International Journal of application or innovation in engineering and management

Ijaiem.com/June 2025/ Volume 14/Issue 1/Article No-1/857-874

ISSN: 2319-4847

Making COVID-19 Predictions in China with a Hybrid AI Model

[1] P. Premchand, [2] P. Sucharitha,

[1]Assistant Professor, Megha Institute of Engineering & Technology for Women, Ghatkesar.
[2] MCA Student, Megha Institute of Engineering & Technology for Women, Ghatkesar.

Article Info

Received: 29-04-2025 ***Revised: 06-06-2025*** ***Accepted: 17-06-2025*** ***Published:28/06/2025***

Abstract

Though the 2019 coronavirus illness (COVID-19) outbreak began in late December 2019 and is now being contained in China, it continues to spread quickly in many other parts of the globe. Research on the epidemiology and potential future developments of the pandemic must be undertaken immediately. To forecast the spread of COVID-19, this paper suggests a hybrid AI model. Firstly, an improved susceptible-infected (ISI) model is suggested to estimate the diversity of infection rates in order to analyze transmission laws and development trends, as standard epidemic models regard all persons with coronavirus as having the same infection rate. The second step in developing a hybrid AI model for COVID-19 prediction is to include the ISI model with a natural language processing (NLP) module and an LSTM network. This will allow us to account for the impact of control and preventive efforts as well as the rise in public awareness of the need of prevention. The experimental findings, based on epidemic data from many typical Chinese provinces and cities, demonstrate that the real laws of epidemic transmission are more accurately reflected by the greater infection rate experienced by infected people between the third and eighth days after infection. In comparison to conventional epidemic models, the suggested hybrid AI model achieves MAPPs of 0.52% for the next six days in Wuhan, 0.38% in Beijing, 0.05% in Shanghai, and 0.86% nationwide. This is in addition to a significant reduction in prediction errors.

Index Terms

Coronavirus disease 2019 (COVID-19) prediction, epidemic model, hybrid artificial-intelligence (AI) model, natural language processing (NLP).

I. INTRODUCTION

The 2019 coronavirus disease outbreak (COVID-19) occurred during China's spring festival season, coinciding with its rapid national expansion. There was a shortage of medical resources, a medical community that was unfamiliar with the novel coronavirus, and the very irregular nature of the main stage of the epidemic, all of which contributed to the ineffective suppression of the COVID-19 throughout its transmission [1]. It was officially confirmed on January 20, 2020 [2] that the COVID-19 may be communicated from person to person. As a result, China's cities and provinces have taken unprecedented steps to prevent and control the spread of the virus, with the airport and train terminals in Wuhan being closed on January 23, 2020. These efficient methods of prevention and management

have led to a steady rise in public understanding of the need of taking precautions against epidemics. The rate of new infections is now declining sharply. For sixteen days in a row, the number of newly confirmed cases outside of Hubei fell from February 3, 2020 to February 19, 2020. Meanwhile, new infections in Hubei have been steadily declining since February 12, 2020, and the number of patients who have been cured has been on the rise. Although there has been some progress in preventing and controlling the pandemic in China, other nations and areas are still facing a grave scenario, particularly in Europe, Iran, South Korea, the US, and Japan. If the pandemic is to be successfully contained, every nation or area must devise specific plans for prevention and control. Research into the causes and dynamics of epidemics is, hence, essential. To effectively avoid and manage this pandemic, it is important to analyze the development law and anticipate the trajectory of COVID-19. The use of epidemic models allows for the study and prediction of the illness's development trend, which in turn guides the creation of control and preventative measures in the event of a large-scale infectious disease epidemic and the initiation of a significant public health emergency. The most popular traditional models for epidemics are susceptible-infected (SI), susceptible-exposed-infected-recovered (SEIR) and susceptible-infected-recovered (SIR) [3]-[5]. In these models, "S," "E," "I," and "R" stand for the numbers of susceptible individuals, incubation period participants, infectious cases, and recovered individuals, respectively. The SI, SIR, and SEIR models use differential equations to depict the I-S connection. These A number of illnesses, including Ebola and SARS, have been effectively predicted using models due to their robust disease prediction capabilities [6]-[10]. The critical nature of the COVID-19 pandemic makes it all the more crucial to track changes in the daily confirmed case count in order to deduce the epidemic's trajectory. Hence, the effect of the pattern of new infections on epidemic propagation must be our primary concern. In addition, the article does not take into account the impact of death and cure rates on the epidemic trend as these two variables are not directly related to the number of new confirmed cases per day. Conventional epidemic models project the outbreak's trajectory by first analyzing the infection rate in relation to the changing number of infections. On the other hand, these models assume that the infection rate for coronavirus is constant throughout all cases. There are limits to their prediction findings since they can only provide broad trends. Transparent reporting of the epidemic, implementation of prevention and control measures, and reinforcement of residents' prevention awareness have accelerated the containment of the virus. The government's prevention and control measures have a significant impact on the containment of the epidemic's development trend. Clearly, reliable prediction cannot be achieved with just epidemic data. In order to handle public health crises, we need to create an epidemic model that is based on data. By including news information elements, we may enhance the accuracy of model prediction, overcome the constraint of classic epidemic models that rely on a single component, and confirm that the government's preventative and control methods are working. In order to address this issue, our epidemic model incorporates a long short-term memory (LSTM) network with a natural language processing (NLP) module. This allows us to update the infection rate and enhance the model's predicted accuracy. According to Hochreiter and Schmidhuber [11], Long Short-Term Memory (LSTM) is a traditional RNN. The continuous error carousel unit is an LSTM innovation that helps with training-related issues including gradient explosion and disappearance. For long-sequence data classification, processing, and prediction, LSTM is the way to go [13]-[16] since it outperforms classical RNN [12] in capturing sequence dependencies over the long term. Many tasks have seen increased application of LSTMs in recent years, including natural language processing [17]–[20], picture production [21], [22], and video analysis [23], [24]. This article presents an improved susceptible-infected (ISI) model based on an examination of the coronavirus infection rate, models the capacity of viruses to infect susceptible persons according to various times after infection, and focuses on the analysis of the infection rate of individuals. This article presents the important information about the great efforts led by the central and local governments, as well as the massive support participation from the public into the prediction calculation process, and explains how the hybrid artificial intelligence (AI) model based on the proposed ISI model predicts the COVID-19. The model includes an NLP module and an LSTM network. In addition, the study forecasts the epidemic's trajectory by analyzing its evolution through the lens of the suggested hybrid prediction model. Experimental results using epidemic data from multiple representative Chinese cities and provinces demonstrate that the suggested hybrid

model outperforms more conventional approaches to epidemic modeling in terms of accuracy and robustness, and can serve as a foundation for estimating the law of virus spread. Furthermore, by incorporating news information into our hybrid AI model, we were able to produce prediction results that were more in line with the actual trend of epidemic development. This highlights the importance of openness, transparency, and efficiency in data releasing for the establishment of a modern epidemic prevention system. The following is the structured rest of the article. The suggested AI model's structure is shown in Section II. In Section III, the ISI epidemic model is suggested for the purpose of studying the transmission laws of epidemics. An LSTM model that is based on natural language processing (NLP) is provided in Section IV for accurate prediction. Results from experiments using epidemic data from a number of representative Chinese provinces and localities are presented in Section V. You may find the conclusion in Section VI.

II. FRAMEWORK OF THE HYBRID AI MODEL

Current epidemic models predict that in the future, unquarantined coronavirus patients will be the source of infection for newly confirmed cases every day. Predicting the number of new confirmed cases each day is therefore accomplished by multiplying the expected infection rate by the number of sick patients who are not quarantined, which is considered the basis by most epidemic models [25]-[27]. On the other hand, coronavirus infection rates change at various points of time after infection [28]. Due to their assumption that all infected people have the same infection rate, traditional epidemic models fail to capture the pattern of an epidemic's progression. The majority of newly confirmed cases right now are due to infections that occurred in the last few days, despite efforts to prevent and limit the spread of the disease. Since cured and died patients do not affect the number of new confirmed cases in any way, they are not included in the epidemic model that is established in this article. On this premise, we provide a grouped multiparameter model for the ISI pandemic that is retrospective in nature. To determine the infection rate and build an epidemic model, the retrospective method relies on a ratio of the number of newly confirmed cases at time t to the total number of new confirmed cases across all time scales prior to time t. In addition, the model's prediction results are used to examine the significance of various time scales with respect to the newly confirmed instances at time t. To measure the infection rate of infected cases at various periods, the ISI model uses grouped multiparameter variables that figure out how confirmed cases at different times before t affect confirmed cases at time t. The next step is to examine the infectious illness development law using the upgraded model. Aside from that, the suggested ISI model is employed in conjunction with the LSTM network to calculate the number of infections and the variation in infection rates in the epidemic model.

International Journal of application or Innovation in engineering and management

Ijaiem.com/June 2025/ Volume 14/Issue 1/Article No-1/857-874

ISSN: 2319-4847

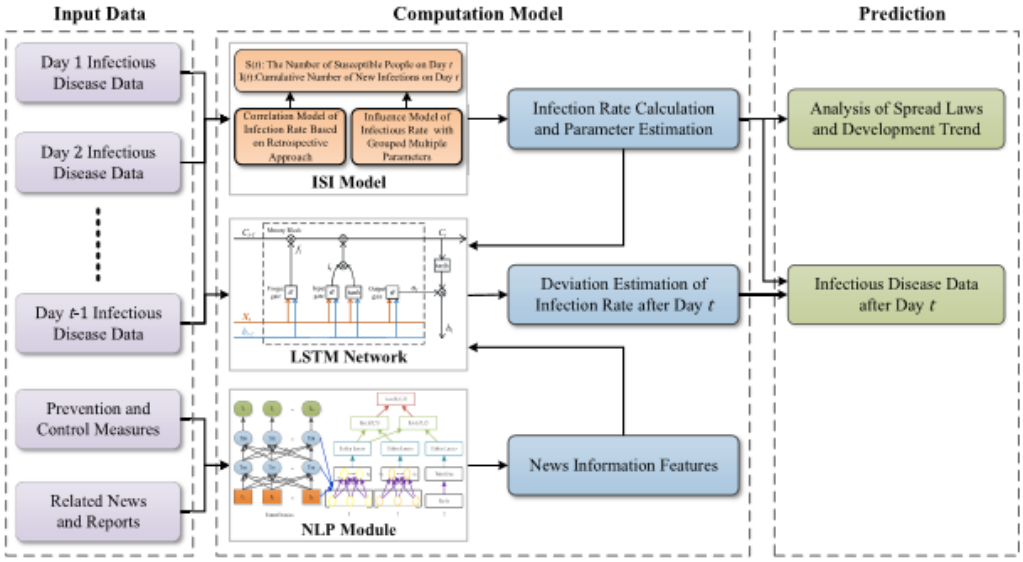

Fig. 1. Hybrid AI model for COVID-19 prediction by using all historical data.

sick individuals. In order to examine the impact of government control measures, the media's open reporting, and the growing public awareness of the need to avoid epidemics, this article employs pretrained natural language processing models to extract characteristics from pertinent news stories in different provinces and cities. After that, the LSTM network is used to fix the discrepancy in the infection rate projected by the ISI model. This model could forecast the number of cases based on the laws of transmission and the pattern of development. Figure 1 depicts the suggested structure.

III. ANALYSIS OF THE LAWS OF EPIDEMIC TRANSMISSION

There is still a lack of thorough examination in traditional epidemic models, which assume that the number of new infectious cases is proportional to the number of infected and susceptible individuals. Infectious illnesses have distinct life cycles that people experience [29]. In order to study the epidemic's infection law, it is necessary to identify the temporal distribution of the infectious sources of new confirmed cases everyday. Using fresh confirmed data as a basis, this essay models the epidemic's propagation rules and development tendency. This article does not take into account death and cure rates since they are unrelated to the amount of newly confirmed cases. We may assume that almost all newly confirmed cases of COVID-19 are caused by individuals who were confirmed during the last 14 days, as the observation period for this virus is 14 days [30]. A majority of the patients who are now being studied have been placed in quarantine, closely monitored, and tested using a nucleic acid reagent. Most of the confirmed patients were quarantined at least three days before the confirmation, so they cannot infect others. This means that most of the confirmed patients cannot be infected by another confirmed case that was confirmed eleven days ago, since patients are required to obtain a minimum of two positive results before they are confirmed as positive for COVID-19. So, this article looks at the infection rate of new daily confirmed cases in the last 10 days compared to the confirmed cases of day t for every day t. The following symbols are defined for an improved analysis: S(t) is the count of susceptible individuals on day t, I(t) is the total number of confirmed cases up to that point (inclusive), and $I(t) = I(t) - I(t-1)$ is the count of newly confirmed cases on that day. We need to find the window of opportunity for the confirmed cases to infect the newly confirmed patients at day t in order to have a full picture of how the infected cases affect subsequent affected people. The rules of transmission may then be determined by comparing and contrasting them. In this essay, we will assume that from day $t-1$ to day $t-10$, a confirmed individual infects the patients who confirmed on day t. We employ the retrospective technique to examine the

International Journal of application or Innovation in engineering and management

Ijaiem.com/June 2025/ Volume 14/Issue 1/Article No-1/857-874

ISSN: 2319-4847

temporal laws of epidemic transmission in the last few days in order to establish at what stage the present new daily confirmed patients are infected at a high infection rate. We analyze the temporal laws of COVID-19 transmission in detail and apply it to many infection periods to construct an improved multiparameter epidemic model over the previous 10 days. The computer model of the ISI outbreak is shown in Fig. 2.

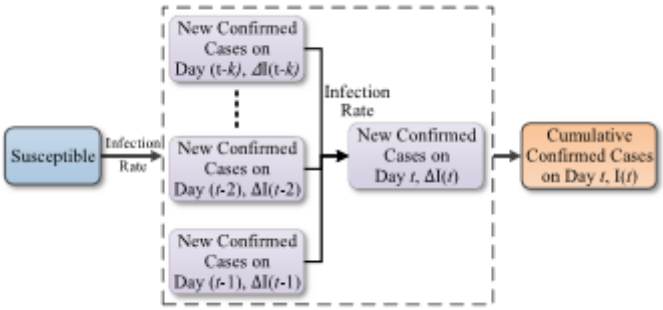

Fig. 2. ISI model.

based on the last k days. But these models don't look at the transmission of epidemics in detail. The general rules of epidemic development state that, on day t, early confirmed patients (e.g., day t−5) are more likely to infect individuals diagnosed on day t, in comparison to patients confirmed at the adjacent period (e.g., day t−1). In order to obtain the laws of COVID-19 transmission with improved macroscopic guiding significance for the overall trend estimation of epidemic development, one can model on the infection rate of the cumulative number of confirmed cases in the past k days relative to the confirmed cases on day t. We examine the impact of cumulative confirmed cases at various dates on the calculation of the infection rate using the retrospective technique. Based on the total number of confirmed cases over the last k days, we calculate the infection rate of newly confirmed cases and compare it to the infection rate in other locations and time periods. The following is the equation:

$$I(t) = I(t-1) + \beta_1(t,k)\sum_{i=1}^{k} \Delta I(t-i), \quad k = 1, 2, \ldots$$

A. Infection Rate Correlation Model Based on the Retrospective Approach The prevailing assumption in traditional epidemic models is that confirmed cases on a given day are a direct result of confirmed cases in the preceding several days. Previous studies on epidemics have often used models that assume transmission is impacted by a fixed number of days, as shown in [31], [32].

In this context, β1(t,k) represents the infection rate, which is defined as the sum of all confirmed cases from day t−k to day t−1, divided by the total number of confirmed cases on day t, as a function of k, where i=1. The sentence illustrates the connection between the quantity of newly confirmed cases I(t) on day t and the quantity of new confirmed cases k j=1 I(t−k) on the previous k days. In order to examine the effect of the total number of confirmed cases over the last k days on the number of new confirmed instances on day t, the first step is to find out the generally stable connection between I(t) and k i=1 I(t−i). This is known as β1(t,k). Our goal is to clarify the rules of epidemic propagation so that they may be fairly used and to provide assistance with further research. Secondly, all provinces and the nation as a whole may have their parameter ˆ β1(t,k) determined by (1). Since the rate of infection during an epidemic evolves exponentially, this article estimates the spread of the epidemic by using the exponential function L(t) = a×e−bt to fit ˆ β1(t,k). Both parameters, a and b, must be greater than zero for the exponential function to be defined in the formula. And lastly, since patients cannot be adequately confined during the incubation period, the infection rate is high. Hence, the section estimates β1(t,k) by progressively raising k, and the model is updated to reflect the number of additional confirmed instances

International Journal of application or Innovation in engineering and management

Ijaiem.com/June 2025/ Volume 14/Issue 1/Article No-1/857-874

ISSN: 2319-4847

at an earlier time point. whether the model's predictions hold, we may next determine whether the newly confirmed cases at each time point will infect the newly confirmed patients on day t. As the epidemic spreads, it is also possible to determine the rules of patient evolution at various time intervals.B. Grouped multiparameter influence model of infection rate According to this article, the stringent control and quarantine procedures ensure that infected cases cannot infect as many vulnerable individuals once they are contained. Newly confirmed patients in the previous k days are therefore likely to infect newly confirmed cases on day t. The rate of infection is strongly correlated with the duration of patient infection [33]. Infectious rates may therefore vary among newly confirmed cases at various points in the previous k days, as shown on day t. Based on (1), we may deduce that the most recent couple days are likely the earliest feasible infection time. From day t−k to day t−1, we assign various weights to the number of new confirmed cases each day in order to quantify the contribution of new confirmed cases at different periods to the infection rate at time t. This allows us to further evaluate the difference. The next step is to use the weighted cumulative confirmed number, the epidemic model, to estimate the infection rate. Two days next to each other are considered a propagation unit in order to simplify the model, and each day is given the same weight αi. After that, as seen in (2), multiparameter pandemic modeling is executed. While reducing the search area of the weight and the model's complexity, the model makes it more resilient by avoiding the abrupt change in weight induced by a single data irregularity.

$$I(t) = I(t-1) + \beta_2(t,k)\sum_{i=1}^{k/2}(\alpha_i(\Delta I(t-2i+1) + \Delta I(t-2i)))$$

where $2\sum_{i=1}^{k/2}\alpha_i = 1$.

when i equals 1, αi equals 1. This section proposes a model that, building on the previous epidemic model, takes into account the transmission correlation between the total number of confirmed cases over the past k days and the number of new cases on day t by taking into account the difference in the infection rate of new confirmed cases over the past k days compared to the new confirmed cases on day t. The process starts with randomly initializing many groups with various weights αi, and then a multiparameter epidemic model is set up using (2). The more accurate the model's predictions, the more closely the associated weights reflect the actual infection law. Lastly, by comparing the weights given to various time periods, we may deduce the infection rate that significantly contributes to the viral infection. Using the value of αi derived from (1) and (2), we can determine the link between the newly confirmed cases on day t and the new confirmed cases on days t−10 to t−1. On the other hand, underfitting may occur with insufficient parameters (i.e., (1)) and overfitting is easy with an excess of parameters (i.e., (2)). Accordingly, we use the aforementioned findings to further equalize the number of factors. The collection of days designated by {t−i|i=1,2,...,10}separated into two sets, with set A consisting of the days that had a stronger influence on the number of new confirmed cases on day t and set B consisting of the other days. Like in, set A is assigned a weight of γ1, while set B is assigned a weight of γ2.

$$I(t) = I(t-1) + \beta_3(t)\left(\gamma_1\sum_{t_1\in A}\Delta I(t_1) + \gamma_2\sum_{t_2\in B}\Delta I(t_2)\right) \quad (3)$$

In a set, |·| represents the number of elements, and γ1|A|+γ2|B|=1. The infection rate is determined using the formula (3). Method for Data Preprocessing (C) According to the Suggested Model As a result of a lack of training and resources and an incomplete knowledge of the new coronavirus's symptoms, diagnostic criteria for patients were revised nationwide during the first stages of the COVID-19 epidemic. All provinces' epidemic data had a significant amount of noise due to these causes. The diagnostic criteria in Hubei Province were updated to include clinical diagnosis with the introduction of the fifth edition of the treatment and diagnosis plan on February 12, 2020. The new daily confirmed cases of Wuhan surged to 13,436 on that particular day because to this clinical diagnosis. Subsequent modeling is greatly hindered by these out-of-the-ordinary and noisy data values. Two common ways to handle out-of-the-ordinary data points include data cleaning, which involves deleting them, and the interpolation-based technique. But there are a lot of problems with these approaches. Due to the very short time scale of the epidemic data, data cleaning results in significant data loss and lowers the accuracy of the overall trend estimate of the epidemic model. However, the accuracy of short-term parameter estimate is affected and the dynamic

International Journal of application or Innovation in engineering and management

Ijaiem.com/June 2025/ Volume 14/Issue 1/Article No-1/857-874

ISSN: 2319-4847

development laws of irregular dates are lost when using the interpolation-based technique, even if it does not cause data loss. As a result, the majority of the newly confirmed cases each day derived from anomalous data points are actually missed diagnoses during the early stages of the pandemic. Leaving out this patient count will lead the model to overestimate the severity of the epidemic in its early stages, which will impact its ability to represent rules of evolution that follow. This article suggests a "data balance" approach using the epidemic model as a data preprocessing module to mitigate the effects of diagnostic criterion modifications for the out-of-the-ordinary data points close to February 12, 2020. In order to forecast the amount of new confirmed cases on the anomalous dates, an epidemic model is first constructed using data collected before to February 12, 2020. Secondly, the total number of patients that were overlooked in the early stages is the sum of all the data points that deviate from the prediction findings. Third, in order to create "trend balance" in the total data, these patients are equally split into abnormal and normal dates. What follows are the specifics of the implementation. Start the date with ts and finish it with te if it contains anomalous data. In order to construct an ISI epidemic model and forecast the amount of new confirmed cases on out-of-the-ordinary dates, use the formula I(t0)⋯I(ts). 2) Add up the number of newly confirmed cases each day and the number of missed diagnoses (M) to get the total number of

the early stage N

$$M = \sum_{t=t_s}^{t_e} (\Delta I(t) - \Delta \hat{I}(t_s))$$

$$N = \sum_{t=t_0}^{t_s-1} \Delta I(t) + \sum_{t=t_s}^{t_e} \Delta \hat{I}(t).$$

3) Let $\alpha = M/N$. Then, the rebalanced data before t_e be obtained by the following equation:

$$\begin{cases} \Delta I'(t) = (1+\alpha)\Delta I(t), & t = t_0, \ldots, t_s - 1 \\ \Delta I'(t) = (1+\alpha)\Delta \hat{I}(t), & t = t_s, \ldots, t_e. \end{cases}$$

Two major benefits characterize the data balance preprocessing approach. 1) The calculating technique of the infection rate β(t) will not alter the evolution trend of β(t) if the number of new confirmed cases before ts is raised α times, as stated in (1)-(3). After all the data points before te have been expanded, the number of new daily confirmed cases I(t) before and after the anomalous date may keep its development trend; hence, the long-term fitting result of β(t) becomes progressively stable. Ts is set to February 12, 2020, while te is set to February 13, 2020.

IV. PREDICTION OF THE DEVELOPMENT TREND OF THE EPIDEMIC

While the epidemic model does a good job of predicting when infectious illnesses will spread, it ignores important elements like control and preventative efforts. Thus, in order to bring the epidemic model's parameters up to date, new mechanisms must be implemented. A common use of the LSTM network is data prediction, but it is also useful for modeling hidden variables (such as the number of persons who may be infected). When it comes to predicting the number of infected patients, however, investigations have shown that the LSTM network alone is ineffective. In light of the fact that public knowledge of and response to epidemic prevention efforts is highly correlated with actual viral transmission, this article employs natural language processing (NLP) technology to glean semantic features from news articles covering these topics. The LSTM network employs these characteristics thereafter. The conventional epidemic model predicts the number of infections by adjusting the infection rate. In order to increase the accuracy of epidemic prediction, this technique refreshes the infection rate using news information and preserves the long-term trend of infectious disease models. Regarding the current pandemic scenario in China, we compile news reports. Textual data pertaining to control and preventative strategies is retrieved from this data set. A pretrained natural language processing model is used to transform the retrieved profiles and titles into feature vectors. Figure 3 shows the results of our efforts to accurately estimate the number of infections by extracting characteristics from news sources using natural language processing (NLP) and combining them with an LSTM network. This allows us to update the deviation of the infection rate in the ISI model.

International Journal of application or Innovation in engineering and management

Ijaiem.com/June 2025/ Volume 14/Issue 1/Article No-1/857-874

ISSN: 2319-4847

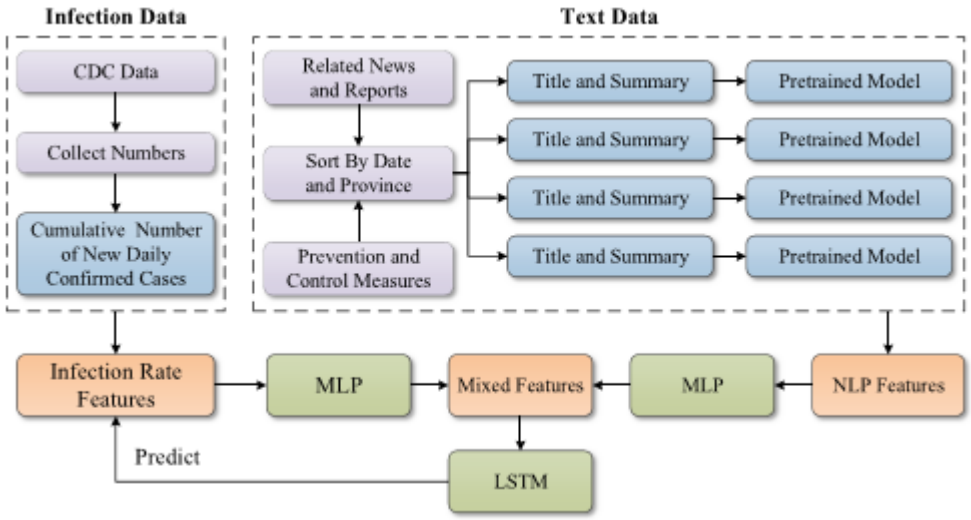

Fig. 3. Prediction model based on the infection rate and NLP features (MLP: multilayer perceptron, NLP: natural language processing, LSTM: long short-term memory network, and CDC: centers for disease control).

A. Extracting News Features We arrange the COVID-19-related material by date, province, and city, and then filter out case reports and linked overseas news in order to obtain the key elements of the news. In order to provide strong and succinct features, feature extraction is limited to the title and major substance of each news piece in practice. Using a pretrained model of the BERT language model (RoBERTa) [34], developed by researchers at the University of Washington and Facebook AI, text characteristics are extracted for each provided Chinese news content. Combining BERT with WordPiece segmentation, the Chinese Whole Word Masking approach, and other methods, this model can break down whole words into smaller ones. This model is capable of producing respectable feature extraction results with very little training. To avoid overloading and accomplish efficient training, the news headlines and primary content are acquired separately as input. The text is encoded using the final hidden layer of the pretrained model. Next, we combine the 768-dimensional title and text encodings to create a 1536-dimensional natural language processing feature vector, where each vector represents a news item. In order to obtain precise daily forecasts throughout the country and in various cities and regions, the dataset is split into two parts: one that includes news from every area, and another that contains news from every province. To make sure there's news every day, it's sorted by day, and the NLP feature vector is the average of all the news features from that day.
A Long Short-Term Memory (LSTM) Network With the Use of Natural Language Processing and the Infection Rate Despite their ability to fit complicated distributions, deep neural networks have a tendency to overfit when not adequately supervised. Features of infection rates do not change with time since they are dependent on the proportion of each element that grows. Epidemic models that rely on infection rates, on the other hand, are ill-equipped to foretell the effects of policy shifts and emergencies, or to make adjustments to their predictions in the near term. So, let us present the LSTM.

International Journal of application or Innovation in engineering and management

Ijaiem.com/June 2025/ Volume 14/Issue 1/Article No-1/857-874

ISSN: 2319-4847

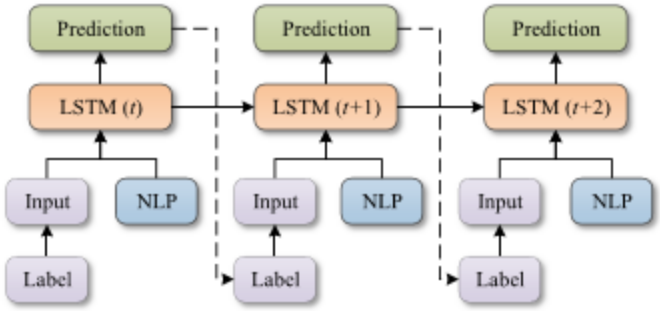

Fig. 4. LSTM network based on NLP features.

Figure 4 depicts a network that uses natural language processing (NLP) elements to predict both the present policy landscape and social media. It will then be possible to guarantee both the short-term flexibility and the long-term stability. We expect the real infection rate to be β(t) in the ISI model, and the regression-affected infection rate to be ˆ β(t) according to the exponential function. To forecast the discrepancy between the regression and actual infection rates, we use the neural network. As the bias feature for prediction, we use the label of day t, denoted as y(t) = β(t)− ˆ β(t). Consequently, the LSTM network may be used in conjunction with the ISI model. We integrate the natural language processing (NLP) characteristics presented in Section IV-A with the bias features to account for the influence of news and policy. For the purpose of encoding hidden states and temporal information, we use LSTM. In order to convert the infection and NLP characteristics into 32-dimensional vectors, we use a one-layer perception model that includes a fully linked layer and a leaky ReLU activation function. This method guarantees that our network is enhanced by identical characteristics. Let W1 and W2 be the weights of the first two perception models, given infection features s1 and NLP features s2. Here is the function g(·) that combines convolution with leaky ReLU:

$$f_1 = g(s_1; W_1)$$
$$f_2 = g(s_2; W_2). \qquad (6)$$

F1 and f2 are the processed features that are combined to form f, a mixed feature. For each time stamp t, let f(t) be the mixed feature, supposing that the hidden state from timestamp t−1 is equal to isht−1. Function In order to convert the hidden state into a prediction, LSTM incorporates a fully linked layer and an LSTM network. The updated hidden state h(t) and the network output x(t) are both defined at time t. After that

$$(x(t), h(t)) = \text{lstm}(f(t), h(t-1); W_l) \qquad (7)$$

where the network's weight is denoted by Wl. During training, we optimize using gradient descent and the Adam optimizer [35]. Subsequently, the loss function is defined as the mean-square error between the prediction and the label.

V. EXPERIMENTAL RESULTS

Here, using data from two sources, we assess how well the suggested model fits the pandemic. First, the health commissioners at the federal and provincial levels are the primary sources of information about cases of infection, cases of suspected infection, cases of cure, and cases of death. Two, natural language processing (NLP) data comes from dxy.com [36], social media, and news media. Prior to categorizing the media by dates and applicable provinces, we filter sickness reports and overseas news. A. Examining the Relationship Between the Infection Rate and the Cumulative Daily Confirmed Cases Current epidemic models assume a periodic infection rate for viruses [37]. Since identified patients are no longer contagious, we may infer that the bulk of the infection on day t originated from the sum of all newly confirmed cases during the last k days, since they are medically separated. In order to delve further into the ever-changing transmission rules of the virus,

we use an epidemic model that is based on a retrospective approach to examine the epidemic data from Beijing, Shanghai, and Hunan. Due to changes in diagnostic criteria and a shortage of medical resources, several individuals were overlooked or misdiagnosed in the early stages of COVID-19. The pandemic statistics became somewhat contaminated due to these circumstances. We investigate the laws of development of the COVID-19 infection rate and choose Shanghai as our research object because of its relatively extensive public health facilities, which helps to limit the influence of noise. At first, we choose k time scales to represent the relationship between the infection rate β1 and the total number of confirmed cases. Then, based on the outcomes of the predictions, we examine the infection rate of the confirmed cases at various time intervals. In order to assess the rules of viral infection and deduce the epidemic's evolutionary tendency, the experimental data are very important. Figure 5 displays the outcomes of the exponential fitting curves of the predicted infection rate in Shanghai vs the various values of k. We also utilize the expected cumulative confirmed cases to find the optimal value of k for calculating the infection rate, so we can evaluate this value objectively. Figure 6 displays the mean absolute error (MAE) curves for Shanghai, which compare the number of actual cumulative confirmed cases with the number of expected cumulative confirmed cases. Using data collected between January 23, 2020, and February 18, 2020, the infection rate for each epidemic model is calculated using exponential fitting. Fig. 5 and the Shanghai curve in Fig. 6 demonstrate that when k is small (k = 1-3), there is no apparent regularity in the distribution of the infection rate β1, and there is a relatively substantial inaccuracy in the forecast of the number of cumulative confirmed cases. According to these results, the effect on the infection rate of newly confirmed cases of dates close to day t is small. The distribution of the infection rate β1 gets more concentrated when k climbs progressively from 4 to 6, and the epidemic model's estimate error falls significantly. This discovery establishes that the trend of the infection rate β1 is becoming closer to reality, and the dates that have a major impact on day t are being included in the model. The distribution of β1 stops changing significantly when k is more than 7, yet the epidemic model's MAE curve starts to rise, indicating that the trend of β1 starts to differ from reality. This discrepancy suggests that the model has been contaminated with noisy data, meaning that the patients at day t −k have been separated and do not infect the S group at day t. And to make sure the regulations up there are applicable elsewhere, we set up two epidemic models using patient data from Beijing and Hunan. Figure 6 also displays the MAE curves for the two areas. The epidemic models' output shows that the two areas' infection rates exhibit an inverted bell curve-like exponential fitting, further confirming that the influence of newly confirmed cases each day on the infection rate changes depending on the date. During the interval from t−10 to t−1, the infection rate of time t is greatly affected by the new confirmed cases in the intermediate phase. Although the experimental data from the aforementioned cities and provinces show that the pandemic is generally trending upwards, Wuhan had a dramatic spike in the number of newly confirmed cases per day after a modification in diagnostic criteria on February 12, 2020. In order to address this issue, we first construct an epidemic model using data collected in Wuhan between January 23, 2020, and February 11, 2020, and we use the exponential function to determine the general trend of the infection rate's progression. Figure 7 shows that Wuhan's infection rate fitting curve is comparable to Shanghai's and Beijing's, indicating that the epidemic pattern in Wuhan is likewise steady. Consequently, it is fair to preprocess the Wuhan anomaly data points using the data balancing approach outlined in Section III-C. B. Examining Various Time Intervals and Their Impact on the Infection Rate of New Confirmed Cases The infection rates of patients in incubation at various time intervals vary [38], [39]. The infection rate of newly confirmed patients on day t may be affected differently by the new daily confirmed cases from day t−k to day t∗1. In this article, we look at the impact and timing

International Journal of application or Innovation in engineering and management

Ijaiem.com/June 2025/ Volume 14/Issue 1/Article No-1/857-874

ISSN: 2319-4847

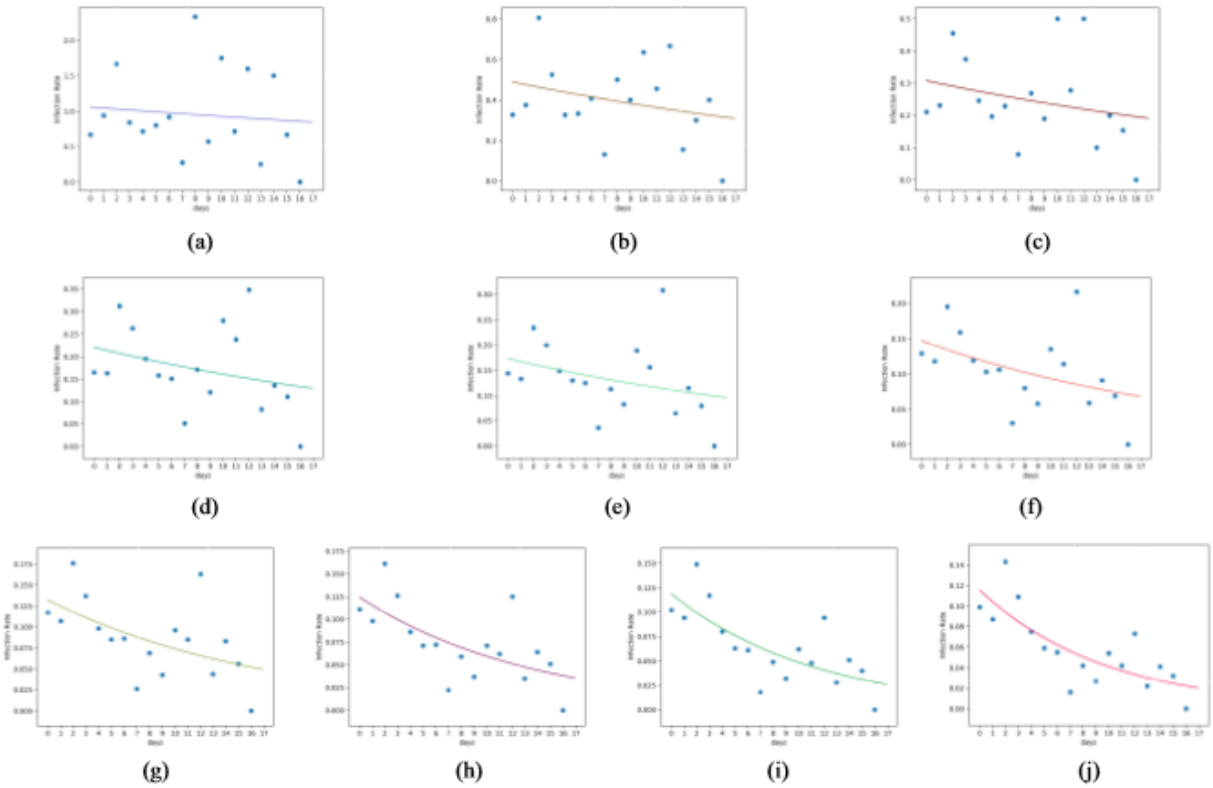

Fig. 5. Fitting curves of infection rate β1 in Shanghai. (a) k = 1. (b) k = 2. (c) k = 3. (d) k = 4. (e) k = 5. (f) k = 6. (g) k = 7. (h) k = 8. (i) k = 9. (j) k=10.

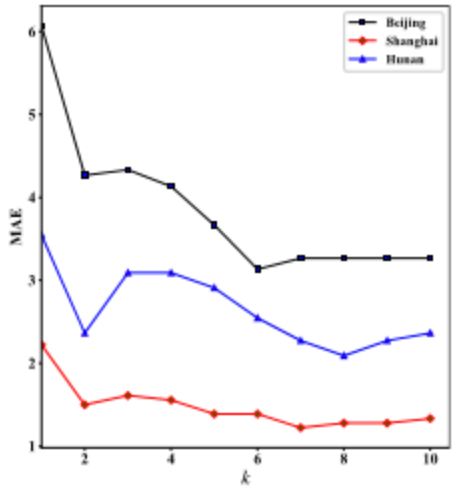

Fig. 6. MAE curves between the number of actual cumulative confirmed cases and the number of predicted cumulative confirmed cases in Shanghai, Beijing, and Hunan.

rules governing the spread of epidemics in various urban areas and regions by means of (2). We start by looking at the correlation between the previous 10 days' worth of confirmed cases and the number of new cases in Beijing,

International Journal of application or Innovation in engineering and management

Ijaiem.com/June 2025/ Volume 14/Issue 1/Article No-1/857-874

ISSN: 2319-4847

Shanghai, Zhejiang, and Hunan on day t. Just like the findings in the previous section, when the distribution of weights is taken into account, the curve of parameter α typically resembles a bell curve. Figure 8(a) shows that the new confirmed cases from days t −8 to t −3 contribute more to the new confirmed cases on day t than the new confirmed cases from days t−10 to t−9 and from days t−2 to t−1. The distribution of αi exhibits a tendency where the value is tiny on either side and big in the center when (2) is applied to the estimated parameter β2(t). On the other hand, the fact that the value of αi is almost zero on day t−10 suggests that the earlier confirmed instances little impact the confirmed cases on day t. For most provinces and cities, the research finds that αi is bigger on days t−8 tot−3, but less on days t−10 to t−9 and days t−2 tot−1. Thus, around 5.5 days is the typical duration of an illness. We use a grouped multiparameter technique to balance the parameters in order to prevent the under- or over-saturation phenomena discussed in Section V-A. In accordance with (3), the weights for the dates ranging from t−8 to t−3 may be defined as γ1, and similarly, for the days t−10 to t−9 and t−2 tot−1, the weights can be defined as γ2. Then, using the formula (3), we can

$$I(t) = I(t-1) + \beta_2(t)\gamma_1 \sum_{i=3}^{8} \Delta I(t-i) + \beta_2(t)\gamma_2 \left(\sum_{i=1}^{2} \Delta I(t-i) + \sum_{i=9}^{10} \Delta I(t-i) \right) \quad (8)$$

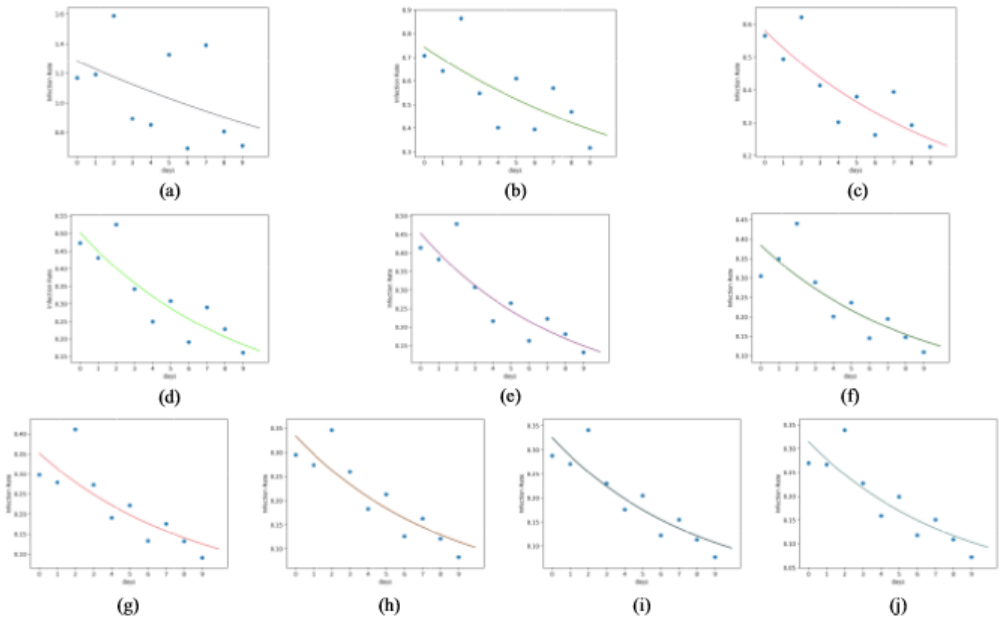

Fig. 7. Fitting curves of infection rate β1 in Wuhan. (a) k = 1. (b) k = 2. (c) k = 3. (d) k = 4. (e) k = 5. (f) k = 6. (g) k = 7. (h) k = 8. (i) k = 9. (j) k=10.

in which 6γ1+4γ2= 1. These findings are shown in Figure 8 as a consequence of this equation. Figure 8 indicates that it is consistently distributed throughout several provinces and cities. It is evident from all the curves in Figure 8 that the values of γ2 are consistently lower than the values of γ1. In the case of Hunan and Zhejiang, γ2 is almost nil. We treat the values of γ2 for other cities as random noise and put γ2 equal to zero. At last, we may rewrite (3) in the following way:

International Journal of application or innovation in engineering and management

Ijaiem.com/June 2025/ Volume 14/Issue 1/Article No-1/857-874

ISSN: 2319-4847

$$I(t) = I(t-1) + \beta_4(t) \sum_{i=3}^{8} \Delta I(t-i). \qquad (9)$$

C. Estimation of the Total COVID-19 Cases We validate our approach throughout the nation, including in Wuhan, Beijing, and Shanghai. From January 23, 2020 to February 18, 2020, the preprocessed infection numbers are used as the training data to forecast the infection counts from February 19, 2020 to February 24, 2020. We evaluate the conventional IS model, the ISI model, the ISI model with the LSTM network, and the ISI model with NLP features and the LSTM network to confirm our model's efficacy and the impact of public awareness and government regulation on epidemic prevention. The LSTM network makes advantage of natural language processing characteristics retrieved from both recent and historical news. Mean absolute percentage error (MAPE), MAE, and daily projection for Wuhan, Beijing,

TABLE I COMPARISON OF ACTUAL CONFIRMED NUMBER AND PREDICTED NUMBER IN WUHAN

	SI	ISI	ISI+LSTM	ISI+NLP +LSTM	GT
Feb. 19	45794	45260	45175	**44970**	45027
Feb. 20	47030	45997	46307	**45504**	45346
Feb. 21	48130	46628	46918	**45872**	45660
Feb. 22	49106	47138	47665	**46163**	46201
Feb. 23	49970	47532	48045	**46265**	46607
Feb. 24	50732	47842	48538	**46439**	47071
MAE	2475.00	747.50	1122.67	**239.83**	0
MAPE	5.35%	1.62%	2.43%	**0.52%**	0

TABLE II COMPARISON OF ACTUAL CONFIRMED NUMBER AND PREDICTED NUMBER IN BEIJING

	SI	ISI	ISI+LSTM	ISI+NLP +LSTM	GT
Feb. 19	396	394	394	**395**	395
Feb. 20	398	395	395	**396**	396
Feb. 21	**400**	396	396	397	399
Feb. 22	402	396	396	**397**	399
Feb. 23	403	**397**	**397**	**397**	399
Feb. 24	404	397	**398**	397	400
MAE	2.50	2.17	2.00	**0.50**	0
MAPE	0.63%	0.54%	0.50%	**0.38%**	0

Shanghai and around the nation. For the sake of clarity, we summarise the prediction findings, and Tables I–IV provide the resulting comparisons.

International Journal of application or innovation in engineering and management

Ijaiem.com/June 2025/ Volume 14/Issue 1/Article No-1/857-874

ISSN: 2319-4847

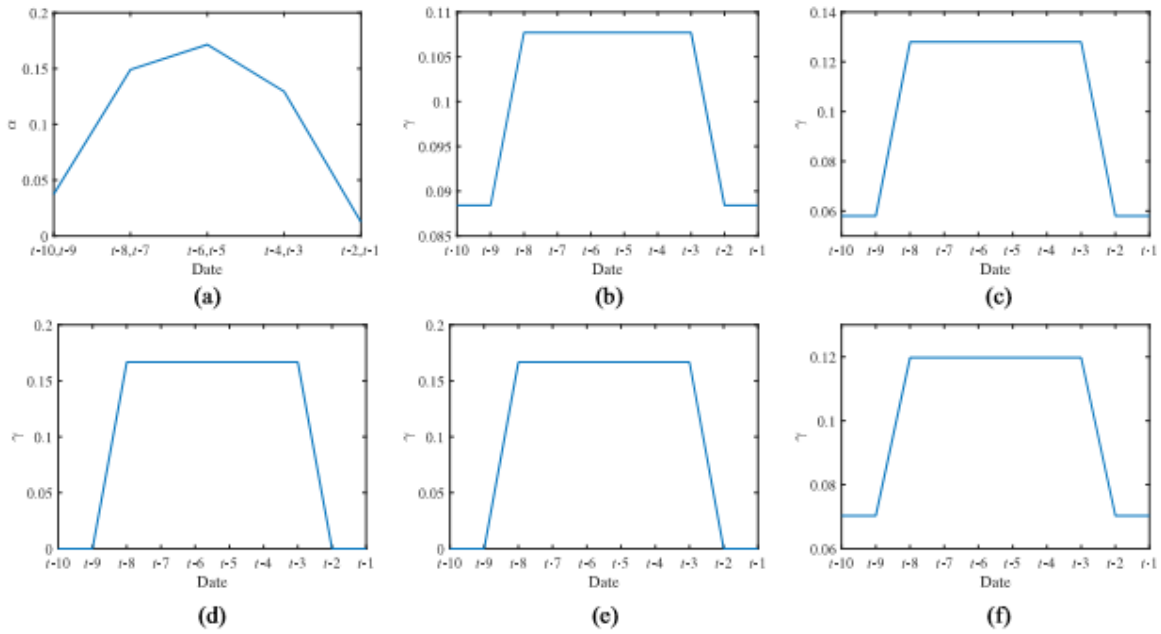

Fig. 8. Infection rate of the new confirmed cases from day t−10 to t−1 to new confirmed cases on day t in the different provinces or cities and the average effect, where "Average" denotes the average contribution of newly confirmed cases from t−10 to t−1 to new confirmed cases on day t in four regions: Beijing, Shanghai, Zhejiang, and Hunan. (a) Average. (b) Beijing. (c) Shanghai. (d) Zhejiang. (e) Hunan. (f) Wuhan.

TABLE III COMPARISON OF ACTUAL CONFIRMED NUMBER AND PREDICTED NUMBER IN SHANGHAI

	SI	ISI	ISI+LSTM	ISI+NLP +LSTM	GT
Feb. 19	336	**334**	**334**	**334**	333
Feb. 20	338	335	335	**334**	334
Feb. 21	340	335	335	**334**	334
Feb. 22	341	336	336	**335**	335
Feb. 23	343	336	337	**335**	335
Feb. 24	344	336	337	**335**	335
MAE	6.00	1.00	1.33	**0.17**	0
MAPE	1.79%	0.30%	0.40%	**0.05%**	0

TABLE IV COMPARISON OF ACTUAL CONFIRMED NUMBER AND PREDICTED NUMBER AT THE COUNTRYWIDE SCALE

International Journal of application or innovation in engineering and management

Ijaiem.com/June 2025/ Volume 14/Issue 1/Article No-1/857-874

ISSN: 2319-4847

	SI	ISI	ISI+LSTM	ISI+NLP +LSTM	GT
Feb. 19	75902	75368	75536	**75270**	74576
Feb. 20	77431	76396	76668	**76116**	75465
Feb. 21	78787	77210	77621	**76807**	76288
Feb. 22	79988	77817	78393	**77432**	76936
Feb. 23	81049	78290	79009	**77970**	77150
Feb. 24	81984	78667	79510	**78432**	77658
MAE	2844.67	945.83	1444.00	**659.00**	0
MAPE	3.71%	1.24%	1.89%	**0.86%**	0

For the three example cities shown in Figure 9, our model produces respectable predictions. The conventional SI model is much outdone by our ISI model. The LSTM network is unstable since it does not consistently improve compared to the ISI model. Among the models tested, the ISI+NLP+LSTM model produced the most accurate forecast. This discovery demonstrates that natural language processing characteristics provide further data and direction for illness prognosis.

To summarize, this article proposes a hybrid AI model for COVID-19 prediction based on the ISI model. The model incorporates an NLP module, which brings crucial information about the major public support and government efforts into the prediction calculation process. As a result, the predicted outcomes are more in line with the actual trend of the epidemic's development.

Section D: The Base Reproduction Number R0 An epidemiologic measure that is often used to characterize the transmissibility of an infected patient is the basic reproduction number R0. Here, R0(t) is defined as the mean number of secondary cases that one confirmed case at time t would generate in an all susceptible population. The following formulation is based on (9):

$$I(t+j) = I(t+j-1) + \beta_4(t+j)\sum_{i=3}^{8}\Delta I(t+j-i). \quad (10)$$

The secondary cases infected by the new daily confirmed cases at time t consist of β4(t+3)I(t), β4(t+4)I(t),..., β4(t+8)I(t), as shown in the equation above. So, the fundamental reproduction number at time t is

$$R_0(t) = \frac{\sum_{i=3}^{8}[\beta_4(t+i)\Delta I(t)]}{\Delta I(t)} = \sum_{i=3}^{8}\beta_4(t+i). \quad (11)$$

Figure 10 shows the results of our analysis of the fundamental reproduction number R0's evolutionary patterns in Beijing, Shanghai, Zhejiang, Hunan, and Wuhan. The values of

International Journal of application or Innovation in engineering and management

Ijaiem.com/June 2025/ Volume 14/Issue 1/Article No-1/857-874

ISSN: 2319-4847

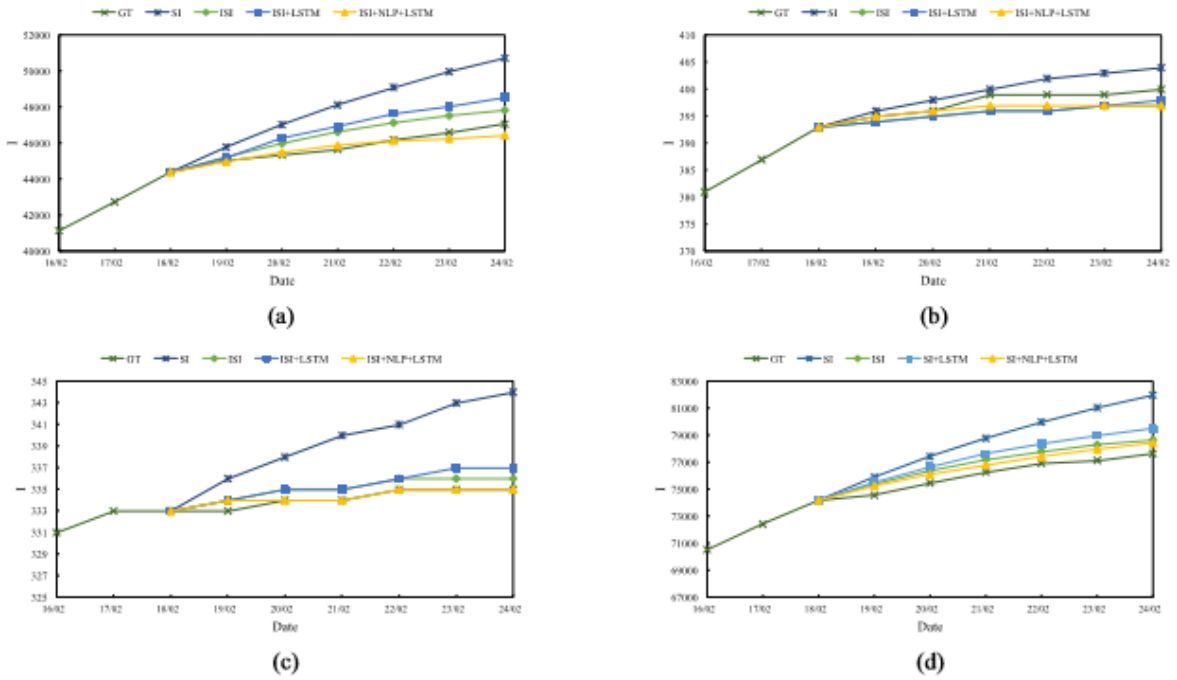

Fig. 9. Comparison of actual confirmed number and predicted number in three typical cities and at the countrywide scale. (a) Wuhan. (b) Beijing. (c) Shanghai. (d) Countrywide

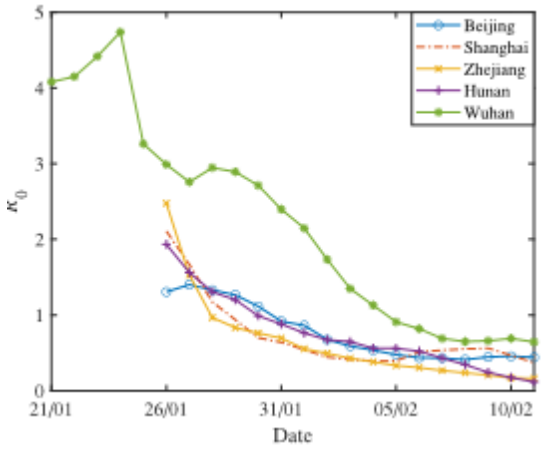

Fig. 10. Curves of the basic reproduction number R0 for different provinces and cities in China.

When control and preventative measures are put into place, R0 for all areas decreases over time. At a pivotal juncture in the 2020 COVID-19 pandemic, the Wuhan region was quarantined on January 23. We look at additional R0 numbers for Wuhan to see how the city lockdown affected R0. Locking down the city was crucial in containing the COVID-19 outbreak, as seen in Fig. 10, where the R0 curve in Wuhan peaked on January 24, 2020, and then decreased quickly. We also forecast the total number of confirmed cases in Wuhan and China using the suggested hybrid AI model; the data used for this purpose was gathered between January 23, 2020, and February 18, 2020. Based on the cumulative confirmed cases prediction curves provided in Figure 11, the total number of cases for

International Journal of application or innovation in engineering and management

Ijaiem.com/June 2025/ Volume 14/Issue 1/Article No-1/857-874

ISSN: 2319-4847

Wuhan up to the end of March would be 482,247. Nevertheless, the figure would rise to 102769 if Wuhan were to be quarantined on January 27, 2020, four days after the actual time.

VI. CONCLUSION

New daily confirmed cases at various time intervals make varying contributions to susceptible infections, according to this paper that seeks to anticipate the trajectory of the COVID-19. An analysis is conducted to determine the effect of confirmed cases in the days leading up to time t on the newly confirmed cases at time t. We use this information to suggest a grouped multiparameter approach that categorizes confirmed cases' infection rates according to time. We continue by deriving the multi-parameter ISI model that was suggested. Using natural language processing (NLP) technology, this article extracts relevant news items, such as steps to control epidemics and residents' knowledge of the need to avoid epidemics, and encodes them into semantic characteristics. In order to update the infection rate provided by the ISI model, these characteristics are then supplied into the LSTM network. To sum up, this article proposes a hybrid AI model for COVID-19 prediction based on the ISI model. The model incorporates an NLP module, which has brought important information facilitated by the joint efforts of federal and state governments, as well as the public's massive support in the prediction calculation process. Consistent with real epidemic cases, the model's prediction results demonstrate that the suggested hybrid model outperforms earlier models in analyzing the virus's transmission law and development trend, and that related news language information processing can

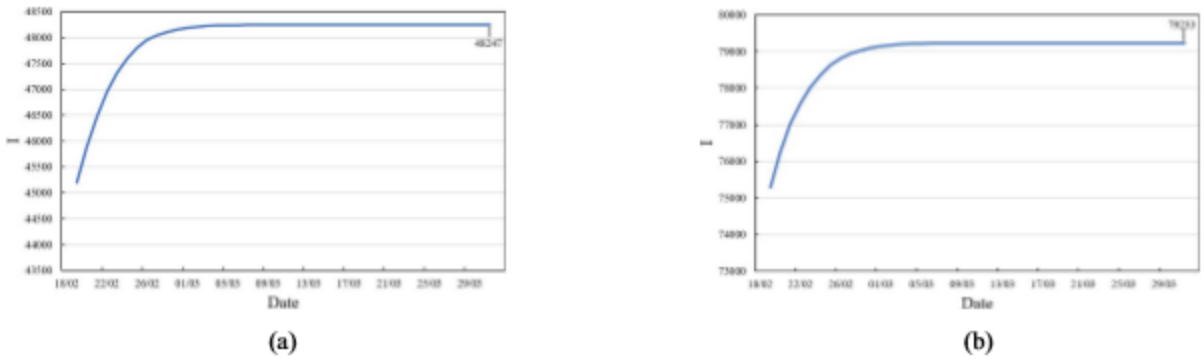

Fig. 11. Prediction curves of the cumulative confirmed cases in (a) Wuhan and at the (b) countrywide scale.

contribute to making the prediction model more accurate. Furthermore, we provide a reliable approach to forecasting future public health events' transmission laws and development trends. In order to set up a state-of-the-art system for preventing epidemics, this paper demonstrates that data release efficiency, openness, and transparency are crucial.

REFERENCES

[1] S. Ying et al., "Spread and control of COVID-19 in China and their associations with population movement, public health emergency measures, and medical resources," p. 24, Feb. 2020. [Online]. Available: https://doi.org/10.1101/2020.02.24.20027623

[2] Y. Bai et al., "Presumed asymptomatic carrier transmission of COVID-19," JAMA, vol. 323, no. 14, pp. 1406–1407, 2020.

[3] W. O. Kermack and A. G. McKendrick, "A contribution to the mathematical theory of epidemics," Proc. Royal Soc. London Ser. A,

Contain. Papers Math. Phys. Character, vol. 115, no. 772, pp. 700–721, 1927.

[4] M. Y. Li, J. R. Graef, L. Wang, and J. Karsai, "Global dynamics of a SEIR model with varying total population size," Math. Biosci., vol. 160, no. 2, pp. 191–213, 1999.

[5] Z. Yang et al., "Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions," J. Thorac. Dis., vol. 12, no. 23, pp. 165–174, 2020.

[6] T. Berge, J.-S. Lubuma, G. Moremedi, N. Morris, and R. Kondera-Shava, "A simple mathematical model for Ebola in Africa," J. Biol. Dyn., vol. 11, no. 1, pp. 42–74, 2017.

[7] C. Rizkalla, F. Blanco-Silva, and S. Gruver, "Modeling the impact of Ebola and bushmeat hunting on Western Lowland Gorillas," EcoHealth, vol. 4, no. 2, pp. 151–155, 2007.

[8] T. W. Ng, G. Turinici, and A. Danchin, "A double epidemic model for the SARS propagation," BMC Infect. Dis., vol. 3, no. 1, p. 19, 2003.

[9] M. Small, P. Shi, and C. K. Tse, "Plausible models for propagation of the SARS virus," IEICE Trans. Fund. Elect. Commu. Comput. Sci., vol. 87, no. 9, pp. 2379–2386, 2004.

[10] O. Zakary, M. Rachik, and I. Elmouki, "On the impact of awareness programs in HIV/AIDS prevention: An SIR model with optimal control," Int. J. Comput. Appl., vol. 133, no. 9, pp. 1–6, 2016.

[11] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997.

[12] T. Mikolov, M. Karafiát, L. Burget, J. ˇCernock`y, and S. Khudanpur, "Recurrent neural network based language model," in Proc. Interspeech, 2010, pp. 1045–1048.

[13] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space Odyssey," IEEE Trans. Neural Netw. Learn. Syst., vol. 28, no. 10, pp. 2222–2232, Oct. 2017.

[14] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," Neural Comput., vol. 12, no. 10, pp. 2451–2471, 2000.

[15] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," 2014. [Online]. Available: arXiv:1409.1259.