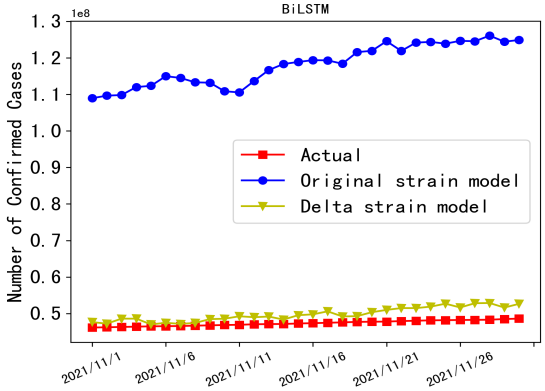


(d)

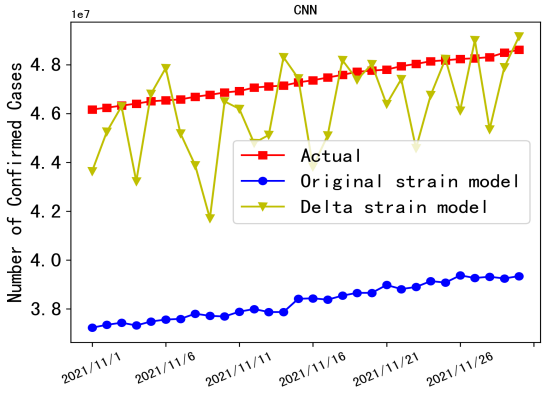
Fig. 18. Long-term prediction of the Omicron variant



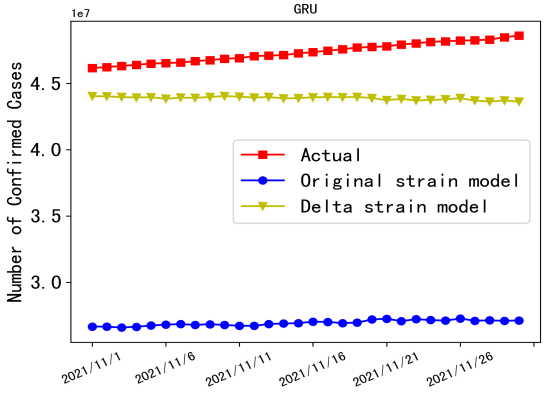
(a)



(b)

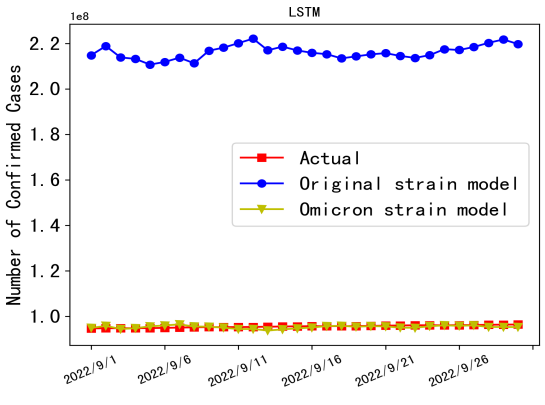


(c)

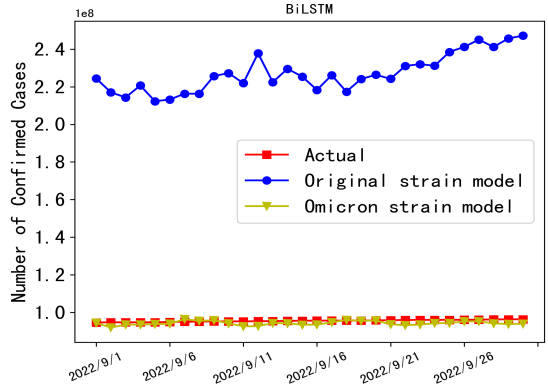


(d)

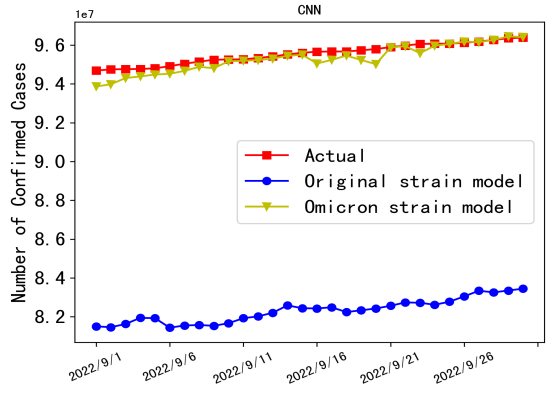
Fig. 19. Period of Delta variant transmission



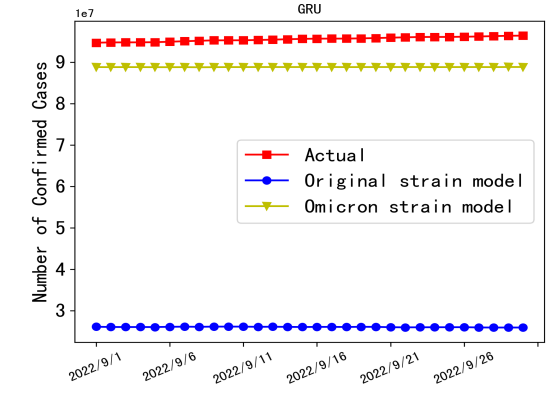
(a)



(b)

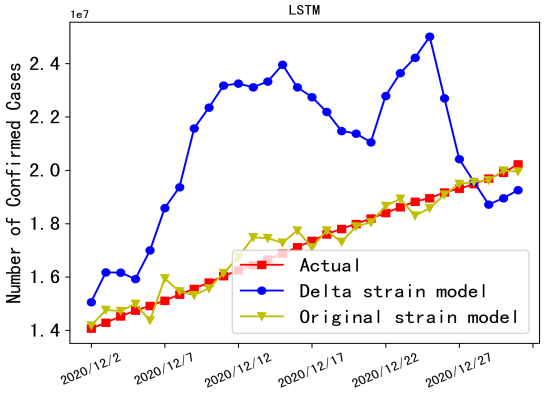


(c)

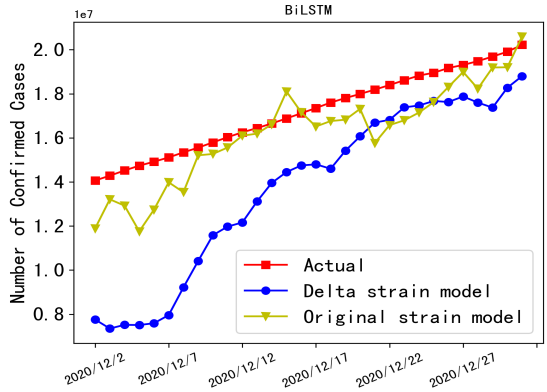


(d)

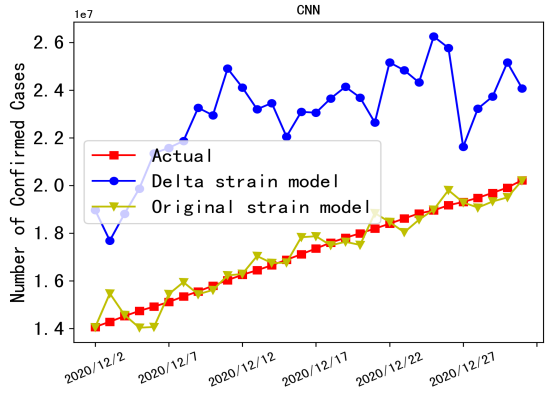
Fig. 20. Period of Omicron variant transmission



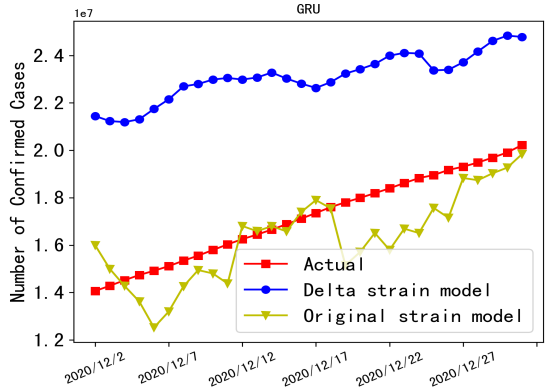
(a)



(b)

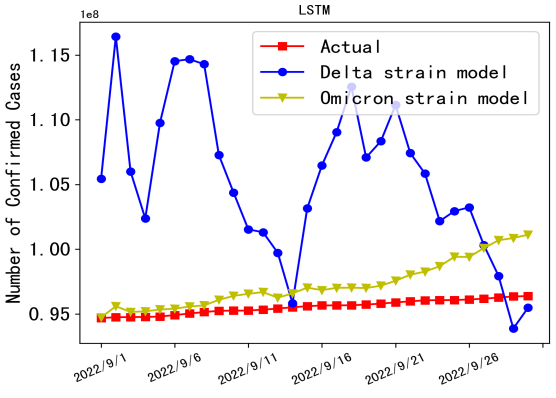


(c)

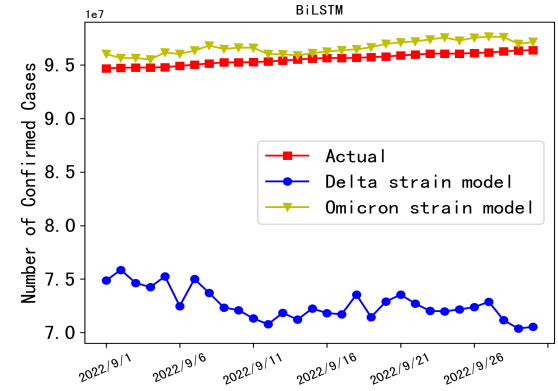


(d)

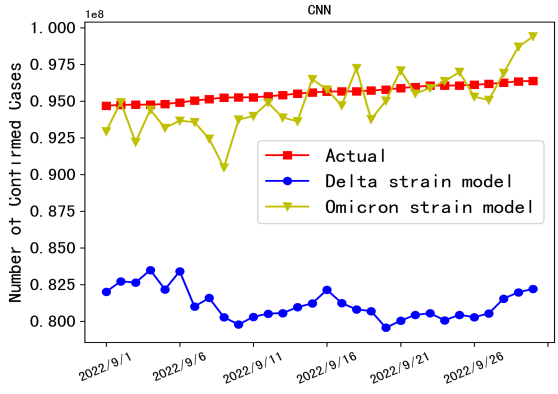
Fig. 21. Period of original virus transmission



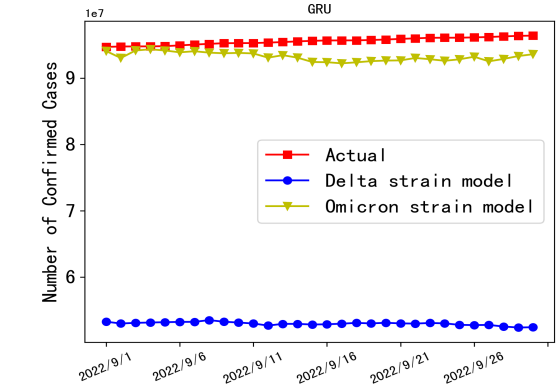
(a)



(b)

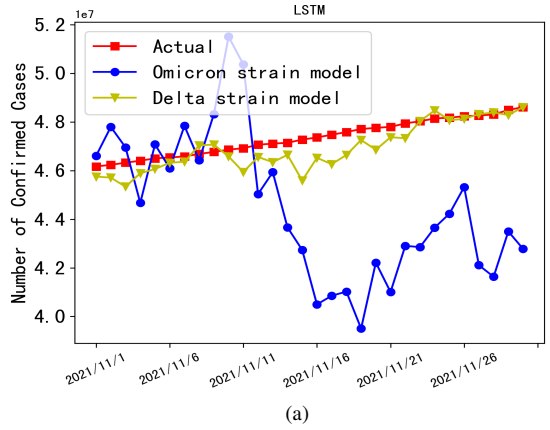


(c)

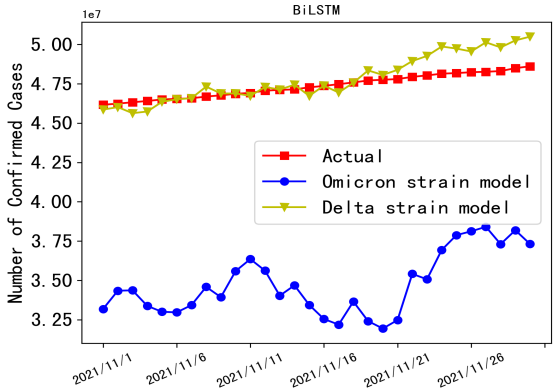


(d)

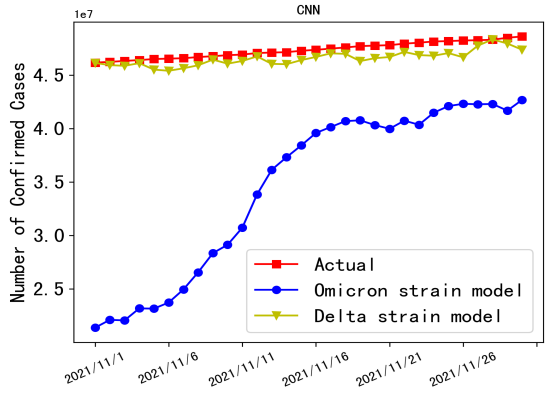
Fig. 22. Period of Omicron variant transmission



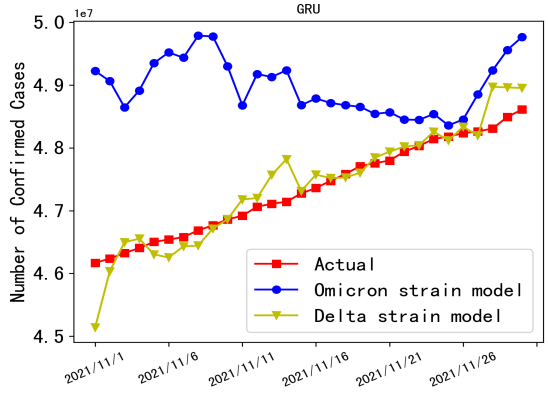
(a)



(b)

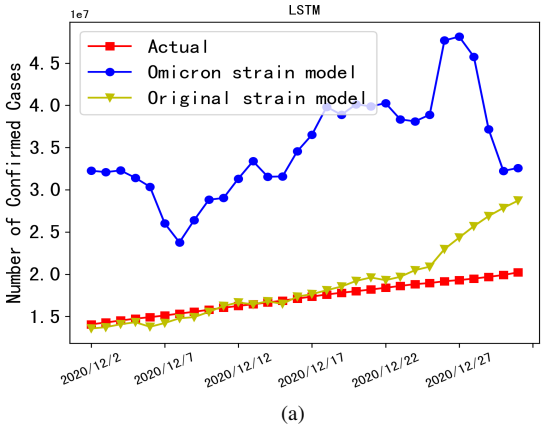


(c)

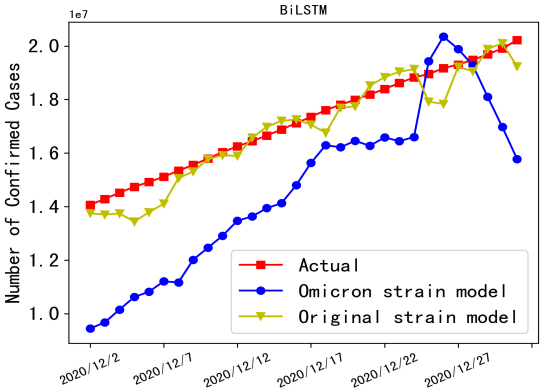


(d)

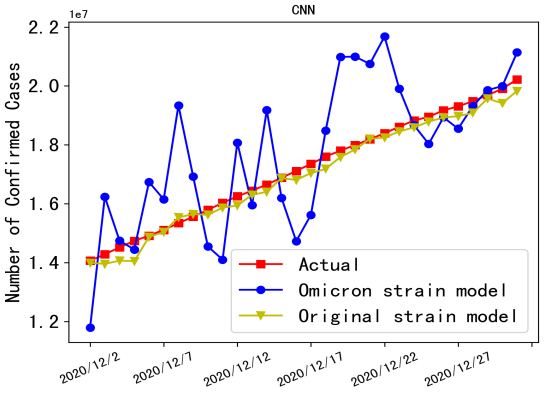
Fig. 23. Period of Delta variant transmission



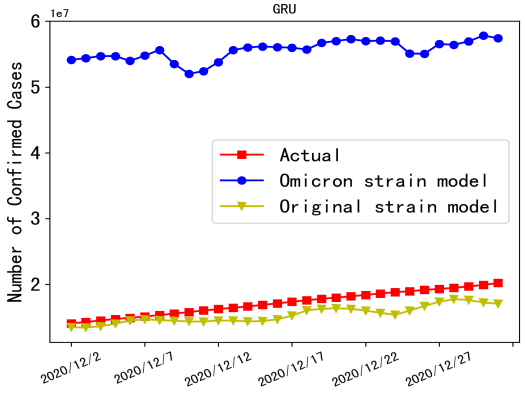
(a)



(b)



(c)



(d)

Fig. 24. Period of original virus transmission

TABLE V
ANALYSIS OF TEXT SENTIMENT BY THE ROBERTA MODEL

Text	Negative	Neutral	Positive	Classification Results
We went searching world for Omicron, but it was right here in our California all along.	0.029	0.693	0.279	neutral
pay attention to the news. the US finna go into a whole shut down again with this new omicron shit	0.862	0.13	0.008	negative
We're excited to continue to work towards creating inclusive communities in a post covid era.	0.001	0.019	0.98	positive

TABLE VI
PERFORMANCE METRICS OF THE MODEL (ORIGINAL VIRUS)

Strain Type	Experiment Classification	Index	Models			
			LSTM	CNN	BiLSTM	GRU
Original virus	Short-term forecasting (no sentiment data)	mse	1.142	0.089	0.073	0.598
		mae	1.055	0.289	0.267	0.769
		mape	0.391	0.107	0.1	0.286
	Short-term forecasting (Added sentiment data)	mse	**0.113**	**0.022**	**0.042**	**0.25**
		mae	**0.285**	**0.103**	**0.172**	**0.483**
		mape	**0.108**	**0.038**	**0.065**	**0.179**
	Long-term forecasting (no sentiment data)	mse	0.424	0.065	0.04	0.18
		mae	0.484	0.227	0.185	0.388
		mape	0.192	0.099	0.081	0.167
	Long-term forecasting (Added sentiment data)	mse	**0.352**	**0.003**	**0.014**	**0.147**
		mae	**0.356**	**0.046**	**0.092**	**0.349**
		mape	**0.141**	**0.021**	**0.043**	**0.151**

TABLE VII
PERFORMANCE METRICS OF THE MODEL (DELTA VARIANT)

Strain Type	Experiment Classification	Index	Models			
			LSTM	CNN	BiLSTM	GRU
Delta variant	Short-term forecasting (no sentiment data)	mse	0.201	0.065	0.181	0.24
		mae	0.447	0.252	0.423	0.49
		mape	0.229	0.129	0.217	0.251
	Short-term forecasting (Added sentiment data)	mse	**0.018**	**0.065**	**0.002**	**0.134**
		mae	**0.114**	**0.233**	**0.041**	**0.363**
		mape	**0.058**	**0.119**	**0.021**	**0.186**
	Long-term forecasting (no sentiment data)	mse	0.095	0.029	0.073	0.164
		mae	0.233	0.162	0.249	0.403
		mape	0.125	0.091	0.14	0.224
	Long-term forecasting (Added sentiment data)	mse	**0.01**	**0.02**	**0.022**	**0.085**
		mae	**0.08**	**0.127**	**0.107**	**0.283**
		mape	**0.045**	**0.07**	**0.057**	**0.156**

TABLE VIII
PERFORMANCE METRICS OF THE MODEL (OMICRON VARIANT)

Strain Type	Experiment Classification	Index	Models			
			LSTM	CNN	BiLSTM	GRU
Omicron variant	Short-term forecasting (no sentiment data)	mse	0.042	0.02	0.092	0.044
		mae	0.204	0.141	0.303	0.211
		mape	0.165	0.114	0.245	0.17
	Short-term forecasting (Added sentiment data)	mse	**0.032**	**0.006**	**0.036**	**0.007**
		mae	**0.176**	**0.061**	**0.168**	**0.083**
		mape	**0.142**	**0.049**	**0.135**	**0.067**
	Long-term forecasting (no sentiment data)	mse	0.012	0.157	0.024	0.024
		mae	0.084	0.375	0.128	0.128
		mape	0.069	0.313	0.106	0.106
	Long-term forecasting (Added sentiment data)	mse	**0.003**	**0.011**	**0.007**	**0.012**
		mae	**0.047**	**0.095**	**0.082**	**0.095**
		mape	**0.04**	**0.079**	**0.069**	**0.079**

TABLE IX
SUMMARY OF MODEL ERROR RESULTS

Experimental group	Experiments: Grouping of experiments	Index	Models: LSTM	CNN	BiLSTM	GRU
Fig. 19	Models trained using data from original virus wave periods -> Epidemiological trends caused by Delta variant	MSE	58.133	3.055	187.745	15.585
		MAE	7.612	1.748	13.67	3.946
		MAPE	0.945	0.217	1.694	0.489
	Models trained using data from Delta variant wave periods -> Epidemiological trends caused by Delta variant (Adjustment for the number of variables)	MSE	0.127	0.144	0.271	0.473
		MAE	0.297	0.291	0.467	0.669
		MAPE	0.037	0.036	0.058	0.083
Fig. 20	Models trained using data from original virus wave periods -> Epidemiological trends caused by Omicron variant	MSE	544.123	6.558	650.556	180.673
		MAE	23.32	2.56	25.439	13.441
		MAPE	1.342	0.147	1.463	0.773
	Models trained using data from Omicron variant wave periods -> Epidemiological trends caused by Omicron variant (Adjustment for the number of variables)	MSE	0.023	0.586	0.107	1.715
		MAE	0.122	0.729	0.283	1.306
		MAPE	0.007	0.042	0.016	0.075
Fig. 21	Models trained using data from Delta variant wave periods -> Epidemiological trends caused by original virus	MSE	0.808	0.911	0.422	0.926
		MAE	0.595	0.926	0.553	0.949
		MAPE	0.185	0.305	0.174	0.31
	Models trained using data from original virus wave periods -> Epidemiological trends caused by original virus (Adjustment for the number of variables)	MSE	1.376	0.332	0.151	0.158
		MAE	1.046	0.401	0.285	0.364
		MAPE	0.357	0.14	0.099	0.124
Fig. 22	Models trained using data from Delta variant wave periods -> Epidemiological trends caused by Omicron variant	MSE	3.433	5.371	13.601	46.677
		MAE	1.61	2.307	3.676	6.831
		MAPE	0.169	0.242	0.386	0.717
	Models trained using data from Omicron variant wave periods -> Epidemiological trends caused by Omicron variant (No adjustment to the number of variables is required.)	MSE	0.003	0.011	0.007	0.012
		MAE	0.047	0.095	0.082	0.095
		MAPE	0.04	0.079	0.069	0.079
Fig. 23	Models trained using data from Omicron variant wave periods -> Epidemiological trends caused by Delta variant	MSE	0.12	1.271	0.968	0.158
		MAE	0.293	0.986	0.975	0.364
		MAPE	0.116	0.383	0.383	0.124
	Models trained using data from Delta variant wave periods -> Epidemiological trends caused by Delta variant (No adjustment to the number of variables is required.)	MSE	0.01	0.02	0.022	0.085
		MAE	0.08	0.127	0.107	0.283
		MAPE	0.045	0.07	0.057	0.156
Fig. 24	Models trained using data from Omicron variant wave periods -> Epidemiological trends caused by original virus	MSE	2.046	1.15	0.32	55.141
		MAE	1.379	1.071	0.509	7.422
		MAPE	0.284	0.219	0.25	3.431
	Models trained using data from original virus wave periods -> Epidemiological trends caused by original virus (No adjustment to the number of variables is required.)	MSE	0.352	0.003	0.014	0.147
		MAE	0.356	0.046	0.092	0.349
		MAPE	0.141	0.021	0.043	0.151

REFERENCES

[1] N. Zheng, S. Du, J. Wang, H. Zhang, W. Cui, Z. Kang, T. Yang, B. Lou, Y. Chi, H. Long *et al.*, "Predicting COVID-19 in China using hybrid AI model," *IEEE Transactions on Cybernetics*, vol. 50, no. 7, pp. 2891–2904, 2020.

[2] S. Mohan, A. K. Solanki, H. K. Taluja, A. Singh *et al.*, "Predicting the impact of the third wave of COVID-19 in India using hybrid statistical machine learning models: A time series forecasting and sentiment analysis approach," *Computers in Biology and Medicine*, vol. 144, p. 105354, 2022.

[3] N. S. Sattar and S. Arifuzzaman, "COVID-19 vaccination awareness and aftermath: public sentiment analysis on Twitter data and vaccinated population prediction in the USA," *Applied Sciences*, vol. 11, no. 13, p. 6128, 2021.

[4] X. He, W. Hong, X. Pan, G. Lu, and X. Wei, "SARS-CoV-2 Omicron variant: characteristics and prevention," *MedComm*, vol. 2, no. 4, pp. 838–845, 2021.

[5] Z. Zhao, W. Chen, X. Wu, P. C. Chen, and J. Liu, "LSTM network: a deep learning approach for short-term traffic forecast," *IET Intelligent Transport Systems*, vol. 11, no. 2, pp. 68–75, 2017.

[6] R. Wang, Y. Hozumi, C. Yin, and G.-W. Wei, "Mutations on COVID-19 diagnostic targets," *Genomics*, vol. 112, no. 6, pp. 5204–5213, 2020.

[7] D. Fanelli and F. Piazza, "Analysis and forecast of COVID-19 spreading in China, Italy and France," *Chaos, Solitons & Fractals*, vol. 134, p. 109761, 2020.

[8] R. M. A. Velásquez and J. V. M. Lara, "Forecast and evaluation of COVID-19 spreading in USA with reduced-space Gaussian process regression," *Chaos, Solitons & Fractals*, vol. 136, p. 109924, 2020.

[9] R. Salgotra, M. Gandomi, and A. H. Gandomi, "Time series analysis and forecast of the COVID-19 pandemic in India using genetic programming," *Chaos, Solitons & Fractals*, vol. 138, p. 109945, 2020.

[10] H. M. Paiva, R. J. M. Afonso, I. L. de Oliveira, and G. F. Garcia, "A data-driven model to describe and forecast the dynamics of COVID-19 transmission," *PloS One*, vol. 15, no. 7, p. e0236386, 2020.

[11] M. Maleki, M. R. Mahmoudi, D. Wraith, and K.-H. Pho, "Time series modelling to forecast the confirmed and recovered cases of COVID-19," *Travel Medicine and Infectious Disease*, vol. 37, p. 101742, 2020.

[12] D. Painuli, D. Mishra, S. Bhardwaj, and M. Aggarwal, "Forecast and prediction of COVID-19 using machine learning," in *Data Science for COVID-19*. Elsevier, 2021, pp. 381–397.

[13] S. Zhang, S. Wang, L. Yuan, X. Liu, and B. Gong, "The impact of epidemics on agricultural production and forecast of COVID-19," *China Agricultural Economic Review*, vol. 12, no. 3, pp. 409–425, 2020.

[14] G. Grasselli, A. Pesenti, and M. Cecconi, "Critical care utilization for the COVID-19 outbreak in Lombardy, Italy: early experience and forecast during an emergency response," *Jama*, vol. 323, no. 16, pp. 1545–1546, 2020.

[15] R. Katoch and A. Sidhu, "An application of ARIMA model to forecast the dynamics of COVID-19 epidemic in India," *Global Business Review*, p. 0972150920988653, 2021.

[16] T. Saba, I. Abunadi, M. N. Shahzad, and A. R. Khan, "Machine learning techniques to detect and forecast the daily total COVID-19 infected and deaths cases under different lockdown types," *Microscopy Research and Technique*, vol. 84, no. 7, pp. 1462–1474, 2021.

[17] H. Tandon, P. Ranjan, T. Chakraborty, and V. Suhag, "Coronavirus (COVID-19): ARIMA-based Time-series Analysis to Forecast near Future and the Effect of School Reopening in India," *Journal of Health Management*, vol. 24, no. 3, pp. 373–388, 2022.

[18] S. Venkatramanan, J. Cambeiro, T. Liptay, B. Lewis, M. Orr, G. Dempsey, A. Telionis, J. Crow, C. Barrett, and M. Marathe, "Utility of human judgment ensembles during times of pandemic uncertainty: A case study during the COVID-19 Omicron BA. 1 wave in the USA," *MedRxiv*, pp. 2022–10, 2022.

[19] O. Gaidai, Y. Xing, and X. Xu, "COVID-19 epidemic forecast in USA East coast by novel reliability approach," 2022.

[20] M. Goic, M. S. Bozanic-Leal, M. Badal, and L. J. Basso, "COVID-19: Short-term forecast of ICU beds in times of crisis," *Plos One*, vol. 16, no. 1, p. e0245272, 2021.

[21] J. Fang, X. Zhang, Y. Tong, Y. Xia, H. Liu, and K. Wu, "Baidu index and COVID-19 epidemic forecast: evidence from China," *Frontiers in Public Health*, vol. 9, p. 685141, 2021.

[22] Y. Sugishita, J. Kurita, T. Sugawara, and Y. Ohkusa, "Forecast of the COVID-19 outbreak, collapse of medical facilities, and lockdown effects in Tokyo, Japan," *MedRxiv*, pp. 2020–04, 2020.

[23] S. Singh, K. S. Parmar, J. Kumar, and S. J. S. Makkhan, "Development of new hybrid model of discrete wavelet decomposition and autoregressive integrated moving average (ARIMA) models in application to one month forecast the casualties cases of COVID-19," *Chaos, Solitons & Fractals*, vol. 135, p. 109866, 2020.

[24] S. Shastri, K. Singh, S. Kumar, P. Kour, and V. Mansotra, "Deep-LSTM ensemble framework to forecast COVID-19: an insight to the global pandemic," *International Journal of Information Technology*, vol. 13, pp. 1291–1301, 2021.

[25] E. Orihuel, J. Sapena, and J. Navarro-Ortiz, "An empirical algorithm for COVID-19 nowcasting and short-term forecast in Spain: a kinematic approach," *Applied System Innovation*, vol. 4, no. 1, p. 2, 2021.

[26] A. A. Davidescu, S.-A. Apostu, and L. A. Stoica, "Socioeconomic Effects of COVID-19 Pandemic: Exploring Uncertainty in the Forecast of the Romanian Unemployment Rate for the Period 2020–2023," *Sustainability*, vol. 13, no. 13, p. 7078, 2021.

[27] M. Carlsson, J. Wittsten, and C. Söderberg-Nauclér, "A note on variable susceptibility, the herd-immunity threshold and modeling of infectious diseases," *Plos One*, vol. 18, no. 2, p. e0279454, 2023.

[28] K. Sarkar, S. Khajanchi, and J. J. Nieto, "Modeling and forecasting the COVID-19 pandemic in India," *Chaos, Solitons & Fractals*, vol. 139, p. 110049, 2020.

[29] M. Shawaqfah and F. Almomani, "Forecast of the outbreak of COVID-19 using artificial neural network: Case study Qatar, Spain, and Italy," *Results in Physics*, vol. 27, p. 104484, 2021.

[30] S. Singh, P. Raj, R. Kumar, and R. Chaujar, "Prediction and forecast for COVID-19 Outbreak in India based on Enhanced Epidemiological Models," in *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*. IEEE, 2020, pp. 93–97.

[31] 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," *Computers in Biology and Medicine*, vol. 146, p. 105560, 2022.

[32] G. Soldi, N. Forti, D. Gaglione, P. Braca, L. M. Millefiori, S. Marano, P. K. Willett, and K. R. Pattipati, "Quickest detection and forecast of pandemic outbreaks: Analysis of COVID-19 waves," *IEEE Communications Magazine*, vol. 59, no. 9, pp. 16–22, 2021.

[33] K. Prieto, "Current forecast of COVID-19 in Mexico: A Bayesian and machine learning approaches," *Plos One*, vol. 17, no. 1, p. e0259958, 2022.

[34] N. Bakhtiarvand, M. Khashei, M. Mahnam, and S. Hajiahmadi, "A novel reliability-based regression model to analyze and forecast the severity of COVID-19 patients," *BMC Medical Informatics and Decision Making*, vol. 22, no. 1, p. 123, 2022.

[35] A. Adiga, D. Dubhashi, B. Lewis, M. Marathe, S. Venkatramanan, and A. Vullikanti, "Mathematical models for COVID-19 pandemic: a comparative analysis," *Journal of the Indian Institute of Science*, vol. 100, no. 4, pp. 793–807, 2020.

[36] N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi, R. Lu *et al.*, "A novel coronavirus from patients with pneumonia in China, 2019," *New England Journal of Medicine*, 2020.

[37] S. Gunay, G. Can, and M. Ocak, "Forecast of China's economic growth during the COVID-19 pandemic: a MIDAS regression analysis," *Journal of Chinese Economic and Foreign Trade Studies*, vol. 14, no. 1, pp. 3–17, 2021.

[38] F. J. J. Joseph, "Time series forecast of Covid 19 pandemic using auto recurrent linear regression," *Journal of Engineering Research*, 2022.

[39] A. Bhattacharyya, D. Bhowmik, and J. Mukherjee, "Forecast and interpretation of daily affected people during 21 days lockdown due to COVID 19 pandemic in India," *MedRxiv*, pp. 2020–04, 2020.

[40] S. M. Zandavi, T. H. Rashidi, and F. Vafaee, "Dynamic hybrid model to forecast the spread of COVID-19 using LSTM and behavioral models under uncertainty," *IEEE Transactions on Cybernetics*, vol. 52, no. 11, pp. 11 977–11 989, 2021.

[41] R. Megna, "First month of the epidemic caused by COVID-19 in Italy: current status and real-time outbreak development forecast," *Global Health Research and Policy*, vol. 5, no. 1, pp. 1–7, 2020.

[42] C. Foroni, M. Marcellino, and D. Stevanovic, "Forecasting the Covid-19 recession and recovery: Lessons from the financial crisis," *International Journal of Forecasting*, vol. 38, no. 2, pp. 596–612, 2022.

[43] T. A. Makarovskikh and M. S. Abotaleb, "Automatic selection of ARIMA model parameters to forecast COVID-19 infection and death cases," *Vestnik Yuzhno-Ural'skogo Gosudarstvennogo Universiteta. Seriya" Vychislitelnaya Matematika i Informatika"*, vol. 10, no. 2, pp. 20–37, 2021.

[44] V. K. R. Chimmula and L. Zhang, "Time series forecasting of COVID-19 transmission in Canada using LSTM networks," *Chaos, Solitons & Fractals*, vol. 135, p. 109864, 2020.

[45] 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, p. 110212, 2020.

[46] M. Arbane, R. Benlamri, Y. Brik, and A. D. Alahmar, "Social media-based COVID-19 sentiment classification model using Bi-LSTM," *Expert Systems with Applications*, vol. 212, p. 118710, 2023.

[47] S. R. Vadyala, S. N. Betgeri, E. A. Sherer, and A. Amritphale, "Prediction of the number of COVID-19 confirmed cases based on K-means-LSTM," *Array*, vol. 11, p. 100085, 2021.
[48] F. Liu, J. Wang, J. Liu, Y. Li, D. Liu, J. Tong, Z. Li, D. Yu, Y. Fan, X. Bi *et al.*, "Predicting and analyzing the COVID-19 epidemic in China: Based on SEIRD, LSTM and GWR models," *PloS One*, vol. 15, no. 8, p. e0238280, 2020.
[49] A. Akhtar, S. Akhtar, B. Bakhtawar, A. A. Kashif, N. Aziz, and M. S. Javeid, "COVID-19 detection from CBC using machine learning techniques," *International Journal of Technology, Innovation and Management (IJTIM)*, vol. 1, no. 2, pp. 65–78, 2021.
[50] F. M. Salman, S. S. Abu-Naser, E. Alajrami, B. S. Abu-Nasser, and B. A. Alashqar, "COVID-19 detection using artificial intelligence," 2020.
[51] S. Pokhrel and R. Chhetri, "A literature review on impact of COVID-19 pandemic on teaching and learning," *Higher Education for the Future*, vol. 8, no. 1, pp. 133–141, 2021.
[52] S. Rashid and S. S. Yadav, "Impact of COVID-19 pandemic on higher education and research," *Indian Journal of Human Development*, vol. 14, no. 2, pp. 340–343, 2020.
[53] B. A. Sethi, A. Sethi, S. Ali, and H. S. Aamir, "Impact of Coronavirus disease (COVID-19) pandemic on health professionals," *Pakistan Journal of Medical Sciences*, vol. 36, no. COVID19-S4, p. S6, 2020.
[54] S. Cheval, C. Mihai Adamescu, T. Georgiadis, M. Herrnegger, A. Piticar, and D. R. Legates, "Observed and potential impacts of the COVID-19 pandemic on the environment," *International Journal of Environmental Research and Public Health*, vol. 17, no. 11, p. 4140, 2020.
[55] E. Monchatre-Leroy, F. Boué, J.-M. Boucher, C. Renault, F. Moutou, M. Ar Gouilh, and G. Umhang, "Identification of alpha and beta coronavirus in wildlife species in France: bats, rodents, rabbits, and hedgehogs," *Viruses*, vol. 9, no. 12, p. 364, 2017.
[56] M. Scudellari, "How the coronavirus infects cells—and why Delta is so dangerous," 2021.
[57] V. M. Corman, O. Landt, M. Kaiser, R. Molenkamp, A. Meijer, D. K. Chu, T. Bleicker, S. Brünink, J. Schneider, M. L. Schmidt *et al.*, "Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR," *Eurosurveillance*, vol. 25, no. 3, p. 2000045, 2020.
[58] V. Shah, R. Keniya, A. Shridharani, M. Punjabi, J. Shah, and N. Mehendale, "Diagnosis of COVID-19 using CT scan images and deep learning techniques," *Emergency Radiology*, vol. 28, pp. 497–505, 2021.
[59] G. A. Miller, "WordNet: a lexical database for English," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.
[60] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *ArXiv Preprint ArXiv:1907.11692*, 2019.
[61] H. Sen, "Time Series Prediction based on Improved Deep Learning," *IAENG International Journal of Computer Science*, vol. 49, no. 4, pp. 1133–1138, 2022.
[62] X. Song, T. Wang, Z. Xu, and Z. Shi, "Multistep Forecasting Method of Short-term Power Load Based on VMD-Prophet-Seq2seq" *IAENG International Journal of Computer Science*, vol. 49, no. 3, pp. 720–727, 2022.
[63] L. Kang, X. Li, L. Wu, Y. Li, and X. Zhao, "Predicting Stock Closing Price with Stock Network Public Opinion Based on AdaBoost-IWOA-Elman Model and CEEMDAN Algorithm," *Surge*, vol. 1, p. 4, 2022.
[64] A. A. Syed, F. L. Gaol, W. Suparta, E. Abdurachman, A. Trisetyarso, and T. Matsuo, "Prediction of the Impact of COVID-19 Vaccine on Public Health Using Twitter," *IAENG International Journal of Computer Science*, vol. 49, no. 1, pp. 19–29, 2022.
[65] J. Che and H. Zhai, "Wt-arima combination modelling for short-term load forecasting," *IAENG International Journal of Computer Science*, vol. 49, no. 2, pp. 542–549, 2022.