scientific reports

Check for updates

OPEN

Analysis and prediction of infectious diseases based on spatial visualization and machine learning

Yunyun Cheng[1], Yanping Bai[2]✉, Jing Yang[3], Xiuhui Tan[2], Ting Xu[2] & Rong Cheng[2]

Infectious diseases are a global public health problem that poses a threat to human society. Since the 1970s, constantly mutated new infectious viruses have been quietly attacking humanity, and at least one new type of infectious disease is discovered every year. Therefore, early warning of infectious diseases will greatly reduce the socio-economic harm of infectious diseases. This study is based on the data of COVID-19 epidemic in China (except Macau and Taiwan Province) from 2020 to 2022. Firstly, we used ArcGIS software to analyze the spatial agglomeration pattern of the number of patients in various regions of China through global spatial autocorrelation analysis, local spatial autocorrelation analysis, center of gravity trajectory migration algorithm and other statistical tools; In addition, the areas with serious COVID-19 epidemic and requiring special attention were screened out. Then, autoregressive integrated moving average model (ARIMA), extreme learning machine (ELM), support vector regression (SVR), wavelet neural network (Wavelet), recurrent neural network (RNN) and long short-term memory (LSTM) were used to predict COVID-19 epidemic data in Guangdong Province, China; And the prediction performance of each model was compared through prediction accuracy indicators. Finally, a multi algorithm fusion learning model based on stacking technology is proposed to address the problem of poor generalization ability of single algorithm models in prediction; Furthermore, radial basis function network (RBF) was used as a two-level meta learner to fuse the above models, and particle swarm optimization (PSO) was used to optimize RBF parameters to reduce generalization error. The experimental results show that the performance of the integrated model is better than that of the single model in the COVID-19 dataset. In order to better apply the stacking model to the prediction of new infectious diseases, we applied the prediction model based on the COVID-19 dataset to the prediction of the number of AIDS and pulmonary tuberculosis (PTB) cases, and verified the wide applicability of our model in the prediction of infectious diseases.

Keywords COVID-19, Spatiotemporal analysis, ArcGIS, Machine learning, Fusion learning model, Epidemic prediction

Infectious diseases have long plagued humanity. Due to its widespread "contagiousness" and "epidemic", infectious diseases have brought far more death and trauma to all humanity from beginning to present than wars and conflicts. For example, in recent years, COVID-19 has become a public health problem concerned by the government, the people and researchers, and its development trend still has high variability and uncertainty. In early December 2019, there was an explosive epidemic caused by SARS-CoV-2 infection all over the world. When people are infected with SARS-CoV-2 virus, it will combine with ACE of human alveolar cells to produce a large number of inflammatory cells, which will lead to imbalance of human immune system and acute lung injury[1]. It has a strong human-to-human transmission ability, and its modes of transmission mainly include droplet transmission and contact transmission[2]. Symptoms are mainly manifested as fever, cough, myalgia, fatigue, dyspnea, occasional cough, headache, hemoptysis and other symptoms[3]. At the beginning of the outbreak of COVID-19, the spread of the epidemic has had a huge impact on the economy, people's lives and property of all countries. And it brought major changes to the world as we know it. For example, the Olympic Games and the European Cup were delayed, human life was limited, and lives was greatly threatened. At the same time, it led to a more severe global economic recession than during the 2008 global financial crisis[4]. In order to fight against COVID-19, a large number of researches related to COVID-19 have been carried out around the world,

[1]School of Information and Communication Engineering, North University of China, Taiyuan 030051, China. [2]School of Mathematics, North University of China, Taiyuan 030051, China. [3]Department of Science, Taiyuan Institute of Technology, Taiyuan 030008, China. ✉email: baiyp666@163.com

involving pathogen traceability, viral transmission routes, vaccine research and so on. However, so far, no clear conclusion has been reached on the issue of pathogen traceability.

At present, the virulence of the latest mutant strain is obviously weakened, which is very close to seasonal influenza. COVID-19 has become a disease that can be prevented, controlled and treated. People have come to the conclusion that coexistence between humans and viruses is inevitable for a long time in the future[5]. Therefore, after controlling the spread of the virus to a certain extent, the government and experts believe that restrictions can be gradually relaxed to ensure the normal operation of society and economy. At present, people are gradually recovering their lifestyles and returning to public places for study, leisure and other activities. Although the situation has improved since the release of epidemic control measures, virus research still faces many challenges and unresolved issues. Governments around the world are still searching for ways to control this disease to mitigate its catastrophic consequences for people's health and the economy[6].

In addition, the spread of infectious diseases is considered one of the inevitable public health problems in the world. For example, common seasonal infectious diseases, such as tuberculosis[7] and malaria[8], will also bring heavy burdens to public health and social economy[9]. In this situation, it is very important to strengthen the monitoring of infectious diseases to reduce their impact on our society[10]. The research on infectious disease analysis and prediction is beneficial for improving the initiative and predictability of infectious disease prevention and control[11], thereby enabling relevant health departments to allocate medical resources rationally. This is also crucial for the formulation of epidemic prevention policies and the decision to resume work and production. However, the spread of infectious diseases involves both biological and social factors. These factors interact and influence each other, forming a complex system[12]. The error of the infectious disease analysis and prediction model based on traditional methods is relatively large, and the effect is not ideal in practical applications. Therefore, this study will propose an analysis and prediction model of COVID-19 based on spatio-temporal data visualization and stacking structure.

According to the characteristics of epidemic prevention and control, we first use big data ArcGIS software to analyze the spatial distribution and development trend of the epidemic in real time. Then, aiming at the problem that the generalization ability of a single prediction model is not strong, a stacking strategy is adopted to integrate multiple prediction algorithms. At the same time, we use three times day forward-chaining to adjust and optimize the super parameters of base learners, and use Pearson correlation analysis to select the best combination of base learners to improve the prediction performance of the stacking algorithm. Our work contributions are as follows:

(1) We collected the data of COVID-19 epidemic in Chinese mainland (except Macau and Taiwan Province) from 2020 to 2022, and made a data visualization analysis of the epidemic data. Then, we used ArcGIS software to analyze the spatial agglomeration pattern of the number of patients in various regions of China through global spatial autocorrelation analysis, local spatial autocorrelation analysis, center of gravity trajectory migration algorithm and other geostatistics tools. The above experiments measured the correlation between the spatial distribution of the number of COVID-19 epidemic cases, reflected the spatial distribution characteristics of the number of patients in China, and screened out Guangdong Province, where the epidemic was serious and needed further analysis.
(2) Autoregressive integrated moving average model (ARIMA), extreme learning machine (ELM), support vector regression (SVR), wavelet neural network (Wavelet), recurrent neural network (RNN) and long short-term memory (LSTM) were used to predict COVID-19 epidemic data in Guangdong Province, China, and we compared the predictive performance of each base learners through three evaluation criteria, including mean absolute error (MAE), root mean square error (RMSE), and mean absolute percentage error (MAPE).
(3) Based on stacking technology, we selected algorithms with good prediction performance and low correlation as the primary base learner of the integrated algorithm through Pearson correlation analysis method, and used the prediction results of the base learner as a new dataset. Then, radial basis function network (RBF) was used as the meta learner of stacking model to correct the deviation of various base learners from the training set, and particle swarm optimization (PSO) algorithm was used to optimize the three parameters of RBF. Finally, by comparing different integrated models with single model, the prediction performance of integrated models were better than that of the single model.
(4) In order to better apply the model proposed in this paper to the study of infectious diseases, we applied the stacking model with better prediction effect on the COVID-19 dataset to build the prediction model of AIDS and pulmonary tuberculosis (PTB), and compared the prediction performance of the model. The results indicated that the migrated model still has good generalization ability.

The rest of this article is structured as follows. In Sect. 2, we will review the literature findings related to spatiotemporal analysis and trend prediction of COVID-19. In Sect. 3, we introduce the spatiotemporal analysis and prediction models proposed in this study. In Sect. 4, we will conduct experiments on several infectious disease datasets, and analyze and compare the performance of the model we proposed. In Sect. 5, we will discuss our proposed model.

Literature review

Spatiotemporal analysis

Tobler's first law of geography holds that all things on Earth are interconnected, and the closer they are to each other, the stronger the connection between them[13]. According to relevant epidemiological studies, various epidemiological mapping systems were used to track the spatio-temporal patterns of cholera, influenza, pestis and other infectious diseases many years ago[14]. For many years, in disaster and crisis management environments, complex data has often made it difficult for people to infer relationships, patterns, and correlations in epidemics

through other methods. Geographic information monitoring technology (GIS) has been widely used to solve such problems[15]. Geographic spatial analysis technology can visualize location-based data, which is helpful for epidemiological modeling. This technology can provide real-time and accurate information, and make significant progress in this field[16]. During the current pandemic, various GIS software and methods have been adopted and widely used. For example, measures such as blocking and tracking close contacts can be taken to prevent the spread of the virus[17]. The best example of a GIS application is a web-based near real-time COVID-19 dashboard created by Johns Hopkins University, which provides visual project services for government agencies, enterprises, and academia[18].

The risk of a pandemic is not only related to the characteristics and modes of transmission of pathogens and hosts, but also to transportation, personnel mobility, population density, social factors, vaccine development, and other factors[19]. In areas with sparse populations and weak transportation networks, where the risk of epidemics is generally low. GIS methods such as buffer zones, interpolation, and numerical analysis were used to estimate and visualize the ongoing COVID-19 pandemic and determine the trajectory of the virus transmission path[20]. Gephi diagrams were also used to visualize the interactions between the spread and flow of viruses, achieving a layout from macro to micro levels[21]. ArcGIS is a platform with powerful data processing and spatial analysis function, which is introduced by ESRI company of the United States[22]. This software can visually display the occurrence and spread of COVID-19 in any region, as well as the correlation between population density and COVID-19 variables[23]. There was a significant correlation between population density, incidence rate of asthma and the death cases of COVID-19. At the same time, some researchers found that the incidence rate of chronic diseases was also high in areas with high incidence of COVID-19, and this synchronous health burden was caused by COVID-19[24]. Community environment may also affect the spread of COVID-19. In some studies, GSV images and computer vision were used to detect the characteristics of the building environment, and then Poisson regression model was used to determine the relationship between the characteristics of the building environment and COVID-19 cases[25]. The relationship between urban land and the spread of COVID-19 is an effective tool to prevent the spread of viruses between cities. According to the socio-economic standards of land use, some scholars used machine learning algorithms to draw a risk map of COVID-19 for Tehran, Iran[26]. The development of vaccines is a necessary step to fight against pandemics and protect high-risk groups[27]. GIS analysis was used to determine vaccination rates in different geographical areas to ensure the fair and effective distribution of vaccination centers[28]. The supply of COVID-19 vaccine is limited. Considering four types of factors, such as population, economy, vulnerability and spatial connectivity, researchers used the hierarchical structure of GIS to calculate and determine the spatial priority of vaccines[29]. However, there are still racial differences in vaccination rates[30]. Community participation and the application of GIS were used to reduce the structural obstacles of COVID-19 vaccination and increase the fair opportunity of COVID-19 vaccination[31]. Population density and weather attributes are also crucial to the study of epidemics such as COVID-19. GIS and Spearman correlation analysis were used to explore the correlation between population density, temperature, humidity and COVID-19 cases[32]. In addition, WHO has formulated the rules of action for health authorities and urged governments to take containment and suppression measures to reduce the spread of COVID-19. Researchers used GIS technology and AHP methods to construct an overall government response and strictness index to evaluate the strictness and effectiveness of implemented policies in reducing disease transmission[33].

In addition, geostatistics methods have been widely applied in the study of COVID-19, providing powerful tools for understanding the spatial distribution, transmission dynamics, and risk assessment of the epidemic. Considering the infection rates in various cities, a spatial model of COVID-19 infection risk based on geostatistical framework was proposed and implemented in mainland Portugal[34]. A geostatistical tool EpiGeostats R package has been developed to simulate the spatial distribution of COVID-19 risk. It integrates geostatistical models and visualization tools, providing a single map that summarizes disease risk and spatial uncertainty[35]. Bayesian spatio-temporal models and area-to-point (ATP) and area-to-area (ATA) Poisson kriging models were used to track the spread of the virus, taking into account spatio-temporal effects and the impact of government interventions on infection risk[36].

Epidemic prediction

However, visual analysis is not sufficient in epidemic prevention and control. It is necessary to use mathematical models to predict the trend of outbreaks and establish a global epidemic early warning mechanism[37]. Currently, dynamic models based on mechanism analysis and machine learning models based on time series data have been widely applied in research on epidemic trend prediction at home and abroad. Considering the latent period, immunity and fatality rate of COVID-19, SEIR is the main framework for the kinetic model study of COVID-19. In SEIR model, the flow of people is divided into four states according to the state of individuals: S (susceptible), E (exposed), I (infectious) and R (recovered)[38]. In order to capture the special dynamics of COVID-19, the fractional modified SEIRF model, the stochastic modified SEIRF model and the fractional stochastic modified SEIF model were used to compare the actual infection and predicted infection in Egypt[39]. Considering travel restrictions and social distance, some scholars proposed a new SEIR model to reclassify individuals who are exposed and infected[40].

However, the SEIR model and its improved versions have some limiting factors, such as fixing the rate, ignoring all factors related to virus causes, using ordinary differential equations, and relying on assumptions such as virus propagation characteristics[41]. Therefore, some scholars studied the development trend of COVID-19 epidemic at home and abroad from the perspective of statistical modeling, such as linear model, machine learning and exponential model. In order to predict the future trend and development trend of COVID-19 epidemic in Bangladesh, ARIMA prediction model was used to fit the epidemic data in March[42]. In addition, the popular LSTM model was also used to predict the COVID-19 in Morocco in the next two months[43]. In relevant literature, statistical methods have also been widely used in forecasting the trend of COVID-19. Exponential growth,

logical growth, Gomperts growth and other models were used to predict the spread of COVID-19 after the announcement of different unlocking stages in India[44]. However, it should be noted that the process of epidemic spread and development is very complex and dynamic, and no single prediction model can accurately predict the epidemic trend and scale of COVID-19[45]. Some scholars combined ARIMA, neural network models, and wavelet technology to reduce the prediction error of the model[46]. A new nonlinear autoregressive neural network time series model (NAR-NNTS) was also used to predict COVID-19 cases[47]. Considering the cases in the 10 most affected countries in the world, researchers used low-pass Gaussian filter to obtain the trend of COVID-19 data before using ARIMA prediction model[48]. Fuzzy inference system (FIS) has advantages over other traditional mathematical methods in achieving fuzzy modeling. The combination of artificial neural networks and fuzzy logic structures were used to predict the infectious effects during the COVID-19 pandemic[49]. Combining the county-level cases and deaths curated by Johns Hopkins University, researchers developed a Seq2Seq prediction model based on an automatic encoder[50]. The Euler iteration method and cubic spline interpolation method were used to predict COVID-19 cases in South Korea, India, South Africa, Germany, and Italy[51]. Recurrent Neural Network (RNNs) is a model based on machine learning, which has been successfully applied to the prediction of time series, and LSTM is the most widely used RNN form[52]. Researchers used RNN, LSTM, BiLSTM, VAE, and GRUs to predict time series of COVID-19 new and recovered cases. The results indicated that the deep learning model has good potential in predicting COVID-19 cases, and the VAE algorithm is superior to other algorithms[53].

Based on the above background, we found that geostatistics methods show strong potential for application in spatial analyses of COVID-19 outbreaks and epidemiological forecasting studies. However, there are still relatively few studies on the spatial clustering patterns and centre of gravity shift trends of epidemic transmission. Therefore, one of the objectives of this study was to investigate the spatial correlation of COVID-19 transmission, describe and analyse the trend of COVID-19 incidence in China and its spatial clustering pattern, in order to fully explore the application value of geostatistics methods in this field. Moreover, machine learning has been widely used in the research of COVID-19 prediction methods. When choosing machine learning models, we prioritize those that can effectively predict COVID-19. ARIMA is suitable for analyzing linear time series data and can capture trends and seasonal variations of COVID-19. ELM has efficient training speed and good generalization ability. Wavelet can effectively process non-stationary signals and extract transient features, which is particularly suitable for capturing the dynamic changes of the epidemic. SVR is insensitive to outliers and can effectively handle complex data patterns. RNN and LSTM can effectively capture the long-term dependencies of sequences and are suitable for analyzing the dynamic changes of COVID-19. In addition, although the above prediction methods are more accurate and efficient than the traditional methods to some extent, they still need to be further improved. Therefore, We choose an algorithm with good prediction performance and low correlation as the main base learner of the integrated algorithm, then use RBF as a two-level meta learner to fuse the above models, and use PSO to optimize the parameters of RBF to reduce generalization error.

Materials and methods

Study area and data

China has a vast territory, undulating terrain and different climates. COVID-19 is widely distributed in China, and its spread varies from place to place. On December 12, 2019, China's first COVID-19 patient was confirmed in Wuhan. As of December 23, 2022, a total of 397,195 confirmed cases and 5241 deaths were officially reported in China. With the support of national epidemic prevention policies, the spread of COVID-19 has been effectively controlled in most parts of China in 2020 and 2021. According to statistics, in 2021, there were 1 confirmed case in Xinjiang and 0 confirmed case in Tibet. There were no more than 20 confirmed cases in Hainan, Qinghai, Anhui, and Guizhou regions. The data used in this study were collected from the National Health Commission of China and other provincial health committees. Since the related data of COVID-19 in Taiwan Province and Macao Special Administrative Region were only available in specific databases in the later stage, we did not collect complete data for the two regions. The actual study area is 32 regions in China, including 22 provinces, 4 municipalities directly under the Central Government, 5 autonomous regions, and 1 special administrative region. We collected daily newly confirmed cases January 20, 2020 to December 25, 2022 in various regions of China. In order to show the incidence of COVID-19 in various regions of China more clearly, we first visualized the number of cases in various regions, thus helping us to observe and analyze the data more deeply in the future. Confirmed cases of COVID-19 in 32 regions of China from 2020 to 2022 are shown in Fig. 1(a). Different fonts and sizes correspond to confirmed cases in different regions. The larger the font, the more confirmed cases in this region. It can be seen from the Fig. 1(a) that Hong Kong has the largest number of confirmed cases, while Ningxia has fewer confirmed cases. The daily changes of newly confirmed cases in China from 2020 to 2022 are shown in Fig. 1(b). During 2020–2022, there were multiple outbreaks. The first outbreak began in January, 2020, and by March, 2020, the epidemic was gradually controlled and the trend became stable. However, in March 2022, the second round of epidemic broke out and lasted about two months. In November 2022, the third round of the epidemic broke out. In this situation, the national government adjusted relevant epidemic prevention policies, moderately relaxed epidemic prevention and control measures, promoted rapid economic recovery, and gradually restored normal production and life for enterprises and citizens.

Evaluation metric

Multiple evaluation criteria can help us comprehensively compare the predictive performance of models. Four evaluation indexes, including MAE, RMSE, MAPE and I_{index} were used to compare the prediction performance of COVID-19 prediction model in this study. Equations (1)-(4) are the calculation formulas of four evaluation indexes respectively.

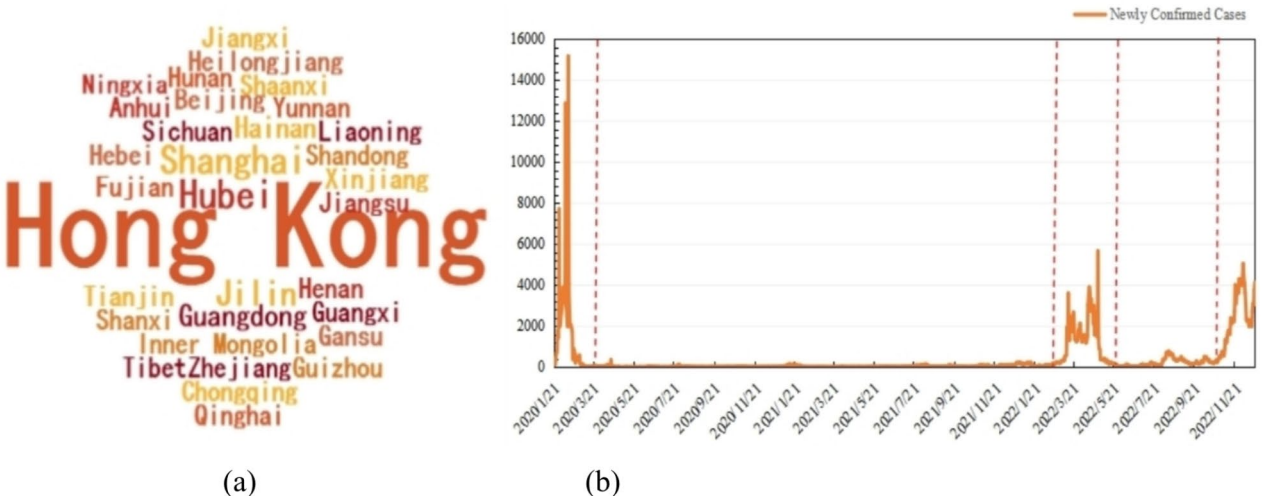

Fig. 1. The status of the epidemic in China in terms of (**a**) confirmed cases in different regions and (**b**) daily changes in the number of new confirmed cases.
Note: Due to the difficulty in obtaining data, Taiwan Province and Macau are not included in this study.

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

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

$$MAPE = \frac{\sum_{i=1}^{n} \left| \frac{\hat{y}_i - y_i}{y_i} \right|}{n} \times 100\% \tag{3}$$

$$I_{index} = \frac{INDEX_{compared} - INDEX_{proposed}}{INDEX_{compared}} \tag{4}$$

Here, n is the number of test sets, y_i is the actual value, and $\hat{y}_i$ is the predicted value. In Eq. (4), $INDEX_{compared}$ and $INDEX_{proposed}$ refer to the two elements to be compared.

Spatial correlation

In spatial data, geographically adjacent location points often have certain spatial dependencies or correlations. Spatial autocorrelation statistics can characterize the potential interdependence or closeness of connections between observed data of variables within the same distribution area[54], reflecting the degree of correlation between random variables in geographical locations. It can help us delve deeper into the spatial distribution pattern of the epidemic, grasp the temporal trend and spatial dynamic evolution of the epidemic. Moran's I[55] is the most common measure of spatial autocorrelation, which reflects the overall distribution of spatial deviation randomness[56]. In this study, global spatial autocorrelation (Global Moran I_g) and local spatial autocorrelation (Local Moran I_t) were used to evaluate the spatial distribution characteristics of COVID-19 cases in different regions of China. Global spatial autocorrelation can determine whether the overall distribution of case numbers exhibits spatial clustering phenomenon, while local spatial autocorrelation can further identify areas with high epidemic incidence in space, with the range of [-1,1].

The expression of the global Moran I_g is as follows:

$$I_g = \frac{N \sum_{i=1}^{N} \sum_{j=1}^{N} \omega_{ij} (x_i - \bar{x}) (x_j - \bar{x})}{\sum_{i=1}^{N} (x_i - \bar{x}) \left(\sum_{i=1}^{N} \sum_{j=1}^{N} \omega_{ij} \right)} \tag{5}$$

The local Moran I_t is expressed as follows:

$$I_t = \frac{\sum_{j=1, j \neq i}^{N} \omega_{ij} (x_i - \bar{x}) (x_j - \bar{x})}{\sum_{i=1}^{N} (x_i - \bar{x})} \tag{6}$$

where N is the number of space observation elements, $N=32$; x_i and x_j are the attribute values of the i-th and j-th elements in the spatial position, respectively. The w_{ij} is the spatial weight matrix. When element i and element j are adjacent, $w_{ij} = 1$. When not adjacent, $w_{ij} = 0$. When Moran's I > 0, there is a positive correlation between the observed values, and the greater the absolute value, the more obvious the spatial correlation; When Moran's I < 0, it means that there is a negative correlation between the observed values, and the greater the absolute value, the greater the spatial correlation.

Center of gravity trajectory transfer

The spatial autocorrelation analysis method can effectively explore the spatial pattern of the epidemic, but it is difficult to intuitively express the spatiotemporal evolution process. In geography, the center of gravity can reflect the spatial and temporal distribution characteristics of a geographical element, and its moving direction and distance can reflect the changing range and spatial difference of a geographical element in a certain period of time. Therefore, the application of the concept of "geographic center of gravity transfer" can further accurately show the temporal and spatial evolution of COVID-19 epidemic in China on the basis of spatial autocorrelation statistics. In this study, the center of gravity migration direction, center of gravity migration distance, stage attribute mean value, and attribute change intensity were used to visually reflect the change status and intensity of the epidemic center in China.

If the study area is composed of n element units, we first calculate the geometric center coordinates of each element, then multiply the center coordinates by the number of confirmed cases in the element unit, and finally divide the accumulated product by the total number of confirmed cases in the study area, so as to find out the coordinates of the epidemic center of the study area at a certain evaluation time. The calculation formula of the center of gravity of the whole area is as follows:

$$\begin{cases} \bar{X} = \sum_{i=1}^{N} w_i x_i / \sum_{i=1}^{N} w_i \\ \bar{Y} = \sum_{i=1}^{N} w_i y_i / \sum_{i=1}^{N} w_i \end{cases} \tag{7}$$

where, $(\bar{X}, \bar{Y})$ is the center of gravity coordinate of newly confirmed cases in the study area at a certain time point; $(x_i, y_i)$is the geometric center of gravity of the i-th planar space element. w_i is its attribute observation value, which is the newly confirmed case at a certain time point in the i-th spatial unit.

$$\theta = \frac{k\pi}{2} + acr\tan\left(\frac{\bar{Y}_t - \bar{Y}_{t-1}}{\bar{X}_t - \bar{X}_{t-1}}\right) \tag{8}$$

$$d = \sqrt{\left(\bar{X}_t - \bar{X}_{t-1}\right)^2 + \left(\bar{Y}_t - \bar{Y}_{t-1}\right)^2} \tag{9}$$

$$M_t = \frac{1}{N}\sum_{i=1}^{N} w_{it} \tag{10}$$

$$\beta_t = \frac{M_t - M_{t-1}}{M_{t-1}} \tag{11}$$

where θ is the center of gravity migration direction, which represents the angular direction of the center of gravity shift at the t-th time point compared to the previous time point. d is the center of gravity migration distance, which calculates the Euclidean distance between the center of gravity at the t-th time point and the previous time point. M_t is the mean of stage attributes, which calculates the mean value of all N planar spatial unit attribute observations at the t time point, and represents the average state of the attribute in the entire study area. β_t is the intensity of attribute change, which represents the intensity of change of the periodic attribute mean value at the t time point relative to the previous time point. The larger the β_t value, the greater the enhancement of the overall state of the attribute compared to the previous time point. The smaller the β_t value, the greater the decrease in the overall state of the attribute compared to the previous time point.

Time-series cross-validation

In addition to many parameters that need to be optimized in the training process, there are a large number of hyperparameter in the machine learning model that need to be manually adjusted before training. Cross-validation is a common method to find a set of hyperparameters that minimize a certain loss function in hyperparametric optimization[57]. However, in the process of building this integrated model, the data set used is time-dependent, and the traditional K-fold cross-validation will cause the problem of information leakage, such as using the later data as the training set and then predicting the previous data[58]. Therefore, we use time-series cross-validation method to improve the integration algorithm[59], so as to accurately simulate "using the current environment to predict the future". As shown in Fig. 2, when we divide the training set and the test set, the timestamp of the test set must be after the training set. All previous data must be allocated to the training set, and the subsequent data should be verified. The step-by-step approach avoids information leakage.

The daily data of new confirmed cases in Guangdong Province from 7 August to 20 December 2022 are selected, containing a total of 136 data points. The dataset is divided into a training set and a test set, in which 80% of the data (the first 111 entries) are used for training and 20% of the data (the remaining 25 entries) are used for testing. A time-series cross-validation method is used, where the training samples are equally divided into four parts, and part of the data is used for training at a time, and the remaining part is used as the validation set to ensure the order of the time series. The 1st folded data (7 August to 2 September 2022) is used as the

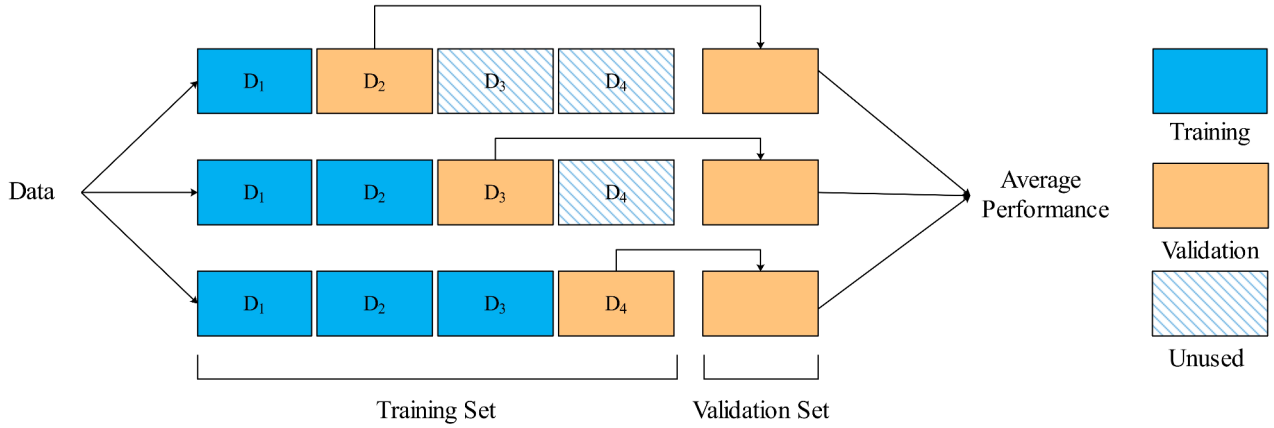

Fig. 2. Time-series cross-validation method overview.

training set and the 2nd folded data (3 September to 1 October 2022) is used as the validation set for the 1st validation. Then the 1st to 2nd folded data (7 August to 1 October 2022) is used as the training set and the 3rd folded data (2 October to 29 October 2022) is used as the validation set for the 2nd validation. And so on, the 4th validation is performed with the data of the 1st to 3rd fold (7 August to 29 October 2022) as the training set and the data of the 4th fold (30 October to 25 November 2022) as the validation set. After completing three validations, we calculate the model evaluation metric MAE for each validation and take the average of these three times to assess the overall performance and predictive ability of the model.

Stacking integrated learning

Stacking is an integration method that connects many different types of base learners through meta-learners. It can integrate the advantages of several weak learners and get a model with strong generalization ability[60]. The selection of learners is a crucial step for the stacking algorithm. Only suitable base learners and meta learners can maximize the effectiveness of learning from each other's strengths and weaknesses[61]. At the same time, the combination method between different learners is also crucial. Therefore, when building a stacking model, we need to analyze the prediction ability of each model and the correlation between each model, and select appropriate base learners, meta learners, and their combination methods to achieve the optimal prediction performance of the stacking model. Figure 3 shows the stacking integrated prediction process.

Base learner

When selecting a base learner for the first layer, two aspects should be considered comprehensively. First, because the effectiveness of the stacking model mainly comes from feature extraction, where different algorithms extract features from different data space angles. Therefore, it is necessary to choose models with significant differences as base learners as much as possible[62]. On the other hand, the base learner with excellent prediction performance can greatly improve the final prediction performance of the fusion model, so we should try to choose the base learner with strong learning ability.

For infectious disease prediction, it is often necessary to analyze and judge the epidemic situation in a short period of time based on the early development trend and level of change of infectious diseases. Traditional time series models require us to master a large amount of statistical knowledge, and modeling a large number of time series also requires tremendous effort. Therefore, in order to ensure the diversity of models, ARIMA, ELM, Wavelet, SVR, RNN, and LSTM were initially selected as the base learners in this study. Among them, ARIMA time series algorithm can better predict time series without constructing features, and can automatically predict large amounts of data through programming. But the shortcomings of ARIMA algorithm are also obvious, and it can't take advantage of the nonlinear features in the data. Compared with traditional neural network algorithms, ELM is faster, simple and easy to implement and use. SVR model has unique advantages in solving small sample, nonlinear and high-dimensional problems, and has been widely used in industrial fields. RNN and LSTM can fully mine the effective information contained in a large number of data, and have the ability of long-term memory and deep learning. Six heterogeneous algorithms have different learning strategies, model forms, and parameter settings, which can leverage different advantages in the dataset. In the stacking algorithm, meta learners use the prediction results of these different base learners to continue learning and optimization, thereby improving overall prediction performance.

Meta learner

When selecting the second layer learner, we choose a simple regression learner with good stability as the meta learner to prevent overfitting. RBF is a feed-forward network, which uses radial basis function as activation function, and can adapt to the processing of nonlinear data and high-dimensional feature space[63]. It has strong adaptability to input data, and the training speed is faster than other neural network algorithms. A RBF neural network consists of input layer, hidden layer and output layer. Figure 4 shows the three-layer structure of RBF

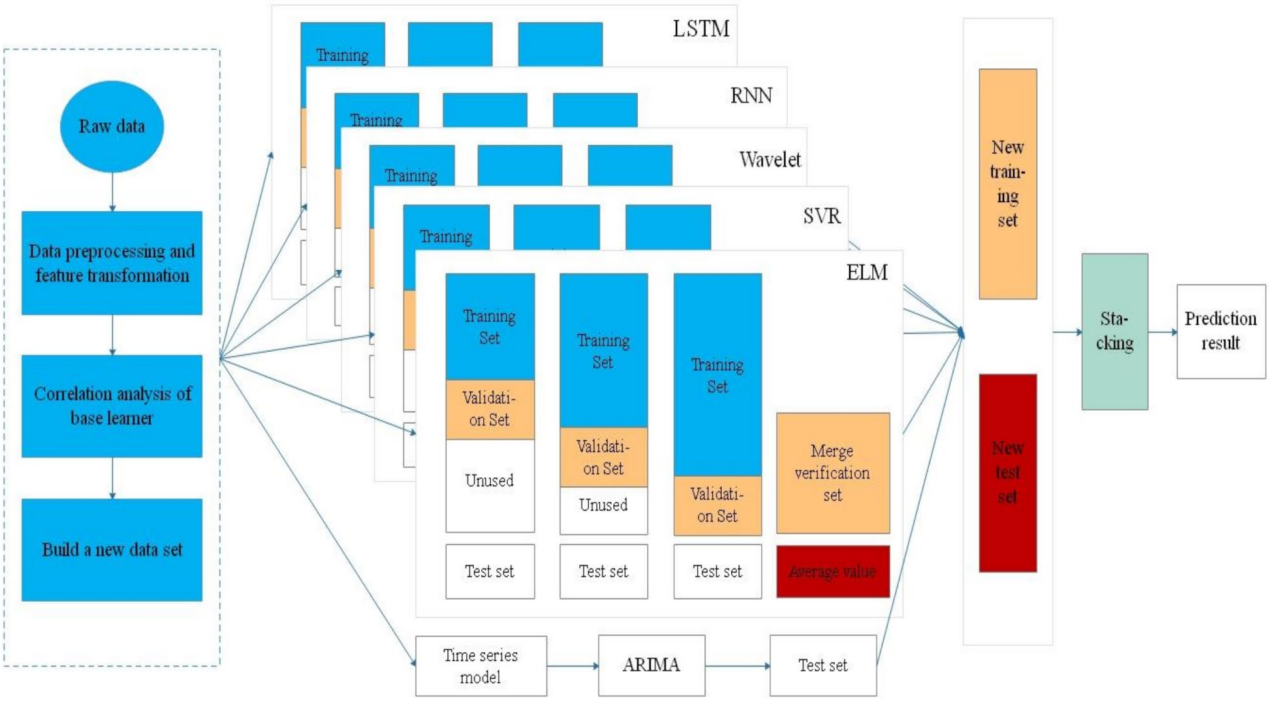

Fig. 3. The stacking integrated prediction process.

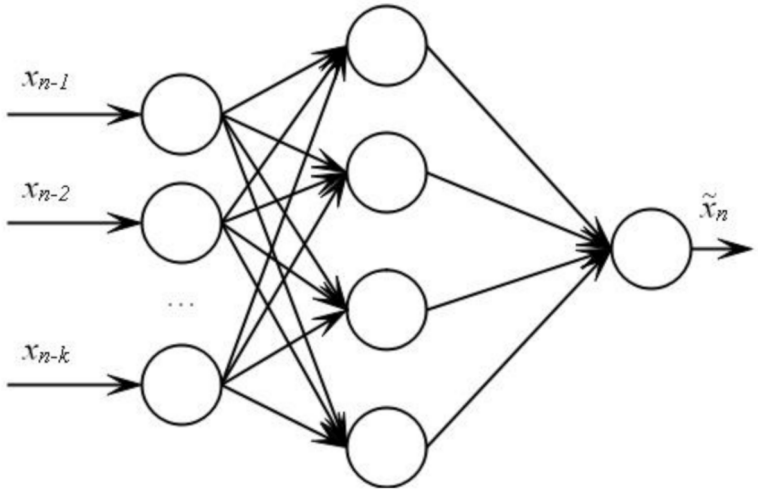

Fig. 4. Structure of RBF neural network.

neural network. The input layer consists of sensing units, the hidden layer converts low-dimensional input data into high-dimensional space, and the output layer responds to input[64].

Due to the special structure and parameter selection of RBF neural networks, the degree of freedom of the network is limited to a certain extent, and it usually requires multiple experiments and adjustments to obtain the optimal parameters[65]. To compensate for the above shortcomings, the PSO algorithm is used to optimize the center c_j of the hidden layer basis function, the width σ_j of the hidden layer nodes, and the weight factor w_j of the output layers connected by the hidden layers in RBF. The PSO algorithm regards these three parameters as free-moving particles, and represents the position of the particles in the form of vectors. The appropriate values of the parameters are determined by PSO algorithm, and finally RBF neural network is established. Table 1 shows the pseudocode of PSO-RBF. The operation process of PSO-RBF neural network algorithm is as follows:

Algorithm: PSO-RBF
Step 1: Initialization the parameters Particle velocity: *V*, Particle position: *X*, best position: *P*, *Vmax*, Minimum error threshold: *MinErr*, and Learning factors: C_1, C_2
Step 2: Particle Swarm Optimization
while Err > MinErr
for i in range(N)
Step 3: Update particle velocity and position
V[i] = *V*[i] + C1 * rand() * (*P*[i] - *X*[i]) + C2 * rand() *(G - *X*[i])
V[i] = clip(*V*[i], -Vmax, Vmax)
X[i] = *X*[i] + *V*[i]
X[i] = clip(*X*[i], Xmin, Xmax)
Step 4: Compute the error
centers = *X*[i][: D]
sigmas = *X*[i][D:2*D]
weights = *X*[i][2*D:]
Err = compute_error(centers, sigmas, weights)
Step 5: Update the particle's personal best position
if Err < best_err[i]:
P[i] = *X*[i]
best_err[i] = Err
Step 6: Update the global best position
if best_err.min() < *G_err*:
G = P[best_err.argmin()]
G_err = best_err.min()
Step 7: Output the final result
print("Optimal solution: ", *G*)
print("error: ", *G_err*)

Table 1. Process of optimizing RBF parameters by PSO.

(1) Combine the center value c_j, width σ_j, and connection weight w_j of the RBF neural network in order to form a particle vector in PSO algorithm, which is written in vector form $X = (c_j, \sigma_j, w_j)$. Initialize the population size, maximum number of iterations, particle position *X*, and velocity *V*.
(2) Continuously update the velocity and position of particles before the error reaches the threshold, and calculate the error *Err*.
(3) Update particle velocity V and position X based on Eq. (12) and Eq. (13):

$$V(t+1) = V(t) + C_1 r_1 (P - X(t)) + C_2 r_2 (G - X(t)) \tag{12}$$

$$X(t+1) = X(t) + V(t) \tag{13}$$

where $X(t)$ is the current position of the particle, P is the personal best position, G is the global best position, and $V(t)$ is the current velocity of the particle.

(4) Update the particle's personal best position P. Take the position with the smallest error value in the particle's own search process as the personal best position, and before the next search step, compare the error value of the current error value with that of the previous personal best position, and select the position of the particle with the smaller error value as the new personal best position.
(5) Update the global best position G. Compare the error values of the ersonal best position of all particles in the search space and select the ersonal best position with the smallest error value as the global best position. Before the next search step, compare the error value of the current global best position with the previous global best position, and select the one with the smallest error value as the new global best position.
(6) Determine whether to end the iteration based on the set end conditions. If it meets the criteria, proceed to the next step. Otherwise, return to step 2 and repeat the iteration.
(7) Record the global extremum G found and end the PSO algorithm.
(8) Construct an RBF neural network using global extremum and train the network.

Result and analysis

This experiment is carried out on a computer with Intel i7-10700 processor and 16 GB of running memory. The operating environment is 64-bit Windows 10 system, and the algorithm implementation tool are MATLAB and ArcGIS.

Time	Field	Moran's I_g	D	Z	*P*
2020	Newly confirmed case	-0.050233	0.0013	-0.5436	0.5867
2021	Newly confirmed case	0.021136	0.0009	1.6667	0.0956
2022	Newly confirmed case	0.033176	0.0010	1.9824	0.0474

Table 2. Global Moran's I_g value of COVID-19 incidence in China from 2020 to 2022.

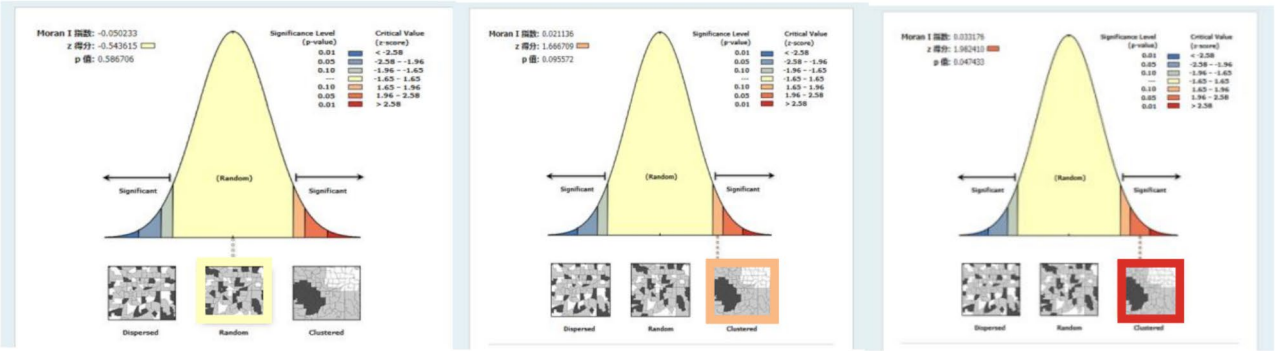

Fig. 5. Spatial distribution pattern of COVID-19 cases in China from 2020 to 2022.

Spatial analysis results

The global Moran's I index of COVID-19 cases in China from 2020 to 2022 was calculated by using the spatial autocorrelation analysis method in ArcGIS. The results are shown in Table 2; Fig. 5. The value of the global Moran's I_g changed from negative to positive, and showed an increasing trend year by year, indicating that the spatial distribution of the epidemic in China has changed from spatial dispersion to spatial agglomeration, and the degree of spatial agglomeration is becoming stronger and stronger. According to data from 2020, there is a negative correlation in the spatial distribution of the epidemic in China, which means that the spread of the epidemic in neighboring areas where the epidemic is severe is not as fast as expected, but relatively slow and stable. This is because at the beginning of the epidemic outbreak, the government's prevention and control measures effectively curbed the "spillover" of the epidemic. The global Moran's I_g for 2021 and 2022 were greater than 0, indicating a certain positive correlation in the spatial distribution of COVID-19 cases in China. The epidemic situation in areas with severe outbreaks and their surrounding areas is also relatively severe. The *p*-value is often used to test whether the Moran's I index is significant. In this study, the *p*-value corresponding to the spatial autocorrelation coefficient in 2022 was less than 0.05, rejecting the original hypothesis. This indicates that COVID-19 data exhibits significant spatial clustering in the study area.

From the results of global autocorrelation analysis, it can be seen that the spatial distribution of epidemic situation in China has shifted from discrete to clustered, and there is a trend of further strengthening. However, the global spatial autocorrelation is a description and analysis of the spatial autocorrelation degree of the epidemic distribution in the entire region, which cannot reflect the differences between the different regions affected by spatial autocorrelation. Based on this, in order to analyze the epidemic situation in various regions of China and deeply explore the spatial interactions and influencing relationships between different regions. We further analyzed the spatial clustering characteristics of the Chinese epidemic using Anselin local Moran I in ArcGIS. Table 3 presents the results of local autocorrelation analysis of COVID-19 incidence in China from 2020 to 2022.

In this study, three time sections of 2020, 2021, and 2022 were selected for local spatial autocorrelation analysis. At a certain level of significance, the newly confirmed cases of COVID-19 in each region was divided into four spatial clustering patterns: high-high, low-low, high-low and low-high. H-H refers to the existence of some high value regions, and the regions around these high value regions are all high value regions. L-L refers to the existence of some low value regions, and the regions around these low value regions are all low value regions. H-L and L-H are just the opposite, indicating that the regions around the high value regions are all low value regions, or the regions around the low value regions are all high value regions.

From the experimental results, it can be seen that only the epidemic in Hong Kong in 2020 belongs to the H-H cluster model. At this stage, the situation of the COVID-19 epidemic situation in this region is relatively severe, the confirmed disease rate is relatively high, and the difference between this region and adjacent regions is small. As an important shipping and trade center in the world, Hong Kong has a high population density and a small per capita housing area. In the early stages of the epidemic outbreak, Hong Kong faced enormous difficulties. The region of L-L cluster pattern experienced two fluctuations, and the regional scope decreased slightly. It moved from Xinjiang, Gansu and Qinghai to Xinjiang and Tibet in turn, and finally to Shanxi, Shaanxi, Ningxia and Sichuan, showing a spatial pattern extending from the north to the northwest and central China. The H-L dispersion area can indicate that the number of confirmed cases in this area is higher than that in the surrounding areas, and there have been a hot spot area of the epidemic. It mainly included Hubei in 2020,

Cluster pattern	2020		2021		2022	
	Province	Ratio	Province	Ratio	Province	Ratio
H-H	Hong Kong	3%	None	0	None	0
H-L	Hubei	3%	Shaanxi	6%	None	0
			Yunnan			
L-H	Guizhou	18%	Hunan	12%	Zhejiang	6%
	Hunan		Jiangxi		Fujian	
	Jiangxi		Anhui			
	Fujian		Zhejiang			
	Anhui					
	Zhejiang					
L-L	Gansu	9%	Xinjiang	6%	Sichuan	12%
	Qinghai		Tibet		Shaanxi	
	Xinjiang				Ningxia	
					Shanxi	

Table 3. Local clustering pattern of COVID-19 incidence in various regions of China from 2020 to 2022.

Type	2020	2021	2022
		Compared with 2020	Compared with 2021
Centre of gravity	30.53°N 112.77°E	29.65°N 114.39°E	25.64°N 115.43°E
Migratory direction	—	Southeast	Southeast
Migration angle (angle with due east)	—	20.1°	65.9°
Center of gravity migration distance/km	—	192.31	469.23
Stage attribute mean	2998	578.66	18988.78
Attribute change intensity	—	-0.81	31.82

Table 4. Calculation results of center of gravity trajectory migration of COVID-19 Incidents in China from 2020 to 2022.

Shaanxi and Yunnan in 2021, and no region belongs to this model by 2022. From 2020 to 2022, the number of L-H regions decreased from 6 to 2, and the scope was greatly reduced, mainly distributed in East China and Central China.

Then, taking the provincial administrative region as the statistical unit, we analyzed the data of COVID-19 confirmed cases in China from 2020 to 2022 with the center of gravity trajectory migration algorithm, and obtained the center of gravity trajectory of China epidemic. The center of gravity, migration direction, migration angle, migration distance, stage attribute mean and attribute change intensity of COVID-19 confirmed cases in each period were calculated by using the formula in Sect. 3.4, as shown in Table 4. In addition, we also mapped the migration path of the center of gravity of COVID-19 epidemic in China, as shown in Fig. 6, which can clearly and intuitively reflect the temporal and spatial evolution characteristics of COVID-19 epidemic in China and the change process of attribute intensity.

From Table 4; Fig. 6, it can be seen that the temporal and spatial evolution of China's COVID-19 in 2020–2022 has the following characteristics. The first stage was from 2020 to 2021, the center of gravity of COVID-19 epidemic slowly moved to the southeast, and it moved from the center of Hubei to the southeast of Hubei. The intensity of its attribute change was −0.81, which was negative. Compared with 2020, the epidemic situation is decreasing. Wuhan, Hubei Province is the first outbreak place in China. At the beginning of the epidemic, due to the lack of knowledge and information, it was difficult to grasp the epidemic situation in time and take necessary prevention and control measures, which led to the spread of the virus. The second stage was from 2021 to 2022, and the focus of the epidemic moved from east to southeast in Hubei to southwest in Jiangxi. Its change intensity reached 31.82, showing an expanding trend of fluctuation, and the epidemic situation intensified. From the overall distribution of the center of gravity, it had undergone a large-scale migration, mainly distributed in the southeast of Hubei and the south of Jiangxi. The track of its movement was becoming faster and faster, and the intensity of attribute changes was changing from negative to positive. This indicated that the epidemic was expanding continuously, and the overall spatiotemporal pattern presented the characteristics of moving towards the southeast.

Data preprocessing

During the process of collecting COVID-19 raw data, there may be data loss and abnormal fluctuations. Not preprocessing can affect the integrity of the dataset and reduce prediction accuracy. In this study, we filled in or replaced missing values and outliers with the average of the corresponding time points before and after the

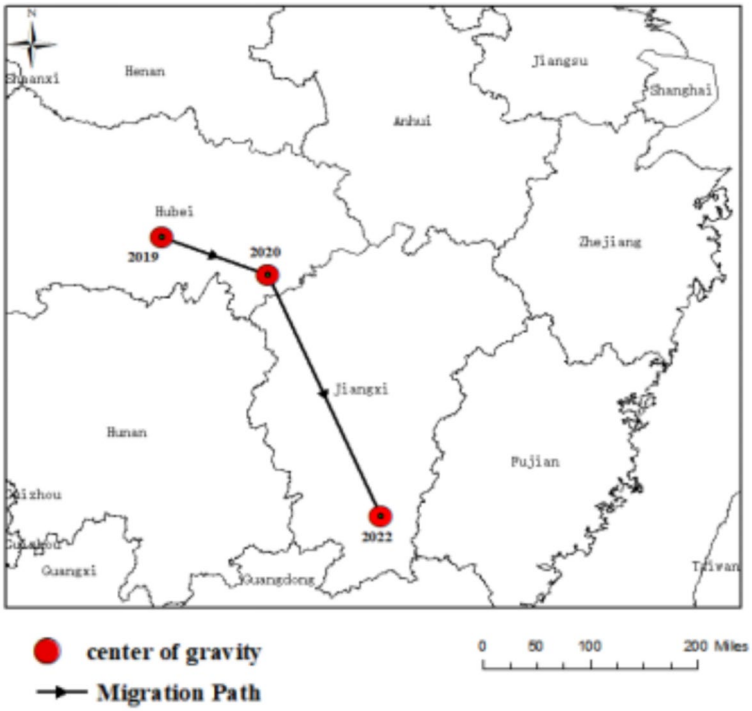

Fig. 6. COVID-19 epidemic center and migration path in China from 2020 to 2022.

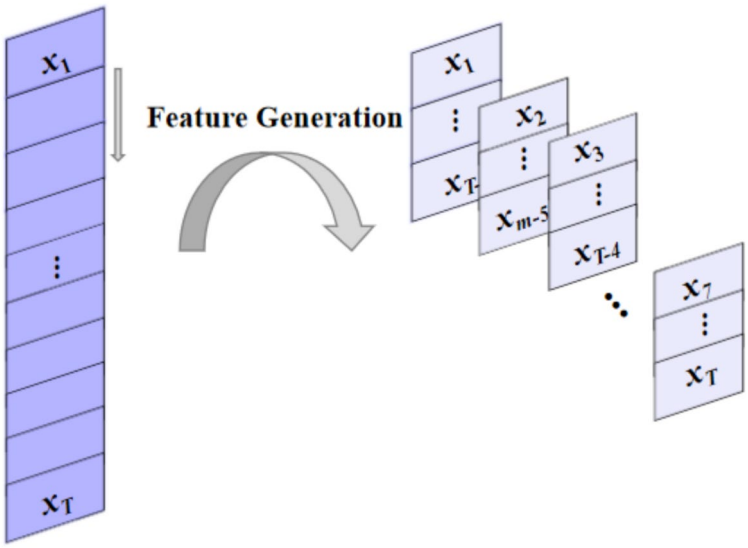

Fig. 7. Feature generation process.

problem point. In addition, normalization was performed on all datasets to limit the data to the range of [-1,1]. The purpose of normalization is to scale the range of values of different features to similar intervals, accelerating the convergence speed of the prediction model. Finally, we used the number of COVID-19 cases in the first seven days to predict the number of cases in the following day. Figure 7 illustrates how to extract and construct features for prediction from historical data.

Single model prediction results

In this section, we selected the daily newly confirmed cases in Guangdong Province from August 7, 2022 to December 20, 2022 as the data set for the prediction model. This study divided 136 sets of COVID-19 data into training and testing sets. The training set accounted for 80% of the total dataset, including 111 sets of data, while the testing set accounted for 20% of the total dataset, including 25 sets of data. In order to prove

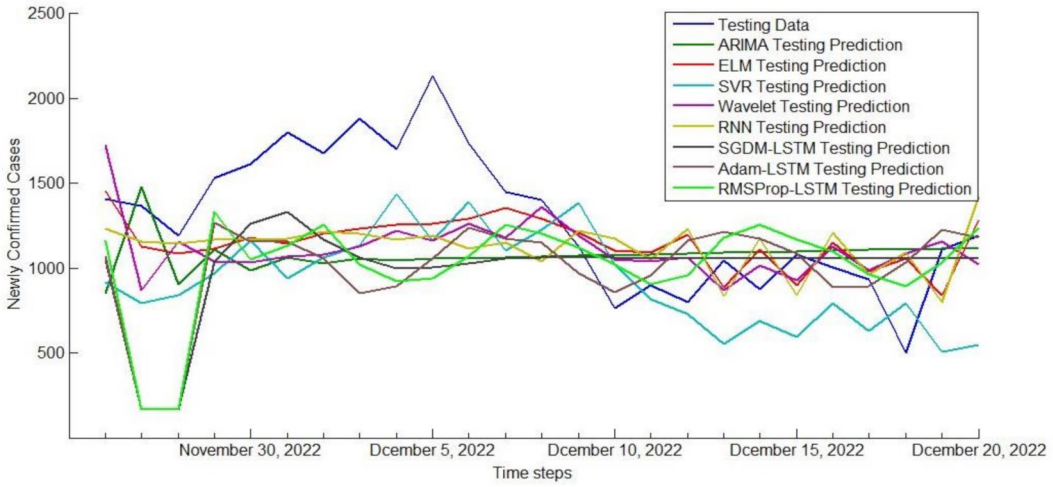

Fig. 8. The newly confirmed cases of COVID-19 in Guangdong Province from November 26th to December 20th, 2022 were compared with those predicted by various models.

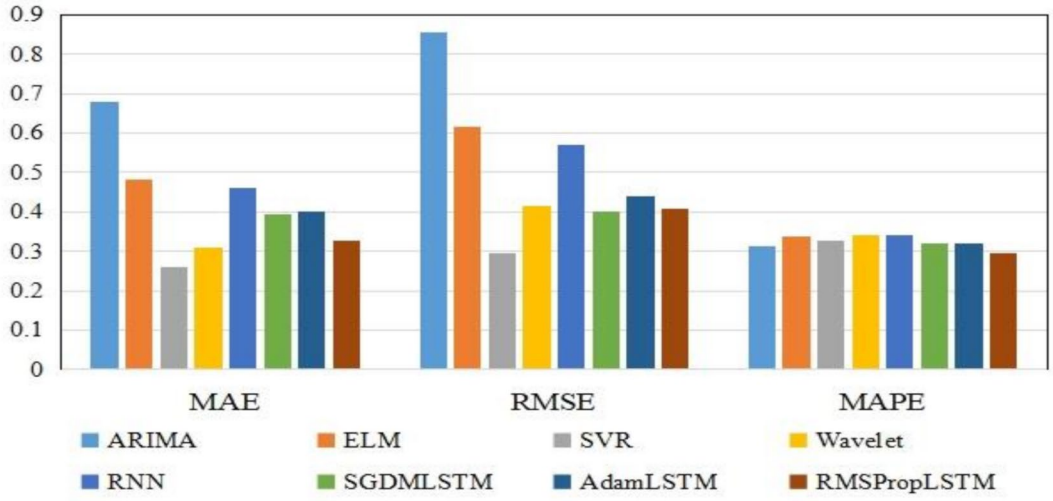

Fig. 9. Predictive performance metrics using the three times day forward-chaining cross-validation.

the rationality of the selection of base learners in the stacking integration model proposed in this study, we first analyzed the prediction performance and differences of each single model. On the COVID-19 dataset, experiments were designed to compare the prediction results of various base learners. All base learners ARIMA, ELM, WNN, SVR, RNN, and LSTM were trained and verified by three times day forward chaining cross-validation. The raw data were divided into the original training set D and the original test set T, then D was divided into four equal training sets D_1, D_2, D_3 and D_4. In the k-th cross-validation, D_{k+1} was the validation set, and $\{D_1, \ldots, D_k\}$ was the training set. For the i-th learning algorithm, the verification set D_{k+1} in the k-th cross-validation can get the corresponding output D^i_{k+1} through this learner. At the same time, after each training, the test set T was put into the training model for prediction, and the average of the three predictions was obtained to form the data set T_i. The above steps were repeated by different base learners. Finally, the prediction results $\left\{D^i_2, D^i_3, D^i_4, i = 1, \ldots, 8\right\}$ of verification set of several base learners were combined to form a new training set D' of the second layer meta-learner of the stacking model, and the prediction results $\{T_1, T_2, \ldots, T_i, i = 1, \ldots, 8\}$ of test set of several base learners were combined to form a new test set T' of the second layer meta-learner of the stacking model. The second layer of meta learners was trained in new training sets and new test sets. We draw the prediction results of each base learner on the test set on the same graph, and used different colors to distinguish different models, as shown in Fig. 8.

In order to make the prediction effect more intuitive, we showed the comparison results of three prediction performance indicators of each model in the form of bar chart in Fig. 9. It can be seen that the SVR model has the lowest MAE and RMSE values, indicating that compared to other models, the predicted results of the SVR

model are closer to the true values and can more accurately predict the propagation and development trend of COVID-19. This is mainly because compared with other models such as neural network, SVR can use fewer free parameters to adjust the model, which also makes it easier to adjust and optimize the model. At the same time, it has also been verified that SVR is indeed an excellent prediction model in the field of machine learning. On the contrary, the values of MAE and RMSE of time series ARIMA model were the highest, reached 0.679 and 0.854 respectively, indicating that its prediction performance is poor. The ARIMA model is based on regular time series for prediction. If the regularity of the time series changes, the predictive performance of ARIMA will be affected. However, there are obvious structural mutations and inflection points in the daily confirmed case dataset of Guangdong Province. For example, the values of the 85th and 102nd groups of data are three times higher than the previous group, which reduces the predictive performance of the ARIMA model.

In addition, from some experimental results, it can be seen that RNN, SGDM-LSTM, Adam-LSTM, and RMSProp-LSTM machine learning models based on cyclic structure have better predictive performance than Wavelet models. And the predictive performance of the LSTM model is superior to RNN model, especially when using the RMSProp optimizer to train the LSTM network, all indicators of predictive performance are smaller. From this experimental result, we can conclude that the LSTM model has stronger ability in predicting epidemic trends than RNN. RNN usually only considers the current input and some previously processed historical inputs, and there are problems with gradient vanishing and gradient explosion. The LSTM network structure introduces gating mechanism, which can more accurately identify the basic patterns and trend behaviors in COVID-19 time series data. However, LSTM is inferior to RNN in training speed and interpretability. For the three optimizers, SGDM is the most common optimizer with almost no acceleration effect. RMSProp is an upgraded version of SGDM, which improves running speed by eliminating oscillations during gradient descent. Adam is an improved version of RMSProp. However, from the experimental results of this study, we can see that the performance of Adam seems to be worse than RMSProp, so it is not that the more advanced the optimizer, the better the results.

Stacking model prediction results

The stacking integrated model needs to integrate diverse prediction algorithms to facilitate a more comprehensive observation of the dataset from different perspectives of data space and data structure. In order to get a better integration effect, we not only analyzed the individual prediction performance of each base learner, but also considered the correlation between base learners. Finally, by comprehensively comparing the prediction effects of each model, we chose the "good but different" algorithm as the first-level base learner of the stacking model. We first made individual predictions on the base learner, comprehensively compared the prediction errors of individual models, and then measured their correlation by calculating the pearson correlation coefficient of the prediction errors between different models. The specific operation can be achieved using the corrMatPlot calculation function provided in the matlab library. The error correlations of each model algorithm are shown in Fig. 10.

The correlation between the prediction errors of the base learners we used is very high, which shows that each model has learned the effective features in the data set during the training process. Among them, SGDM-LSTM, Adam-LSTM, and RMSProp-LSTM have the highest error correlation. This is because three models are all based on LSTM network structure and only optimized using different optimizers. Although the optimization principles are different, the observation methods of the data are generally similar to a large extent. The ARIMA model and other network models have similar prediction results, and can extract information such as trends, seasonality, and residuals from time series data. And the predictive performance of the ARIMA model is poor, so we do not consider using the ARIMA model for fusion. However, the observation mode and principle of SVR are quite different from other algorithms, so the correlation of its prediction error is relatively low. In order to find the best combination of base learners, we used five combination methods to dynamically model based on different base learners: Model I (RMSProp-LSTM, SVR, ELM, RNN, Wavelet), Model II (RMSProp-LSTM, SVR, ELM, Wavelet), Model III (RMSProp-LSTM, SVR, ELM, RNN), Model IV (ELM, Wavelet, SVR), and Model V (RMSProp-LSTM, ELM, SVR). The second layer of meta learners not only needs to correct the deviation of the algorithm, but also needs to maintain a high generalization ability to prevent overfitting. Therefore, we chose a simpler RBF algorithm and used PSO algorithm to optimize it to improve its prediction performance.

After completing the training of each base learner, we used the generated new dataset to train the meta learner and output the final results. In order to more intuitively represent the prediction performance of the stacking integrated model, the real values were compared with the prediction results of five stacking models. Figure 11 shows the prediction results of the stacking model on the test set.

The error evaluation indicators of single model and stacking integrated model are shown in Table 5. By analyzing Fig. 11; Table 5, we can conclude that among all single prediction models, the error evaluation indicators of SVR and RMSProp-LSTM were smaller, indicating better prediction performance. Among all prediction models based on stacking integration, the MAE, RMSE, and MAPE values of Model III were 0.135, 0.162, and 0.204, respectively, with the smallest error index and the best prediction effect. This indicates that the base learners composed of RMSProp-LSTM, SVR, ELM, and RNN models can maximize the prediction accuracy of stacking model fusion. In addition, the MAE values of the five stacking integrated models based on different base learners proposed by us were all around 0.15, the RMSE values were all around 0.18, and the MAPE values were all around 0.23. Compared to all single prediction models, they have better prediction performance. For example, the values of MAE, RMSE, and MAPE in Model I were 0.185, 0.221, and 0.269, respectively. Among the five stacking models, the three error indicators of Model I were all the highest, indicating the worst prediction performance. However, compared to the excellent SVR and RMSProp-LSTM, their prediction effect is more excellent.

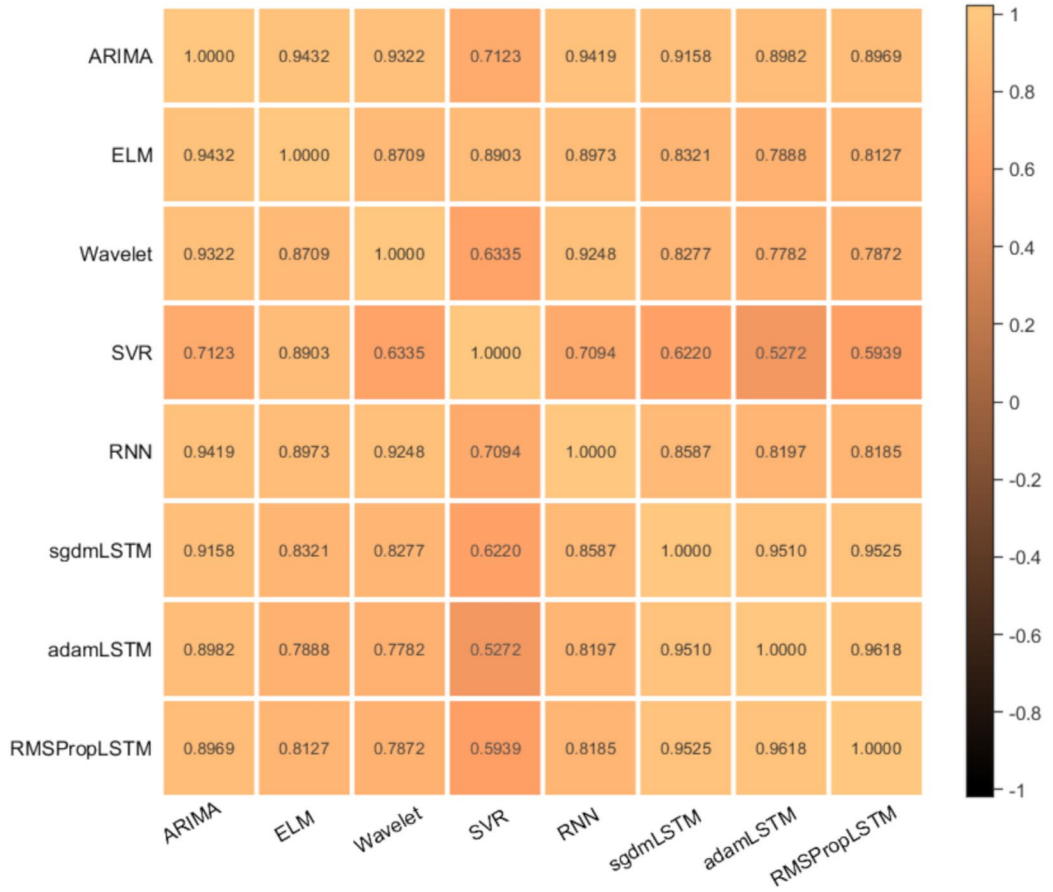

Fig. 10. Correlation plot of base learner prediction error.

In order to quantify the difference in prediction performance between the stacking integrated prediction model and the base learner, as shown in Table 6, we calculated the improvement rates of three evaluation indicators I_{index}. The calculation of improvement rate can help us further understand the prediction accuracy and practical value of integrated models, providing reference for model optimization and improvement. Table 6 records the improvement in prediction performance of the stacking integrated prediction model compared to single prediction models. The results showed that the algorithm set with the lowest prediction accuracy was Model I: RMSProp-LSTM, SVR, ELM, RNN and Wavelet, and the algorithm set with the highest prediction accuracy was Model III: RMSProp-LSTM, SVR, ELM and RNN. For example, for Model III, compared to ARIMA, ELM, SVR, Wavelet, RNN, SGDM-LSTM, Adam-LSTM, and RMSProp-LSTM, the MAE index decreased by 80.12%, 71.99%, 48.28%, 56.45%, 70.65%, 65.65%, 66.25%, and 58.72%, respectively. The RMSE index decreased by 81.03%, 81.03%, 45.27%, 61.06%, 71.58%, 59.70%, 63.27%, and 60.20%, respectively, while the MAPE index decreased by 34.62%, 34.62%, 37.80%, 40.18%, 40.00%, 36.45%, 35.85%, and 30.85%, respectively. In addition, in all integrated experiments, the maximum decreases in the evaluation indicators MAE, RMSE, and MAPE of the stacking integrated prediction model were 80.12%, 81.03%, and 40.18%.

From the above experimental results, it can be seen that using different prediction models can lead to different prediction results. ARIMA prediction model is very flexible, but its prediction accuracy is low for unstable or abrupt trend data sets. ELM is fast in learning and can solve problems quickly and efficiently, but its regularization parameters and the number of hidden neurons are difficult to determine, so we need to constantly adjust the parameters in the experiment. In this experiment, SVR has achieved better prediction accuracy. Compared with other regression algorithms, SVR can perform well even when the training dataset is small, making it possible to train the model in situations of data shortage. Wavelet can predict nonlinear and non-stationary time series data more accurately than traditional prediction algorithms such as ARIMA. Compared with other neural network models, RNN have a slower training speed and require multiple iterations to obtain more accurate prediction results. This is because RNN needs to consider all previous information when processing sequence data, and therefore needs to repeatedly calculate and update parameters to achieve the optimal state of the model. The LSTM model has higher prediction accuracy and better generalization ability, but requires more time and resources for training. In general, compared with the above prediction model, the stacking integrated model can use a variety of algorithms to observe the COVID-19 dataset from different perspectives, which makes the prediction dimension of the integrated model more comprehensive. To some extent, it overcomes

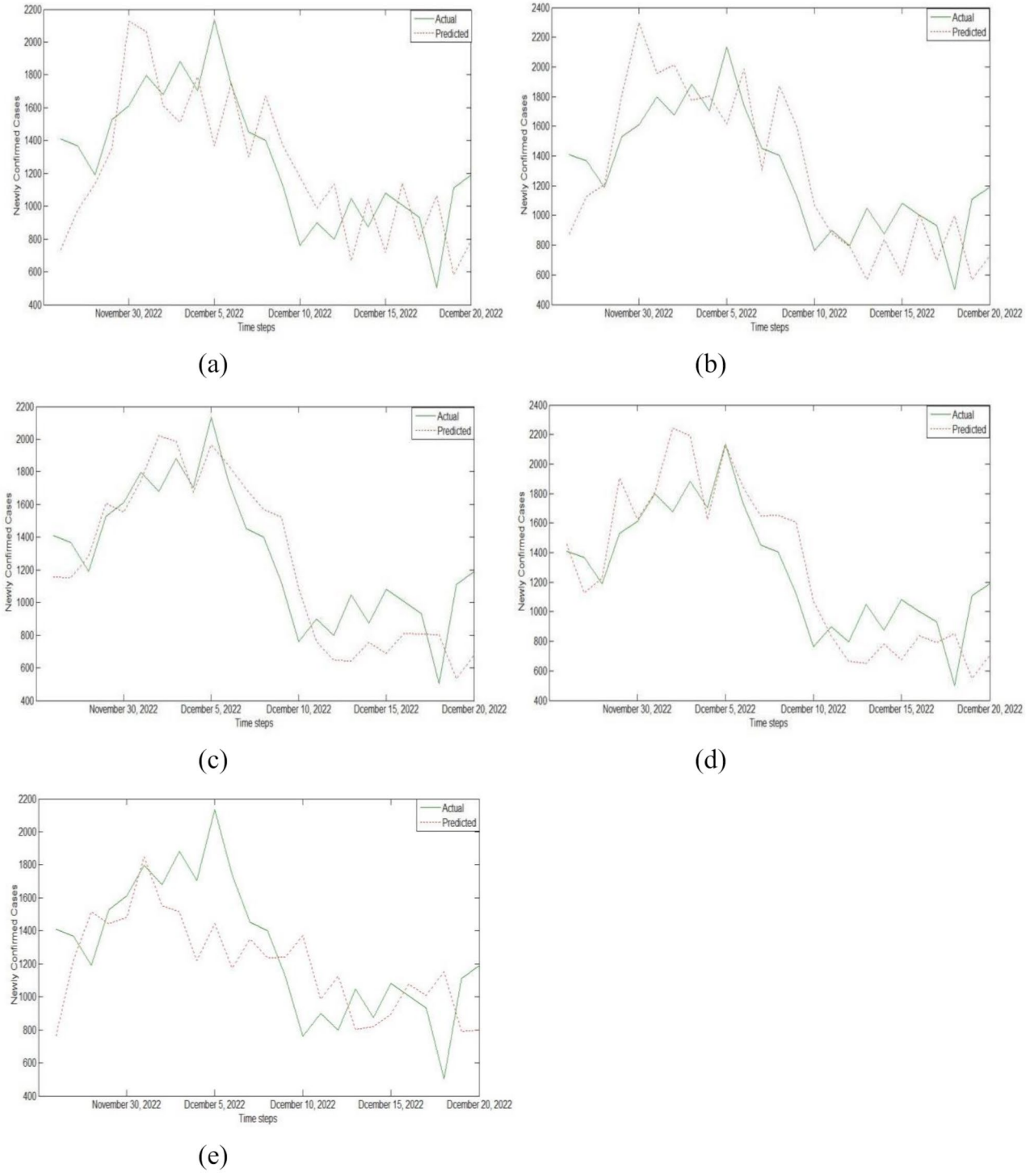

Fig. 11. Prediction results of training on test data sets using (**a**) Model **I**, (**b**) Model II, (**c**) Model III, (**d**) Model IV, and (**e**) Model V.

the limitations of individual learners and can effectively combine the characteristics and advantages of various learners to improve prediction accuracy and stability.

In order to verify the effectiveness of the model in this paper, we collected two groups of data, including the monthly incidence of AIDS in Anhui Province from January 2005 to December 2018 and the monthly incidence of tuberculosis in Shanxi Province from January 2005 to December 2017, and then applied the above stacking framework to establish the prediction model. Figure 12 shows the historical incidence data of AID and PTB. The number of AIDS cases is on the rise, while the number of tuberculosis cases is on the decline, and there is a certain periodicity. Before 2012, the number of new cases of AIDS per month was relatively small, and the

Type	Algorithm	Error evaluation indicators		
		MAE	RMSE	MAPE
Base-learner	ARIMA	0.679	0.854	0.312
	ELM	0.482	0.615	0.338
	SVR	0.261	0.296	0.328
	Wavelet	0.310	0.416	0.341
	RNN	0.460	0.570	0.340
	SGDM-LSTM	0.393	0.402	0.321
	Adam-LSTM	0.400	0.441	0.318
	RMSProp-LSTM	0.327	0.407	0.295
Stacking	Model I	0.185	0.221	0.269
	Model II	0.181	0.219	0.253
	Model III	**0.135**	**0.162**	**0.204**
	Model IV	0.143	0.179	0.212
	Model V	0.172	0.214	0.253

Table 5. Compare the error evaluation indexes of different base learners and Stacking integrated models on the COVID-19 dataset. Note: In this table, the base learner of Model I is composed of RMSProp-LSTM, SVR, ELM, RNN and Wavelet. The base learner of Model II is composed of RMSProp-LSTM, SVR, ELM and Wavelet. The base learner of Model III is composed of RMSProp-LSTM, SVR, ELM and RNN. The base learner of Model IV is composed of ELM, Wavelet and SVR. The base learner of Model V is composed of RMSProp-LSTM, ELM and SVR.

fluctuation was relatively small. However, since 2012, the number of new cases of AIDS per month has increased rapidly and the fluctuation has become more obvious, while the incidence trend of tuberculosis is the opposite.

On two sets of target datasets (AIDS and PTB), the data can be processed similarly to the previous COVID-19 data, divided into a training/validation set and a testing set, and trained through six base learners. The output results of four sub models, including RMSProp-LSTM, SVR, ELM, and RNN, were used as inputs to the stacking model to construct the training and testing sets of RBF. It can be seen from Table 7 that the stacking prediction model's three evaluation indicators on two datasets are still smaller than a single prediction model, and its prediction performance is still optimal.

Discussion

In infectious disease research, geospatial statistical analysis has been widely used, such as tuberculosis. However, in the analysis and research of COVID-19, the spatial statistical analysis method is rarely used. Compared with traditional statistical analysis methods, spatial statistical analysis can study the spatial distribution and spatial relationship of regionalized variables, and focus more on the spatial dependence and spatial correlation of data. Therefore, this analysis method is more suitable for analyzing the spread trend of COVID-19. In this study, spatial autocorrelation technology was used to analyze the global and local spatial correlation characteristics of COVID-19 incidence in China, and the center of gravity migration trajectory algorithm was used to detect the temporal and spatial evolution of epidemic trend in recent years. The results show that the global Moran index of COVID-19 cases in China change from negative to positive in 2020, 2021 and 2022, indicating that the global spatial correlation degree change from weak to strong, and it has strong spatial dependence on the provincial scale. Moreover, the overall prevalence of the epidemic has moved on a large scale, and the spatial and temporal pattern has generally moved to the southeast. This may be closely related to various factors such as population mobility and social intervention. By conducting in-depth analysis and monitoring of epidemic data, we can better understand its dynamic characteristics and evolution patterns, and take corresponding measures and strategies to control the spread and harm of the epidemic.

After spatiotemporal analysis, it was found that Guangdong Province is a high-risk area for the epidemic in 2022, so we chose it as the region to establish the prediction model. This study first used ARIMA, ELM, Wavelet, SVR, RNN, and LSTM models for modeling and predicting, and used three times day forward chaining cross-validation to reduce the risk of overfitting the model. Finally, we adopted the stacking technology to integrate the sub models mentioned above, and further optimized the model using the PSO algorithm. The results show that through this integration technology and optimization method, the problem of weak generalization ability of a single model can be effectively compensated, and the prediction performance of the model can be greatly improved. This proves that the stacking architecture has broad application prospects and value in the field of COVID-19 trend prediction.

In addition, the model in this study can also be used to analyze and predict other infectious diseases, and study the development trend and spatial distribution of different infectious diseases. Based on the transfer learning idea, we applied the trained model to predict the number of AIDS and tuberculosis cases. The experimental results show that the stacking model has a better prediction performance than the single prediction model. For infectious diseases, various regions around the world bear a heavy burden. Establishing a reliable prediction model can provide early understanding of the future epidemic trend and spread scale of infectious diseases,

Algorithm	Evaluation indicators	Stacking vs. Base-learner			
		ARIMA	ELM	SVR	Wavelet
Model I	I_{MAE}(%)	72.75	61.62	29.12	40.32
	I_{RMSE}(%)	74.12	74.12	25.34	46.88
	I_{MAPE}(%)	13.78	13.78	17.99	21.11
Model II	I_{MAE}(%)	73.34	62.45	30.65	41.61
	I_{RMSE}(%)	74.36	74.36	26.01	47.36
	I_{MAPE}(%)	18.91	18.91	22.87	25.81
Model III	I_{MAE}(%)	**80.12**	71.99	48.28	56.45
	I_{RMSE}(%)	**81.03**	**81.03**	45.27	61.06
	I_{MAPE}(%)	34.62	34.62	37.80	**40.18**
Model IV	I_{MAE}(%)	78.94	70.33	45.21	53.87
	I_{RMSE}(%)	79.04	79.04	39.53	56.97
	I_{MAPE}(%)	32.05	32.05	35.37	37.83
Model V	I_{MAE}(%)	74.67	64.32	34.10	44.52
	I_{RMSE}(%)	74.94	74.94	27.70	48.56
	I_{MAPE}(%)	18.91	18.91	22.87	25.81
Algorithm	**Evaluation indicators**	**Stacking vs. Base-learner**			
		RNN	**SGDM-LSTM**	**Adam-LSTM**	**RMSProp-LSTM**
Model I	I_{MAE}(%)	59.78	52.93	53.75	43.43
	I_{RMSE}(%)	61.23	45.02	49.89	45.70
	I_{MAPE}(%)	20.88	16.20	15.41	8.81
Model II	I_{MAE}(%)	60.65	53.94	54.75	44.65
	I_{RMSE}(%)	61.58	45.52	50.34	46.19
	I_{MAPE}(%)	25.59	21.18	20.44	14.24
Model III	I_{MAE}(%)	70.65	65.65	66.25	58.72
	I_{RMSE}(%)	71.58	59.70	63.27	60.20
	I_{MAPE}(%)	40.00	36.45	35.85	30.85
Model IV	I_{MAE}(%)	68.91	63.61	64.25	56.27
	I_{RMSE}(%)	68.60	55.47	59.41	56.02
	I_{MAPE}(%)	37.65	33.96	33.33	28.14
Model V	I_{MAE}(%)	62.61	56.23	57.00	47.40
	I_{RMSE}(%)	62.46	46.77	51.47	47.42
	I_{MAPE}(%)	25.59	21.18	20.44	14.24

Table 6. The improvement rate of the stacking integrated model compared to the three error indicators of different base learners.

help public health institutions take timely measures to curb the spread of diseases, and ensure public health and safety.

The first limitation of our work is that the study relies on COVID-19 epidemic data from various provinces and autonomous regions in China from 2020 to 2022. The diversity of data sources and differences in collection methods may affect the completeness and consistency of the data, thereby affecting the accuracy and predictive ability of the model. The second limitation is that the study did not fully consider the impact of external factors such as socio-economic changes, public health policies, and public behavior on the development of the epidemic. These factors may play a key role in the spread and development of the epidemic, but have not been reflected in the model. In response to these limitations, we will expand the scope of the data, use more complex models, and consider more influencing factors in our future studies to improve our ability to analyse and predict outbreaks.

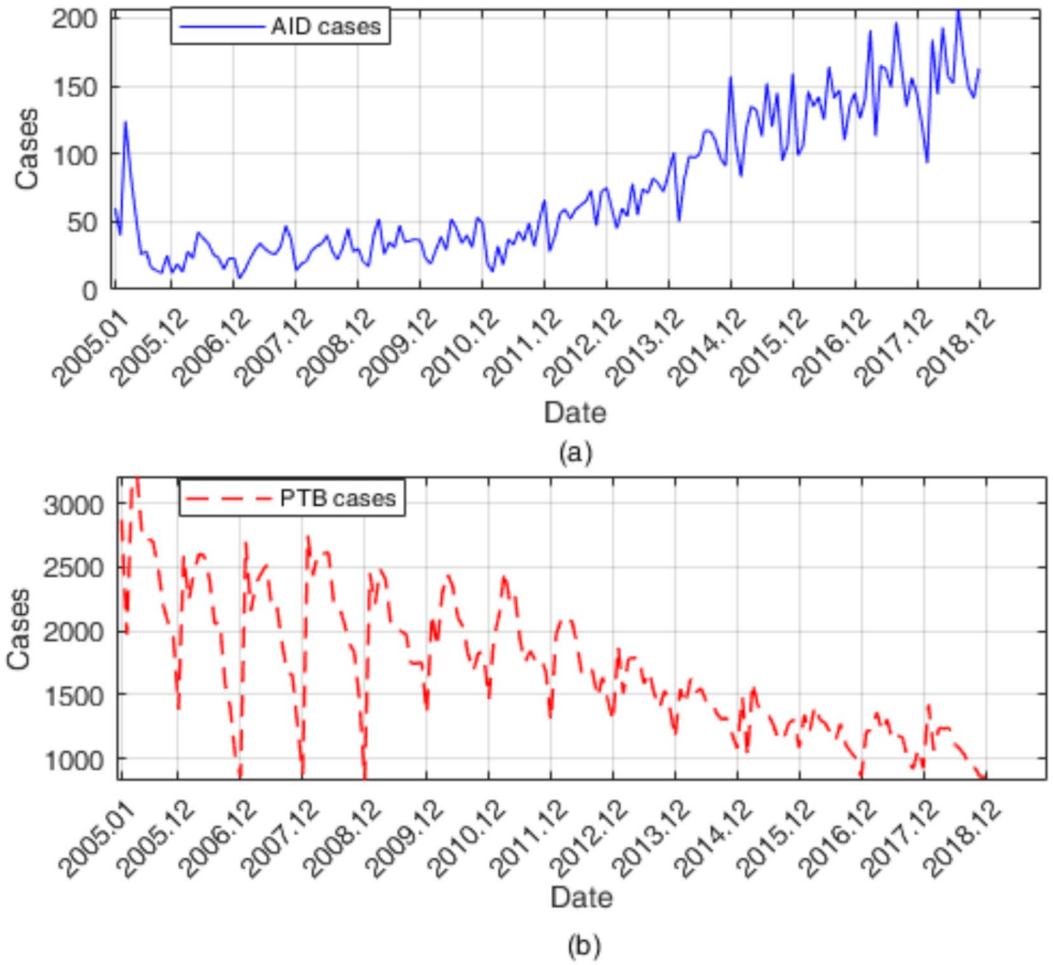

Fig. 12. The monthly incidence trend of (**a**) AIDS and (**b**) PTB over the years.

Algorithm	AID			PTB		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE
ARIMA	0.364	0.380	0.183	0.203	0.265	0.101
ELM	0.420	0.519	0.200	0.221	0.243	0.253
SVR	0.310	0.391	0.180	0.341	0.375	0.393
Wavelet	0.402	0.455	0.251	0.333	0.356	0.286
RNN	0.495	0.609	0.257	0.320	0.236	0.233
SGDM-LSTM	0.453	0.537	0.208	0.323	0.228	0.231
Adam-LSTM	0.482	0.594	0.220	0.354	0.252	0.248
RMSProp-LSTM	0.328	0.405	0.179	0.317	0.235	0.230
Stacking	**0.210**	**0.261**	**0.161**	**0.185**	**0.195**	**0.082**

Table 7. Compare the error evaluation indexes of different base learners and Stacking integrated models on the AID dataset and pulmonary tuberculosis dataset.

Data availability

These data were collected by the author from daily announcements issued by the National Health Commission of China. The data provided in this study can be obtained from the first author (E-mail addresses: 1556275877@qq.com).

Received: 29 March 2024; Accepted: 14 November 2024
Published online: 19 November 2024

References

1. Fu, Y., Cheng, Y. & Wu, Y. Understanding SARS-CoV-2-Mediated inflammatory responses: from mechanisms to potential therapeutic tools. *Virol. Sin.* **35** (3), 266–271 (2020). https://doi.org/10.1007/s12250-020-00207-4
2. Lai, C-C., Shih, T-P., Ko, W-C., Tang, H-J. & Hsueh, P-R. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): the epidemic and the challenges. *Int. J. Antimicrob. Agents*. **3** (55), 1–8 (2020). https://doi.org/10.1016/j.ijantimicag.2020.105924
3. Huang, C. et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet*. **395** (10223), 497–506 (2020). https://doi.org/10.1016/S0140-6736(20)30183-5
4. Musleh Alsartawi, A., Hegazy, M. A. A. & Hegazy, K. Guest editorial: the COVID-19 pandemic: a catalyst for digital transformation. *Managerial Auditing J.* **37** (7), 769–774 (2022). https://doi.org/10.1108/MAJ-07-2022-024
5. Han, L. et al. Exploring the clinical characteristics of COVID-19 clusters identified using factor analysis of mixed data-based cluster analysis. *Front. Med.* **8** (644724), 497–506 (2021). https://doi.org/10.3389/fmed.2021.644724
6. Al-Shargabi, A. A. & Selmi, A. Social Network Analysis and Visualization of Arabic tweets during the COVID-19 pandemic. *IEEE Access*. **9**, 90616–90630 (2021). https://doi.org/10.1109/ACCESS.2021.3091537
7. Ranasinghe, L. et al. Global impact of COVID-19 on childhood tuberculosis: an analysis of notification data. *Lancet Global Health*. **10** (12) (2022). https://doi.org/10.1016/S2214-109X(22)00414-4
8. Mahittikorn, A. et al. Elevation of serum interleukin-1 levels as a potential indicator for malarial infection and severe malaria: a meta-analysis. *Malar. J.* **21** (1) (2022). https://doi.org/10.1186/s12936-022-04325-0
9. Yang, W., Zhang, J. & Ma, R. The prediction of infectious diseases: a bibliometric analysis. *Int. J. Environ. Res. Public Health*. **17** (17), 1–19 (2020). https://doi.org/10.3390/ijerph17176218
10. Guo, K. et al. Traffic data-empowered xgboost-lstm framework for infectious disease prediction. *IEEE Trans. Intell. Transp. Syst.* (2022). https://doi.org/10.1109/TITS.2022.3172206
11. Guo, X. et al. Predicting the trend of infectious diseases using grey self-memory system model: a case study of the incidence of tuberculosis. *Public. Health.* **201**, 108–114 (2021). https://doi.org/10.1016/j.puhe.2021.09.025
12. Chae, S., Kwon, S. & Lee, D. Predicting Infectious Disease using Deep Learning and Big Data. *Int. J. Environ. Res. Public Health*. **15** (8) (2018). https://doi.org/10.3390/ijerph15081596
13. Tobler, W. R. A computer movie simulating urban growth in the Detroit Region. *Econ. Geogr.* **46**, 234–240 (1970). https://doi.org/10.2307/143141
14. Sarfo, A. K. & Karuppannan, S. Application of Geospatial Technologies in the COVID-19 fight of Ghana. *Trans. Indian Natl. Trans. Indian Natl. Acad. Engineering: Int. J. Eng. Technol.* **5**, 193–204 (2020). https://doi.org/10.1007/s41403-020-00145-3
15. Hertelendy, A. J. & Goniewicz, K. The COVID-19 pandemic: how predictive analysis, artificial intelligence and GIS can be integrated into a clinical command system to improve disaster response and preparedness. *Am. J. Emerg. Med.* **45**, 671–672 (2021). https://doi.org/10.1016/j.ajem.2020.10.049
16. Murugesan, M. et al. Epidemiological investigation of the COVID-19 outbreak in Vellore district in South India using Geographic Information Surveillance (GIS). *Int. J. Infect. Dis.* **112**, 669–675 (2022). https://doi.org/10.1016/j.ijid.2022.07.010
17. Ahasan, R., Alam, M. S., Chakraborty, T. & Hossain, M. M. Applications of GIS and geospatial analyses in COVID-19 research: a systematic review [version 1; peer review: 1 approved, 1 approved with reservations]. *F1000Research*. **9**, 1379 (2020). https://doi.org/10.12688/f1000research.27544.2
18. Dong, E. et al. The Johns Hopkins University Center for Systems Science and Engineering COVID-19 dashboard: data collection process, challenges faced, and lessons learned. *Lancet. Infect. Dis.* **22** (12), e370–e376 (2022). https://doi.org/10.1016/S1473-3099(22)00434-0
19. Shadeed, S. & Alawna, S. GIS-based COVID-19 vulnerability mapping in the West Bank, Palestine. *Int. J. Disaster Risk Reduct.* **64**, 102483 (2021). https://doi.org/10.1016/j.ijdrr.2021.102483
20. Valjarevic, A. et al. Modelling and mapping of the COVID-19 trajectory and pandemic paths at global scale: a geographer's perspective. *Open. Geosci.* **22** (1), 1603–1616 (2020). https://doi.org/10.1515/geo-2020-0156
21. Tiwari, A. & Aljoufie, M. A qualitative geographical information system interpretation of mobility and COVID-19 pandemic intersection in Uttar Pradesh, India. *Geospat Health*. **16** (1), 124–136 (2021). https://doi.org/10.4081/gh.2021.911
22. Kidd, D. M. & Liu, X. H. GEOPHYLOBUILDER 1.0: an ARCGIS extension for creating 'geophylogenies'. *Mol. Ecol. Resour.* **8** (1), 88–91 (2008). https://doi.org/10.1111/j.1471-8286.2007.01925.x
23. Haider, M. S. et al. Spatial distribution and mapping of COVID-19 pandemic in Afghanistan using GIS technique. *SN Social Sci.* **2** (5), 59 (2022). https://doi.org/10.1007/s43545-022-00349-0
24. Ramirez, I. J. & Lee, J. COVID-19 emergence and Social and Health Determinants in Colorado: a Rapid spatial analysis. *Int. J. Environ. Res. Public Health*. **77** (11), 3856 (2020). https://doi.org/10.3390/ijerph17113856
25. Nguyen, Q. C. et al. Using 164 million Google Street View images to derive built Environment predictors of COVID-19 cases. *Int. J. Environ. Res. Public Health*. **17** (17), 6359 (2020). https://doi.org/10.3390/ijerph17176359
26. Razavi-Termeh, S. V., Sadeghi-Niaraki, A., Farhangi, F. & Choi, S. M. COVID-19 risk mapping with considering Socio-Economic Criteria using machine learning algorithms. *Int. J. Environ. Res. Public Health*. **18** (18), 9657 (2020). https://doi.org/10.3390/ijerph18189657
27. Faisal, K. et al. Spatial analysis of COVID-19 Vaccine centers distribution: a case study of the City of Jeddah, Saudi Arabia. *Int. J. Environ. Res. Public Health*. **19** (6), 3526 (2022). https://doi.org/10.3390/ijerph19063526
28. Krzysztofowicz, S. & Osinska-Skotak, K. The use of GIS technology to optimize COVID-19 vaccine distribution: a case study of the City of Warsaw, Poland. *Int. J. Environ. Res. Public Health*. **18** (11), 5636 (2021). https://doi.org/10.3390/ijerph18115636
29. Sarkar, S. K. & Morshed, M. M. Spatial priority for COVID-19 vaccine rollout against limited supply. *HELIYON*. **7** (11), e08419 (2021). https://doi.org/10.1016/j.heliyon.2021.e08419
30. Office of the Assistant Secretary for Planning and Evaluation. Disparities in COVID-19 vaccination rates across racial and ethnic minority groups in the United States. (2021). https://aspe.hhs.gov/sites/default/files/private/pdf/265511/vaccination-disparities-brief.pdf
31. Wu, T. Y. et al. Using Community Engagement and Geographic Information Systems to address COVID-19 vaccination disparities. *Trop. Med. Infect. DISEASE*. **7** (8), 177 (2022). https://doi.org/10.3390/tropicalmed7080177
32. Elsheikh, R. F. Covid-19's pandemic relationship to Saudi Arabia's Weather using statistical analysis and GIS. *Comput. Syst. Sci. Eng.* **42** (2), 813–823 (2022). https://doi.org/10.32604/csse.2022.021645
33. Abulibdeh, A. & Mansour, S. Assessment of the effects of Human mobility restrictions on COVID-19 prevalence in the Global South. *Prof. Geogr.* **74** (1), 16–30 (2022). https://doi.org/10.1080/00330124.2021.1970592
34. Azevedo, L. et al. Geostatistical COVID-19 infection risk maps for Portugal. *Int. J. Health Geogr.* **19**, 1–8 (2020). https://doi.org/10.1186/s12942-020-00221-5
35. Ribeiro, M., Azevedo, L. & Pereira, M. J. EpiGeostats: an R Package to facilitate visualization of Geostatistical Disease Risk maps. *Math. Geosci.* **56**, 103–119 (2024). https://doi.org/10.1007/s11004-023-10080-y
36. Alvo, M. & Mu, J. COVID-19 Data Analysis using Bayesian models and nonparametric geostatistical models. *Mathematics*. **11** (6), 1359 (2023). https://doi.org/10.3390/math11061359
37. Wang, Y. L. et al. An intelligent forecast for COVID-19 based on single and multiple features. *Int. J. Intell. Syst.* **37** (11), 9339–9356 (2022). https://doi.org/10.1002/int.22995

38. Jia, L., Li, K., Jiang, Y., Guo, X. & Zhao, T. Prediction and analysis of Coronavirus Disease 2019. (2020). https://doi.org/10.48550/arXiv.2003.05447
39. Omaret, O. A. M., Elbarkouky, R. A. & Ahmed, H. M. Fractional stochastic models for COVID-19: case study of Egypt. *RESULTS Phys.* **23**, 104018 (2021). https://doi.org/10.1016/j.rinp.2021.104018
40. Li, W. et al. An evaluation of COVID-19 transmission control in Wenzhou using a modified SEIR model. *Epidemiol. Infect.* **149** (2021). https://doi.org/10.1017/S0950268820003064
41. Ogunjo, S. T., Fuwape, I. A. & Rabiu, A. B. Predicting COVID-19 cases from Atmospheric Parameter using Machine Learning Approach. *GEOHEALTH*. **6** (4) (2021). https://doi.org/10.1029/2021GH000509
42. Nesa, M. K., Babu, M. R. & Mamun Khan, M. T. Forecasting COVID-19 situation in Bangladesh. *Biosaf. Health*. **4** (1), 6–10 (2022). https://doi.org/10.1016/j.bsheal.2021.12.003
43. Rguibi, M. A., Moussa, N., Madani, A., Aaroud, A. & Zine-Dine, K. Forecasting Covid-19 transmission with ARIMA and LSTM techniques in Morocco. *SN Comput. Sci.* **3** (2), 133 (2022). https://doi.org/10.1007/s42979-022-01019-x
44. Mangla, S., Pathak, A. K., Arshad, M. & Haque, U. Short-term forecasting of the COVID-19 outbreak in India. *Int. HEALTH*. **13** (5), 410–420 (2021). https://doi.org/10.1093/inthealth/ihab031
45. Alsartawi, A. M., Hegazy, M. A. A. & Hegazy, K. Guest editorial: the COVID-19 pandemic: a catalyst for digital transformation. *MANAGERIAL AUDITING J.* **37** (7), 769–774 (2022). https://doi.org/10.1108/MAJ-07-2022-024
46. Biswas, S. Forecasting and comparative analysis of Covid-19 cases in India and US. *Eur. Phys. JOURNAL-SPECIAL Top.* **231** (18–20), 3537–3544 (2022). https://doi.org/10.1140/epjs/s11734-022-00536-3
47. Namasudra, S., Dhamodharavadhani, S. & Rathipriya, R. Nonlinear neural network based forecasting model for Predicting COVID-19 cases. *Neural Process. Lett.* **55** (1), 171–191 (2023). https://doi.org/10.1007/s11063-021-10495-w
48. Fatimah, B., Aggarwal, P., Singh, P. & Gupta, A. A comparative study for predictive monitoring of COVID-19 pandemic. *Appl. Soft Comput.* **22**, 108806 (2022). https://doi.org/10.1016/j.asoc.2022.108806
49. Ly, K. T. A COVID-19 forecasting system using adaptive neuro-fuzzy inference. *FINANCE Res. Lett.* **41**, 101844 (2021). https://doi.org/10.1016/j.frl.2020.101844
50. Zhang-James, Y. et al. A seq2seq model to forecast the COVID-19 cases, deaths and reproductive R numbers in US counties. *Res. Square*. (2021).
51. Appadu, A. R., Kelil, A. S. & Tijani, Y. O. Comparison of some forecasting methods for COVID-19. *ALEXANDRIA Eng. J.* **60** (1), 1565–1589 (2021). https://doi.org/10.1016/j.aej.2020.11.011
52. Alruily, M. et al. Prediction of COVID-19 transmission in the United States using Google Search trends. *CMC-COMPUTERS Mater. CONTINUA*. **70** (1), 1751–1768 (2022). https://doi.org/10.1016/j.aej.2020.11.011
53. Zeroual, A., Harrou, F., Dairi, A. & Sun, Y. Deep learning methods for forecasting COVID-19 time-series data: a comparative study. *CHAOS SOLITONS FRACTALS*. **140**, 110121 (2020). https://doi.org/10.1016/j.chaos.2020.110121
54. Liu, Q., Xie, W. J. & Xia, J. B. Using Semivariogram and Moran's I techniques to evaluate spatial distribution of Soil micronutrients. *Commun. Soil Sci. Plant Anal.* **44** (7), 1182–1192 (2013). https://doi.org/10.1080/00103624.2012.755999
55. Moran, P. A. P. Notes on continuous stochastic phenomena. *BIOMETRIKA*. **37** (1–2), 17–23 (1950). https://doi.org/10.2307/2332142
56. Saffary, T. et al. Analysis of COVID-19 cases' spatial dependence in US counties reveals Health inequalities. *Front. PUBLIC. HEALTH*. **8**, 579190 (2020). https://doi.org/10.3389/fpubh.2020.579190
57. Tahkola, M. & Zou, G. Automated Time Series classification with sequential model-based optimization and nested Cross-validation. *IEEE ACCESS*. **10**, 39299–39312 (2022). https://doi.org/10.1109/ACCESS.2022.3166525
58. Li, H. J. et al. Temporal detection of sharp landslide deformation with ensemble-based LSTM-RNNs and Hurst exponent. *GEOMATICS Nat. HAZARDS RISK*. **12** (1), 3089–3113 (2021). https://doi.org/10.1080/19475705.2021.1994474
59. Aminanto, M. E., Ban, T., Isawa, R., Takahashi, T. & Inoue, D. Threat Alert Prioritization using isolation forest and stacked auto Encoder with Day-Forward-Chaining analysis. *IEEE ACCESS*. **8**, 217977–217986 (2020). https://doi.org/10.1109/ACCESS.2020.3041837
60. Chiu, C. C. et al. Applying an Improved Stacking Ensemble Model to predict the mortality of ICU patients with heart failure. *J. Clin. Med.* **11** (21), 6460 (2022). https://doi.org/10.3390/jcm11216460
61. Jia, J. H., Wu, G. Q. & Qiu, W. R. pSuc-FFSEA: Predicting lysine Succinylation sites in proteins based on Feature Fusion and Stacking Ensemble Algorithm. *Front. CELL. Dev. BIOLOGY*. **10**, 894874 (2022). https://doi.org/10.3389/fcell.2022.894874
62. Wu, W. T., Xia, Y. S. & Jin, W. Z. Boosting Decision Trees. *IEEE Trans. Intell. Transp. Syst.* **22** (4), 2510–2523. https://doi.org/10.1109/TITS.2020.3035647 (2020). Predicting Bus Passenger Flow and Prioritizing Influential Factors Using Multi-Source Data: Scaled Stacking Gradient.
63. Cao, C., Song, S. Y., Chen, J. P., Zheng, L. J. & Kong, Y. Y. An Approach to predict debris Flow Average Velocity. *WATER*. **9** (3), 205 (2017). https://doi.org/10.3390/w9030205
64. Mao, L. et al. Online State-of-Health Estimation Method for Lithium-Ion Battery based on CEEMDAN for feature analysis and RBF neural network. *IEEE J. Emerg. Sel. Top. POWER Electron.* **11** (1), 187–200 (2023). https://doi.org/10.1109/JESTPE.2021.3106708
65. You, D. Z., Lei, Y. M., Liu, S., Zhang, Y. P. & Zhang, M. Networked Control System based on PSO-RBF neural Network Time-Delay Prediction Model. *Appl. SCIENCES-BASEL*. **13** (1), 536 (2023). https://doi.org/10.3390/app13010536

Author contributions

Cheng, Y. Y participated in data analysis and manuscript writing. Cheng, Y. Y and Bai, Y. P proposed the main structure of this study. Tan, X. H., Xu, T and Cheng, R provided useful suggestions and revised the manuscript. All authors have read and approved the final manuscript.

Funding

This research is funded by the Fundamental Research Program of Shanxi Province, China (Grant No. 202103021224195, 202103021223189, 202103021224212, 20210302123019), the National Science Foundation of China, China (Grant No. 61774137).

Declarations

Ethical approval

All procedures of this study were performed in accordance with the 1964 Helsinki Declaration and its later amendments. All procedures of this study were approved by the North University of China institutional review board. This article does not contain any studies with animals performed by the author.

Informed consent

All participants completed informed consent.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to Y.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.