Exploring the influence of COVID-19 on the spread of hand, foot, and mouth disease with an automatic machine learning prediction model

Chuan Yang1 · Shuyi An2 · Baojun Qiao2 · Peng Guan1 · Desheng Huang3 · Wei Wu1

Received: 5 July 2022 / Accepted: 11 October 2022 / Published online: 18 October 2022 © The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

Abstract
Hand, foot, and mouth disease (HFMD) is an important public health problem and has received concern worldwide. Moreover, the coronavirus disease 2019 (COVID-19) epidemic also increases the difficulty of understanding and predicting the prevalence of HFMD. The purpose is to prove the usability and applicability of the automatic machine learning (Auto-ML) algorithm in predicting the epidemic trend of HFMD and to explore the influence of COVID-19 on the spread of HFMD. The AutoML algorithm and the autoregressive integrated moving average (ARIMA) model were applied to construct and validate models, based on the monthly incidence numbers of HFMD and meteorological factors from May 2008 to December 2019 in Henan province, China. A total of four models were established, among which the Auto-ML model with meteorological factors had minimum RMSE and MAE in both the model constructing phase and forecasting phase (training set: RMSE = 1424.40 and MAE = 812.55; test set: RMSE = 2107.83, MAE = 1494.41), so this model has the best performance. The optimal model was used to further predict the incidence numbers of HFMD in 2020 and then compared with the reported cases. And, for analysis, 2020 was divided into two periods. The predicted incidence numbers followed the same trend as the reported cases of HFMD before the COVID-19 outbreak; while after the COVID-19 outbreak, the reported cases have been greatly reduced than expected, with an average of only about 103 cases per month, and the incidence peak has also been delayed, which has led to significant changes in the seasonality of HFMD. Overall, the AutoML algorithm is an applicable and ideal method to predict the epidemic trend of the HFMD. Furthermore, it was found that the countermeasures of COVID-19 have a certain influence on suppressing the spread of HFMD during the period of COVID-19. The findings are helpful to health administrative departments.

Keywords HFMD · Time series · Prediction · Automatic machine learning · COVID-19 · Countermeasures

Introduction
Hand, foot, and mouth disease (HFMD) is an acute infectious disease caused by a variety of human enteroviruses, of which enterovirus 71 (EV71) and Coxsackievirus A16...
(Cox A16) are the most common pathogens (Hu et al. 2021; Puenpa et al. 2019). The susceptible population is mainly preschool children, especially children under 5 years old (Jijun Zhao et al. 2016). The transmission routes of HFMD are complex and diverse, such as close personal contact, contact with objects contaminated by pathogens, respiratory pathway, and fecal–oral way pathway (Qi et al. 2020). Most patients show self-limiting illness typically including fever rash on the hands, feet, and buttocks; mouth ulcers; inappetence; and diarrhea and vomiting, but a small minority may experience more severe complications including myocarditis, pulmonary edema, and aseptic meningoencephalitis, or even death (Yu Wang et al. 2011; J. Xu et al. 2020).

In 1957, HFMD was first identified in New Zealand and then was frequently reported worldwide (Chen et al. 2019; Sumi et al. 2017; Van Pham et al. 2019). Since the EV71 was discovered in Vietnam in 1969, the Asia–Pacific region, including China, has experienced various degrees of HFMD outbreaks (Van Tu et al. 2007). HFMD has been evaluated to cause about 96,900 age-weighted disability adjusted life years (DALYs) in eight high-burden countries in East and Southeast Asia annually (Koh et al. 2018). In China, after the first HFMD cases were reported in Shanghai in 1981, there have been multiple HFMD outbreaks (Y. Wang et al. 2019). In the spring of 2008, the first large-scale and unprecedented HFMD outbreak occurred in Fuyang city of Anhui province, China, which caused extremely serious consequences due to the rapid development of the epidemic beyond control (Y. Zhang et al. 2010). Therefore, the Chinese Ministry of Health attached great importance to HFMD and listed it as a class C notifiable infectious diseases (Zhong et al. 2018). Since the routine surveillance was carried out, an average of about 2 million HFMD cases was reported annually, and the incidences and deaths caused by HFMD rank first in Class C infectious diseases every year (Xing et al. 2014; X. Zhang et al. 2016). Meanwhile, HFMD also ranks first among the susceptible infectious diseases of children under 5 years old (Zhao and Hu 2019). The EV71-inactivated vaccine was approved for the market in China in 2015, but the facts showed that the HFMD cases have not decreased significantly (Mao et al. 2016). HFMD has been around for a long time, which not only brings a heavy burden of disease to children but also some economic burden to society.

In December 2019, the coronavirus disease 2019 (COVID-19) began a pandemic, posing a formidable challenge to global public health (C. Wang et al. 2020). Then, the World Health Organization listed this epidemic as a “public health emergency of global concern,” and countries around the world successively took diverse prevention and control measures to deter the transmission of the epidemic. Because the COVID-19 was firstly identified in China, the national health commission of China has quickly implemented the first-level response to public health emergencies and launched corresponding control measures, including lockdown, use of masks, suspension of work and school, and restriction of crowd flow (Bangura et al. 2020; C. Wang et al. 2020). Several studies have shown that such countermeasures were effective not only in preventing the circulation of COVID-19, but also in the prevalence of other infectious diseases, such as influenza, pneumonia, and natural focal diseases (Cheng et al. 2021; Sakamoto et al. 2020; Wu et al. 2020). Therefore, COVID-19 has brought more complexity and uncertainty to the study of the prevalence tendency of HFMD, which has increased the difficulty of prediction. Meanwhile, this also poses a new challenge to implementing HFMD prevention and control measures that are more in line with the actual situation. At present, few studies have explored the effect of COVID-19 on the HFMD epidemic by combining forecasting model (Niu et al. 2021).

The incidence numbers of infectious diseases are crucial for the health administration departments to optimize decisions, understand the epidemic patterns, implement interventions, and ultimately minimize the harm of infectious diseases. Therefore, it is essential to construct a scientific and credible infectious disease prediction model. The existing time series prediction models for identifying and forecasting the trend of HFMD mainly consist of linear and nonlinear models, along with their hybrids (Y. Wang et al. 2019). Among the linear models, the autoregressive integrated moving average (ARIMA) has been widely used to predict various infectious diseases, such as hemorrhagic fever, hepatitis B, COVID-19, and HFMD (Ceylan 2020; Lv et al. 2021; Tian et al. 2019; Y.-W. Wang et al. 2018). Furthermore, many studies have demonstrated that meteorological factors significantly influenced the incidences of HFMD (Chen et al. 2019; Yi et al. 2019; J. Zhao et al. 2016). However, the ARIMA model only extracts the linear information from the time series, whereas most time series may not satisfy the assumption of linearity in practical (Zou et al. 2019).

Machine learning (ML) has made great improvements in exploiting the nonlinear information, so it has become popular to predict the HFMD (Gao et al. 2021). For example, eXTreme Gradient Boosting (XGBoost) models were built to forecast HFMD in Shenzhen based on daily level clinical data and multiple environment factors (Zhong et al. 2018). The multivariate back propagation neural networks (BP) model comprehensively integrated the autocorrelation of the HFMD incidence series, the climatic variables, and their hysteresis effects to predict the epidemic level of HFMD (W. Liu et al. 2019). Although there has been a big stride in predicting performance, ML is still not easy for non-experts to perform a lot of modeling-related tasks in most cases, such as the reserve of expertise in ML, the selection of algorithms, the optimization of models, and the requirement of more efforts. The automatic machine learning algorithm
(Auto-ML), developed based on traditional ML, provides an automated approach to training, tuning, and testing multiple ML algorithms before selecting the model that performed best according to the model’s evaluation criterion. Auto-ML combining the advantages of ML could save human hours by training the best model in the least amount of time, reduce the need for expertise in ML by shorting the manual code-writing time, and improve the performance of ML models. At present, AutoML has begun to be applied in the medical field to predict the incidences of diseases (Asfahan et al. 2020; Olsavszky et al. 2020). To our knowledge, however, there has been no study published on applying the Auto-ML algorithm to predict the prevalence of HFMD.

Hence, in this study, we have deployed the Auto-ML algorithm to construct predictive models for the incidence numbers of HFMD, and meteorological factors were also added to models to improve the prediction performance. Next, the optimal predictive model, selected by the prediction performance of models, was used to generate predicted values of incidence numbers of HFMD in 2020. The gap between predicted values and reported values could explore the influence of the various COVID-19 countermeasures measures on HFMD. This can help the health administration departments in optimizing decisions and implementing targeted prevention and control measures more efficiently during COVID-19.

Material and methods

Study area

Figure 1 shows the geographic location of Henan province in China. Henan province (110°21′E-116°39′E, 31°23′N-36°22′N), located in central China and bordering many provinces, is an important comprehensive transportation hub of China. Henan province with a total area of 167,000 km², most areas of which lie in the warm temperate zone and the south of which is across the subtropical zone, has a continental monsoonal climate characterized by complexity and diversity. According to the 8th national demographic census in 2020, the population of permanent residents amounts to 99.37 million, with a high population density.
Data sources

HFMD data

The clinical diagnostic criteria of HFMD cases are based on the “Guidelines for the diagnosis and treatment of HFMD” issued by the National Health Commission of the People’s Republic of China (National Health Commission 2018). The confirmed HFMD cases are identified in combination with epidemiological history, clinical manifestations, and pathogenic examinations. HFMD cases are required to be directly reported online to China Information System for Disease Control and Prevention (CISDCP) within 24 h of diagnosis (S. Liu et al. 2018). In this study, the monthly reported incidence numbers of HFMD were gathered from May 2008 to December 2020 which were published by the site of the Health Commission of Henan province. The basic information of HFMD data is shown in Table 1.

Meteorological data

The monthly meteorological data from May 2008 to December 2020 were downloaded from the China Meteorological Data Sharing Service System (http://data.cma.gov.cn/). There are a total of 17 meteorological factors, and their specific information is presented in Table 1. In addition, there are very few missing values in meteorological data, which were interpolated by using the mean of value from the month before and after the month with a missing value.

Model implementation

In this process, we adopted the following analysis strategies. First of all, the decomposition of time series was conducted to identify the patterns and components of the time series of HFMD and to provide meaningful insights. An additive structure was used to express time series data by the following equation:

\[ Y_t = T_t + S_t + I_t \]

where \( Y_t \) represents the observed series at time \( t \) and \( T_t, S_t, \) and \( I_t \) represent the value of trend, seasonal and random fluctuation of series at time \( t \), respectively.

And, the Spearman correlation coefficients, between the monthly HFMD data and the monthly meteorological factors and between the monthly meteorological factors, were calculated to form a correlation heatmap, which was aimed to select meteorological factors significantly associated with the incidence numbers of HFMD and avoid multicollinearity (W. Liu et al. 2019). The remaining meteorological factors filtered by the above steps are utilized as external variables of the prediction models in the subsequent process.

Table 1 Descriptive statistics of monthly HFMD cases and monthly meteorological factors in Henan province from May 2008 to December 2020

| Variables                                | Abbreviation | n  | Mean  | Min   | Median | Max    | SD   |
|------------------------------------------|--------------|----|-------|-------|--------|--------|------|
| Monthly HFMD cases                       |              | 152| 6741.00 | 66.00 | 4166.00 | 37640.00 | 7182.24 |
| Monthly Meteorological factors           |              |    |        |       |        |        |      |
| Minimum atmospheric pressure(hPa)        | APmin        | 152| 991.30 | 979.20 | 990.30 | 1009.80 | 7.34 |
| Mean temperature (℃)                     | Tm           | 152| 15.79  | 15.79 | 16.86  | 29.53  | 9.29 |
| Mean wind velocity in 2 min (m/s)        | WVm          | 152| 1.99   | 1.98  | 1.98   | 2.73   | 0.33 |
| Sunshine duration (hours)                | SD           | 152| 143.20 | 44.30 | 144.90 | 239.10 | 41.17 |
| Precipitation at 20–20 o’clock (mm)      | PRCP         | 152| 66.52  | 0.50  | 48.58  | 309.42 | 62.09 |
| Mean water vapor pressure (hPa)          | WVPm         | 152| 13.63  | 2.20  | 11.94  | 29.95  | 8.17 |
| Mean atmospheric pressure (hPa)          | APm          | 152| 1002.89| 988.00| 1003.75| 1020.63| 8.54 |
| Maximum atmospheric pressure (hPa)       | APmax        | 152| 1014.17| 994.43| 1015.49| 1036.33| 11.00 |
| Minimum temperature (℃)                 | Tmin         | 152| 5.61   | 5.61  | 5.76   | 21.78  | 9.77 |
| Maximum temperature (℃)                 | Tmax         | 152| 28.62  | 9.68  | 21.78  | 39.83  | 8.25 |
| Mean minimum temperature (℃)            | Tm_min       | 152| 11.60  | 11.60 | 12.24  | 25.68  | 9.18 |
| mean maximum temperature (℃)            | Tm_max       | 152| 20.92  | 3.83  | 22.44  | 34.30  | 9.28 |
| Extreme wind velocity (m/s)             | WVex         | 152| 14.45  | 10.35 | 14.39  | 20.80  | 2.11 |
| Maximum wind velocity (m/s)             | WVmax        | 152| 8.11   | 8.11  | 8.19   | 10.53  | 1.20 |
| Mean relative humidity (%)              | RHm          | 152| 65.13  | 41.00 | 64.80  | 84.75  | 9.83 |
| Days with daily ≥ 0.1 mm precipitation (days) | PRCP_d    | 152| 7.55   | 0.75  | 7.25   | 18.25  | 3.53 |
| Maximum daily precipitation (mm)        | PRCPmax      | 152| 27.19  | 0.50  | 22.64  | 140.75 | 23.52 |

SD standard deviation.
Secondly, the data set (152 months in total) was split into three subsets, namely training set, test set, and prediction set. The training set was aimed to train the predictive models, the test set was to evaluate the predictive performance of the models and the prediction set was to analyze changes in the prevalence of HFMD during the COVID-19. In this study, the monthly incidence numbers of HFMD data and monthly meteorological data from May 2008 to December 2018 (128 months in total) were used as the training set, the data for the whole year of 2019 (12 months in total) were as the test set, and the data for the whole year of 2020 (12 months in total) were as the prediction set.

Thirdly, for comparison, we used traditional time series analysis methods (ARIMA models) to serve as the baseline predictive models. Then, we built Auto-ML models to investigate the applicability and prediction performance of the AutoML algorithm for HFMD. All fitted models were validated and the optimal predictive model was selected by the model evaluation criterion.

Finally, based on the optimal model, the independent variables from 2020 were input to predict the incidence numbers of HFMD in 2020. Comparing the differences between predicted values and the reported values of the HFMD incidence numbers in 2020 to analyze and explore the influence of the COVID-19 epidemic and the various COVID-19 prevention and control measures on the spread of HFMD.

All analyses were based on Microsoft Excel 2019 and R software Version 4.1.1. Particularly, the ARIMA models were constructed with R package “forecast,” the ARIMAX model with R package “TSA,” and the Auto-ML models with R package “h2o”.

**ARIMAX model**

The ARIMA with exogenous input (ARIMAX) model is an extension of the ARIMA model. The construction of the ARIMAX model mainly included two steps: fitting multiple ARIMA models and putting the external variables into the best ARIMA model selected by selection criteria to further develop the ARIMAX model.

The ARIMA model is a model family, including three parts: autoregressive process (AR), establishing the relationship between the current values and the past values of the time series; integrated process (I), transforming the non-stationary time series to stationary through differencing; moving average process (MA), and focusing on the accumulation of error terms in the autoregressive process. The basic structure of the ARIMA model, typically expressed by ARIMA \((p,d,q) \times (P,D,Q)_s\) is written with the backshift notation as follows:

\[
\Theta_p(B^s)\theta_p(B)(1 - B^s)(1 - B)^d x_t = \Phi_q(B^s)\phi_q(B)w_t
\]

where \(\Theta_p, \theta_p, \Phi_q, \) and \(\phi_q\) are polynomials of orders \(P, p, Q, \) and \(q,\) respectively, \(d\) is differencing degree and \(s\) is the seasonal period (\(s = 12\) in this study).

The ARIMA model was constructed in two ways: automatic fitting model and manual fitting model. The ARIMA model was developed with the Box-Jenkins approach, which included the following four steps. Firstly, check the stationarity of the original time series, and differencing or seasonal differencing could be adopted if it was nonstationary. Secondly, calculate and plot the autocorrelation coefficient (ACF) and partial autocorrelation coefficient (PACF) of the above stationary time series to determine the possible model orders. Multiple candidate models were fitted with different model orders. Thirdly, identify whether the residual series of candidate models complied with white noise based on the Ljung-Box test. The Akaike Information Criteria (AIC) was used to examine the goodness-of-fit of the candidate models, the one with the lowest AIC values was the most appropriate model. Fourthly, use the test set to validate the model selected in the previous step.

Once the ARIMA model was determined, the ARIMAX model including the remaining meteorological factors as external variables was further constructed. Because certain meteorological factors may exert lagged effects on the incidence numbers of HFMD, cross-correlation analysis was performed to assess associations between them. To avoid one situation where cross-correlation may indicate spurious relationships between the input time series and the output time series caused by autocorrelation of the time series, a pre-whitening process was required before cross-correlation analysis (Dean and Dunsmuir 2016). Then, the lagged terms for these meteorological factors found to be significant were also considered as external variables. Finally, the ARIMAX model with meteorological factors and their lagged terms was constructed. This model was also validated against the test set.

**Auto-ML model**

The H2O AutoML algorithm is an open-source, distributed, and Java-based library for automating the machine learning workflow, which includes automatic training, tuning, and testing of multiple ML models. This integrates a variety of algorithms such as distributed random forest (DRF), deep learning (DL), extremely randomized trees (XRT), generalized linear model with regularization (GLM), XGBoost, and the Stacked Ensemble models. The establishment of AutoML models consists of four processes.

**Feature engineering** According to the preceding analysis, the main characteristics of the time series have been explored and learned about. These insights will be used to build new features as informative inputs for the ML model.
A total of five kinds of new features were created: series trend, seasonal component, series autocorrelation, remaining meteorological factors, and lagged terms of meteorological factors, respectively. The series trend was represented by a numeric index. In addition, as the series trend is not linear, a second polynomial of the index was used to capture the overall curvature of the series trend. To capture the series’ seasonality, we created a categorical variable for the month of the year. As an infectious disease, the current number of HFMD cases may be influenced by the number of past cases in a certain range. According to the result of ACF, we utilized the strong correlation of the series with its seasonal lag and non-seasonal lag as inputs to the model. And the lagged terms of meteorological factors were also taken as new features through the previous cross-correlation analysis.

Setting the hyperparameters. We empirically specify five-fold cross-validation to reduce the chance of overfitting of the models. This process randomly split the training set into 5 folders and then trained the model 5 times. Each time a different folder was left out as a testing partition, and the remaining 4 folders were used to train the model. Throughout this process, the model tuned the model’s parameters. The final model was tested on the testing partition. Then, specify “RMSE” as the optimization metric used to rank the Leaderboard frame of all the trained models at the end of the training process. In addition, to ensure the reproducibility of AutoML, a seed value of the random number generator was set, and DL algorithms that were not reproducible in the AutoML algorithm by default for performance reasons were excluded from Auto-ML built-in algorithms. The stopping strategy of the training process was set by specifying the argument about the maximum number of models of an Auto-ML run, which would exclude the Stacked Ensemble models and ensure reproducibility again.

Training Auto-ML models. “The HFMD incidence numbers data” and “the above new features” were individually identified as the dependent variable and the independent variables. Then, two types of Auto-ML models with different informative inputs were constructed at the end of twice training process. The one type models with series trend, seasonal component, series autocorrelation, and the other type with five all kinds of new features. Eventually, the two training processes each generate a Leaderboard frame, and the leader model selected from the frame was viewed as the best model for each type of model.

Validating the models. The two best models were validated by the test set.

**Model evaluation**

The predictive performances of four models were evaluated by using the root mean square error (RMSE) and the mean absolute error (MAE) to select the optimal model. The smaller the RMSE and MAE, the better the model is for forecasting.

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

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

**Results**

**General analysis**

There were 1024573 HFMD cases total from May 2008 to December 2020 in Henan province. Table 1 shows the summary statistics of the monthly incidence numbers of HFMD and the meteorological factors during the study period. Especially, the monthly average HFMD cases were 6741 (range, 66–37,640).

Figure 2 revealed that the HFMD was not significantly associated with monthly extreme wind velocity, monthly maximum wind velocity, and monthly mean relative humidity, but significantly with other meteorological factors (P < 0.05).

It is worth noting that strong correlations were found among monthly mean atmospheric pressure and monthly maximum atmospheric pressure, and five temperature indicators, between monthly mean atmospheric pressure and monthly maximum atmospheric pressure, and between monthly precipitation at 20–20 o’clock and monthly maximum daily precipitation. Finally, to avoid multicollinearity, monthly mean atmospheric pressure, monthly maximum atmospheric pressure, monthly mean temperature, monthly maximum temperature, monthly minimum temperature, monthly mean maximum temperature, and monthly maximum daily precipitation were excluded according to a threshold of absolute correlation coefficient being 0.9 (W. Liu et al. 2019). Monthly minimum atmospheric pressure, monthly mean temperature, monthly mean wind velocity in 2 min, monthly sunshine duration, monthly maximum daily precipitation, monthly mean water vapor pressure, and six meteorological factors were preserved and used for subsequent analysis.

As shown in Fig. 3, the monthly HFMD incidence showed no long-term trend. There was a bimodal seasonal distribution every year (Fig. 4), with one obvious peak between April and June, and another slight peak between October and December, which indicated that the original time series of HFMD was nonstationary.
Before the development of the ARIMA models, one order seasonal differencing and one ordinary differencing were performed to stabilize the time series of the train set (Fig. 5). As shown in Table 2, eight candidate ARIMA models were fitted based on the ACF and PACF (Fig. 6). The results of the Ljung-Box test illustrated that all the residual series of these models complied with white noise except ARIMA ((2,1,1),(2,1,1)) model ($P < 0.05$). According to the AIC, the most appropriate model was ARIMA (2,1,2)(1,1,1)$_{12}$ model, with a minimum AIC = 2251.02.

Then, six meteorological variables were retained and included in the ARIMA (2,1,2)(1,1,1)$_{12}$ model as external variables. And as shown in Fig. 7, further cross-correlation analysis suggested that monthly mean wind velocity in 2 min and monthly precipitation at 20–20 o’clock related to the HFMD at the lag of 1 month (WVm$_{lag}$) and the lag of 2 months (PRCP$_{lag}$), respectively. Therefore, these two lagged terms were also considered as external variables. Finally, the ARIMAX model was constructed with eight external variables.
Auto-ML model

During the feature engineering process, we created five kinds of new features for the Auto-ML algorithm. These features included series trend represented by a numeric index and a second polynomial of the index, seasonal component by the categorical variables of the months, series autocorrelation by seasonal and non-seasonal lag, remaining meteorological factors, and lagged terms of meteorological factors. It was particularly worth mentioning that, according to the information of ACF in Fig. 6, the incidence numbers of HFMD were found to be correlated with its seasonal lag (lag12) and non-seasonal lag (lag2), so two lagged terms of series were input into the model to provided serial correlation information. And, as already mentioned in the results on the ARIMAX model, the lagged terms of the monthly mean wind velocity in 2 min and monthly precipitation at 20–20 o’clock were also viewed as new features. Therefore, there are a total of thirteen new features.

After the two training processes were over, we selected the best model from the two types of Auto-ML models separately, namely Auto-ML1 which did not contain meteorological variables and only contained information about the time series of HFMD, and Auto-ML2 which contained both above. The variables importance plot (Fig. 8) with the ranking of the input variable’s contribution to the model performance using a scale between 0 and 1 showed the relative importance of each feature of the Auto-ML2 model. Ten of the variables included in the Auto-ML2 model had influences on the model, among which the categorical variable of the month was remarkably the most important, followed by the seasonal lag and one order lag of the monthly mean wind velocity in 2 min, which showed that the month had a large effect on the prevalence of HFMD, and in other words, the prevalence of HFMD had a seasonal pattern.

Prediction performance comparison

The above models were used to predict the monthly incidence numbers of HFMD in Henan province for 12 months between January 2019 and December 2019 and were validated using the test set. The prediction results were plotted in Fig. 9, from which it can be seen that the overall performance of the Auto-ML models was better than that of the ARIMA models, and the Auto-ML2 model can better predict the trend of the prevalence of HFMD.

The comparison results were summarized in Table 3. In terms of the ARIMA models, taking meteorological factors as external variables did not significantly improve the model’s fitting and prediction performance. Comparing the Auto-ML1 model and the ARIMA model, RMSE reduced by 1566.33 and MAE reduced by 827.61 respectively in the training set, and RMES reduced by 1870.54 and MAE reduced by 505.46 respectively in the test set. And comparing the Auto-ML2 model and the ARIMAX model, RMSE reduced by 2085.18 and MAE reduced by 1289.15 respectively in the training set, and RMES reduced by 2142.95 and MAE reduced by 1060.57 respectively in the test set. Therefore, the two Auto-ML models performed much better than both the ARIMA model
and ARIMAX model. And, for the Auto-ML1 model and AutosML2 model, the addition of meteorological factors reduced RMSE and MAE of the model by 716.72 and 838 separately in the training set, and 478.65 and 760.59 separately in the test set. In brief, the Auto-ML2 model had minimum RMSE and MAE in the model constructing phase and forecasting phase, which suggested that this model had the best prediction performance in estimating the trend and seasonal fluctuation of HFMD.

**Prediction**

The Auto-ML2 with excellent predictive performance was used to predict the monthly incidence numbers of HFMD in Henan province throughout 2020. The prediction results are presented in Fig. 9. We divided 2020 into the pre-COVID-19 outbreak and the post-COVID-19 outbreak. As shown in Fig. 9, before the COVID-19 outbreak, the predicted values followed the same trend as the reported
values of HFMD. The post-COVID-19 outbreak period was further divided into three stages, namely from the end of January to June (the first stage), from July to September (the second stage), and from October to December (the third stage). In the first stage, the predicted values rose rapidly and began to decline rapidly after reaching the peak in May, while the reported values were extremely small, with an average of only 103 cases per month. In the second stage, the trend of the predicted values and the reported values was opposite. The predicted values continued to decline and reached the bottom in September, while the reported values showed an upward trend and reached the peak in September. In the third stage, the predicted values first rose to the second peak of the year and then dropped, but the reported values kept falling. Overall, after the outbreak of COVID-19, the actual incidence numbers of HFMD have been greatly reduced than expected, and the incidence peak has also been delayed compared with previous years, which has led to significant changes in the seasonality of HFMD.
Table 2 AIC and Ljung-Box test result of the ARIMA models

| Model                  | AIC    | Ljung-box test | Q  | P  |
|------------------------|--------|----------------|----|----|
| Auto-ARIMA             | 2259.94| 18.281         | 0.6311|
| ARIMA(2,1,2)(1,1,1)_12 | 2251.02| 10.735         | 0.9053|
| ARIMA((2,1,2)(1,1,1)_12| 2258.76| 14.706         | 0.6820|
| ARIMA((2,1,1,1)(1,1,1)_12| 2259.77| 12.262         | 0.0065|
| ARIMA((2,1,0)(1,1,1)_12| 2258.98| 14.449         | 0.8070|
| ARIMA((2,1,0)(1,1,1)_12| 2257.33| 14.093         | 0.8257|
| ARIMA((0,1,2)(1,1,1)_12| 2255.98| 14.458         | 0.8065|
| ARIMA((0,1,2)(1,1,1)_12| 2256.78| 14.429         | 0.8081|

Q the statistics of the Ljung-Box test.

Discussion

HFMD, one of the most common infectious diseases, has prevailed globally for many years, and the Asia–Pacific region is most seriously affected by it. At present, the public health problems caused by HFMD have attracted extensive attention. Thus, an early accurate prediction of the epidemic trend and duration in advance is vital to preventing and controlling the epidemic of HFMD. In this study, we compared four models based on RMSE to select the optimal model, and eventually determined the optimized Auto-ML2 model with meteorological factors to predict the incidence numbers of HFMD in 2020.

We found that HFMD seemed to be a bimodal seasonal distribution every year in Henan province. The first obvious peak was from April to June, followed by a slight peak from October to December. However, different regions may show different seasonal patterns. For example, HFMD showed a bimodal change in southern China, peaking in May and October respectively, whereas it peaked in June in the north of the country, but a study on the seasonality of HFMD in China found two peaks annually (Tian et al. 2019; C. Xu et al. 2019). Furthermore, bimodal cycles of epidemics were reported in Japan, and one peak in Finland (Blomqvist et al. 2010; Sumi et al. 2017). This difference may be due to the complicated influencing factors of disease such as climate, geography, socioeconomic conditions, and public health interventions (Zhao and Hu 2019; Zhao and Li 2016). The transmission of HFMD has been proven to be climate sensitive (Yi et al. 2019). The possible mechanism is that meteorological factors may have impacts on the growth and survival of pathogens, the susceptibility and behaviors of humans, and the transmission environment, which further affects the prevalence of HFMD. However, the influences of meteorological factors on HFMD are inconsistent in different studies. Fu et al. indicated that monthly mean relative humidity, monthly mean temperature, and wind speed significantly influenced HFMD (Fu et al. 2019). Van et al. reported that the risk of HFMD was positively associated with higher average temperature, higher rainfall, and longer sunshine duration, and negatively associated with wind speed in Dak Lak province, Viet Nam (Van Pham et al. 2019). In our study, results indicated that monthly minimum atmospheric pressure, monthly mean temperature, monthly mean wind velocity in 2 min, monthly sunshine duration, monthly precipitation at 20–20 o’clock, and monthly mean water vapor pressure were significantly associated with HFMD. This discrepancy suggested that the interventions of HFMD should be adapted to local climate conditions.

To date, substantial studies have used the ARIMA model to predict the epidemic of HFMD, but the results varied across different studies. For example, by using the national monthly HFMD cases from May 2008 to August 2018, Tian et al. indicated that an annual periodicity and seasonal variation of HFMD incidence in China could be predicted well by a SARIMA (1,1,2)(0,1,1)_12 model (Tian et al. 2019). Liu et al. developed a univariate SARIMA model based on the monthly HFMD infection data in Sichuan province, China, and found that the seasonal ARIMA (1,0,1)(0,1,0)_12 model could be applied to forecast HFMD incidence (L. Liu et al. 2016). The other study reported that an ARIMA model with engine query and lagged temperature variables improved the prediction of HFMD (Du et al. 2017). In our results, ARIMA(2,1,2)(1,1,1)_12 with minimum AIC = 2251.02 is the most appropriate model among all ARIMA models, and then the ARIMAX model with six types of meteorological factors and two lagged meteorological factors was constructed based on it. However, according to the RMSE and MAE, the added external variables did not significantly improve the prediction performance of the model compared with the ARIMA model. The possible reasons for this result are as follows: the ARIMA model is essentially a linear method, whereas there is a nonlinear relationship between meteorological factors and HFMD incidence data, which causes the ARIMAX model could not capture this potential nonlinear information well. Consequently, neither of the above two models achieved a satisfied predictive performance. As mentioned earlier, the ARIMA model is a classic traditional infectious disease prediction method, so it was considered the benchmark model to judge and evaluate the performance of other prediction models. To further accurate prediction, the Auto-ML algorithm was applied to predict the incidence trend of HFMD. We constructed two Auto-ML models: one model which only mined the information contained in HFMD data, and the other model which simultaneously probed and explored the information of HFMD data and its relationship with meteorological factors. In terms of modeling only with the information of HFMD data, the Auto-ML1 model was superior to the ARIMA model. And in terms of modeling with the information of HFMD data and meteorological factors, the Auto-ML2 model also excelled over the ARIMAX model.
model. Therefore, both the Auto-ML models significantly outperformed the ARIMA model and the ARIMAX model, indicating that the Auto-ML model is more appropriate than the ARIMA model to forecast the incidence numbers of HFMD. Meanwhile, the Auto-ML2 model performed better than the Auto-ML1 model, which suggested that the addition of meteorological factors can further improve the prediction accuracy. Eventually, in this study, we consider the Auto-ML2 model as the optimal model among four models to predict the trend of HFMD in Henan province, China.

Therefore, we decided to use the Auto-ML2 model to further predict the incidence numbers of HFMD in 2020 to explore and analyze the change in HFMD epidemic pattern and the possible causes of the change after the outbreak of COVID-19.

Since the COVID-19 was firstly reported in December 2019, China has experienced multiple stages of epidemic prevention and control (Shi et al. 2020). After the first case of COVID-19 in Henan province was confirmed on January 21, 2020, the local government adopted a series...
of epidemic prevention and control measures, including advocating frequent hand washing and wearing masks, suspension of school, closure of entertainment venues, home quarantine, restriction of population mobility, and cancellation of assembly activities (Adhikari et al. 2020). At present, the overall epidemic situation in China has entered a stage of normalization. In this study, we used the time point at which COVID-19 appeared to divide 2020 into two periods, and the second period was further divided into three stages.

Before the outbreak of the epidemic in Henan province, we found that although the predicted values were slightly higher than the reported values, the model predicted the trend of HFMD well, which proved that the model still has good prediction performance in the longer-term prediction.

In the first stage after the outbreak of COVID-19, the actual monthly incidence numbers of HFMD were extremely small, while the predicted values increased month by month until the incidence peak and then slowly decreased. The research results of Niu et al. also exhibited that the transmission ability of HFMD dropped to 0 after the outbreak of COVID-19 (Niu et al. 2021). This phenomenon shows that the above prevention and control measures not only effectively restrained the transmission of COVID-19 but also suppressed the spread of HFMD. The involved explanations may be as follows. Firstly, the propaganda of epidemic prevention and control measures in COVID-19 made people improve their self-protection awareness and health awareness, especially hand hygiene and environmental sanitation, so the number of HFMD pathogens was reduced in the environment and the transmission route of HFMD was also blocked effectively. Secondly, the suspension of schools and the closure of entertainment venues reduced the contact opportunities between people to some extent, thus reducing the possibility of HFMD transmission. Thirdly, home quarantine also was an effective strategy to protect susceptible populations. Fourthly, during this stage, the number of people with mild symptoms who took the initiative to seek medical help may have reduced, which led to a decline in reported values. Finally, as the whole medical and health system was overloaded, the testing and surveillance of HFMD may be seriously disturbed. Hence, the reported values of HFMD were minimized.
In the second and third stages after the COVID-19 outbreak, the COVID-19 epidemic in Henan province has been well controlled, and the province has entered the normalization period. During these two stages, the reported HFMD cases had a rapid upward trend, and then a downswing after reaching a peak in September. However, the predicted values were quite different. There was a very large gap between the predicted values and the reported values. During this period, work and school have begun to resume nationwide, and daily life has gradually returned to the right track. Therefore, the public may relax their vigilance, and susceptible populations may have more opportunities to contact pathogens and populations infected with HFMD, so the increased possibility of infection caused a rebound in the incidence numbers of HFMD. Meanwhile, overall, the incidence peak was postponed due to COVID-19, and the seasonal pattern also changed. All of these suggest that the superposition of the countermeasures during the COVID-19 epidemic has some influence on reducing the transmission of HFMD.

There are also some limitations. First, the epidemic of HFMD is affected by complicated factors such as climate, geography, and socioeconomic conditions. In this study, just meteorological factors were considered in the model to improve the prediction accuracy. However, a sophisticated Auto-ML model with meteorological factors already has high predictive accuracy, and the inclusion of other related factors may cause overfitting. Therefore, in future research, whether to consider other factors remains to be explored. Second, similar to the traditional ML algorithms, the Auto-ML algorithms also exist in black boxes, so it is hard to interrupt the association between predictors and diseases. Nevertheless, it is still important for health administration departments to implement more effective prevention and control measures. Third, the models of this study are based on one province, so these specific models may not be generalizable to other regions with different climates. On the other hand, the Auto-ML algorithms, the ease of use and applicability of which in HFMD prediction have been proved, can be extended to other regions. Finally, we have no more

---

**Table 3** Comparison of four models

| Model                          | RMSE Training set | RMSE Test set | MAE Training set | MAE Test set |
|--------------------------------|-------------------|---------------|------------------|--------------|
| ARIMA (without meteorological factors) | 3707.45           | 4816.37       | 2118.81          | 2760.46      |
| ARIMAX (with meteorological factors)   | 3509.58           | 4250.78       | 2101.70          | 2554.98      |
| Auto-ML1 (without meteorological factors) | 2141.12           | 2945.83       | 1291.20          | 2255.00      |
| Auto-ML2 (with meteorological factors)   | 1424.40           | 2107.83       | 812.55           | 1494.41      |
direct evidence to assess the specific extent of the influence of countermeasures taken for COVID-19 on HFMD.

Conclusion

In our study, a total of four models were established to predict the incidence numbers of HFMD in Henan province, China. From the results of the model evaluation, both the ARIMA model and the ARIMAX model have not achieved satisfactory performance, and the introduction of meteorological variables did not significantly improve the model’s fitting and prediction accuracy. But in comparison, the Auto-ML models both have smaller RMSE and MAE than the ARIMA models in the model constructing phase and forecasting phases, which indicates that the Auto-ML algorithm performs excellently. The addition of meteorological factors and their lagged terms further improved the prediction accuracy of the Auto-ML model in both phase, which made the Auto-ML2 model have the minimum RMSE and MAE (training set: RMSE = 1424.40 and MAE = 812.55; test set: RMSE = 2107.83, MAE = 1494.41). Overall, the AutoML algorithm could base on the minimum error criterion to select a model which extracts available data information as much as possible and achieves satisfactory prediction performance. Therefore, this study proves that the AutoML algorithm is an applicable and ideal method to predict the epidemic trend of the HFMD. Furthermore, after analyzing the possible reasons for the difference between the reported value and predicted value of the incidence numbers of HFMD in 2020, it was found that during the period of COVID-19, the countermeasures of COVID-19 have a certain influence on suppressing the spread of HFMD. The findings can help health administrative departments in optimizing strategies and allocating resources to implement more effective HFMD prevention and control measures in the situation of the normalization of COVID-19.

Abbreviations

HFMD: Hand, foot, and mouth disease; AutoML: Automatic machine learning; ARIMA: Autoregressive integrated moving average; RMSE: Root mean square error; MAE: Mean absolute error; DALYs: Disability adjusted life years; COVID-19: Coronavirus disease 2019; ML: Machine learning; XGBoost: XExtreme Gradient Boosting; CISDPC: China Information System for Disease Control and Prevention; ARIMAX: Auto Regression and Moving Average model with eXogenous input; AR: Autoregressive process; I: Integrated process; MA: Moving average process; ACF: Autocorrelation coefficient; PACF: Partial autocorrelation coefficient; AIC: Akaike Information Criteria; DRF: Distributed random forest; DL: Deep learning; XRT: Extremely randomized trees; GLM: Generalized linear model with regularization

Author contribution

Chuan Yang: conceptualization, methodology, formal analysis, visualization; writing—original draft; Wei Wu: conceptualization, supervision, writing—reviewing and editing; Shuyi An, Baojun Qiao, and Peng Guan: data collection and curation. Desheng Huang, and Peng Guan: formal analysis, visualization.

Funding

This work was supported by the National Natural Science Foundation of China (grant numbers: 81202254 and 71974199) and the Science Foundation of Liaoning Provincial Department of Education (LJQK2021027).

Data availability

All data were from public databases, and the sources have been presented in the manuscript.

Declarations

Ethics approval and consent to participate Not applicable.

Consent for publication All of the authors agreed to publish the manuscript.

Competing interests The authors declare no competing interests.

References

Adhikari SP, Meng S, Wu Y-J, Mao Y-P, Ye R-X, Wang Q-Z et al (2020) Epidemiology, causes, clinical manifestation and diagnosis, prevention and control of coronavirus disease (COVID-19) during the early outbreak period: a scoping review. Infect Dis Poverty 9(1):29. https://doi.org/10.1186/s40249-020-00646-x

Asfahan S, Gopalakrishnan M, Dutt N, Niwas R, Chawla G, Agarwal M et al (2020) Using a simple open-source automated machine learning algorithm to forecast COVID-19 spread: a modelling study. Adv Respir Med 88(5):400–405. https://doi.org/10.5603/ARMa2020.0156

Bangura MS, Gonzalez MJ, Ali NM, Ren R, Qiao Y (2020) A collaborative effort of China in combating COVID-19. Glob Health Res Pol 5(1):47. https://doi.org/10.1186/s12556-020-00174-z

Blomqvist S, Klemola P, Kajalainen S, Paananen A, Simonen M-L, Vuorinen T et al (2010) Co-circulation of coxsackieviruses A6 and A10 in hand, foot and mouth disease outbreak in Finland. J Clin Virol : Off Publ Pan Am Soc Clin Virol 48(1):49–54. https://doi.org/10.1016/j.jcv.2010.02.002

Ceylan Z (2020) Estimation of COVID-19 prevalence in Italy, Spain, and France. Sci Total Environ 729:138817. https://doi.org/10.1016/j.scitotenv.2020.138817

Chen S, Liu X, Wu Y, Xu G, Zhang X, Mei S et al (2019) The application of meteorological data and search index data in improving the prediction of HFMD: a study of two cities in Guangdong Province, China, Sci Total Environment 652:1013–1021. https://doi.org/10.1016/j.scitotenv.2018.10.304

Cheng X, Hu J, Luo L, Zhao Z, Zhang N, Hannah MN et al (2021) Impact of interventions on the incidence of natural focal diseases during the outbreak of COVID-19 in Jiangsu Province. China Parasit Vectors 14(1):483. https://doi.org/10.1186/s13428-015-0611-2

Dean RT, Dunsuur WT (2016) Dangers and uses of cross-correlation in analyzing time series in perception, performance, movement, and neuroscience: the importance of constructing transfer function autoregressive models. Behav Res Methods 48(2):783–802. https://doi.org/10.3758/s13428-015-0611-2

Du Z, Xu L, Zhang W, Zhang D, Yu S, Hao Y (2017) Predicting the hand, foot, and mouth disease incidence using search engine query data and climate variables: an ecological study in Guangdong, China BMJ Open 7(10):e016263. https://doi.org/10.1136/bmjopen-2017-016263
of hand, foot, and mouth disease from 2000 to 2015 in Japan. Epidemiol Infect 145(14):2896–2911. https://doi.org/10.1017/S0950268817001820

Tian CW, Wang H, Luo XM (2019) Time-series modelling and forecasting of hand, foot and mouth disease cases in China from 2008 to 2018. Epidemiol Infect 147:e82. https://doi.org/10.1017/S095026881800362X

Van Pham H, Phan UTN, Pham ANQ (2019) Meteorological factors associated with hand, foot and mouth disease in a Central Highlands province in Viet Nam: an ecological study. West Pac Surveill Response J 10(4):18–23. https://doi.org/10.5365/wpsar.2017.8.1.003

Van Tu P, Thao NTT, Perera D, Truong KH, Tien NTK, Thuong TC et al (2007) Epidemiological and virological investigation of hand, foot, and mouth disease, southern Vietnam, 2005. Emerg Infect Dis 13(11):1733–1741. https://doi.org/10.3201/eid1311.070632

Wang Y, Feng Z, Yang Y, Seif S, Gao Y, Longini IM et al (2011) Hand, Foot, and mouth disease in China patterns of spread and transmissibility. Epidemiology 22(6):781–792. https://doi.org/10.1097/EDM.0b013e1135b65f79

Wang Y-W, Shen Z-Z, Jiang Y (2018) Comparison of ARIMA and GM(1,1) models for prediction of hepatitis B in China. PLoS ONE 13(9):e0201987. https://doi.org/10.1371/journal.pone.0201987

Wang Y, Xu C, Zhang S, Yang L, Wang Z, Zhu Y et al (2019) Development and evaluation of a deep learning approach for modeling seasonality and trends in hand-foot-mouth disease incidence in mainland China. Sci Rep 9(1):8046. https://doi.org/10.1038/s41598-019-44469-9

Wang C, Horby PW, Hayden FG, Gao GF (2020) A novel coronavirus outbreak of global health concern. Lancet (London, England) 395(10223):470–473. https://doi.org/10.1016/S0140-6736(20)30185-9

Wu D, Lu J, Cao L, Ma X, Liu Q, Liu Y et al (2020) Positive effects of COVID-19 control measures on pneumonia prevention. Int J Infect Dis : IJID : Off Publ Int Soc Infect Dis 96:548–549. https://doi.org/10.1016/j.ijid.2020.05.069

Xing W, Liao Q, Viboud C, Zhang J, Sun J, Wu JT et al (2014) Hand, foot, and mouth disease in China, 2008–12: an epidemiological study. Lancet Infect Dis 14(4):308–318. https://doi.org/10.1016/S1473-3099(13)70342-6

Xu C, Zhang X, Xiao G (2019) Spatiotemporal decomposition and risk determinants of hand, foot and mouth disease in Henan, China. Sci Total Environ 657:509–516. https://doi.org/10.1016/j.scitotenv.2018.12.039

Xu J, Yang M, Zhao Z, Wang M, Guo Z, Zhu Y et al (2020) Meteorological factors and the transmissibility of hand, foot, and mouth disease in Xiamen City, China Front Med (lausanne) 7:597375. https://doi.org/10.3389/fmed.2020.597375

Yi L, Xu X, Ge W, Xue H, Li J, Li D et al (2019) The impact of climate variability on infectious disease transmission in China: current knowledge and further directions. Environ Res 173:255–261. https://doi.org/10.1016/j.envres.2019.03.043

Zhang Y, Zhu Z, Yang W, Ren J, Tan X, Wang Y et al (2010) An emerging recombinant human enterovirus 71 responsible for the 2008 outbreak of hand-foot and mouth disease in Fuyang city of China. Virol J 7:94. https://doi.org/10.1186/1743-422X-7-94

Zhang X, Hou F, Qiao Z, Li X, Zhou L, Liu Y et al (2016) Temporal and long-term trend analysis of class C notifiable diseases in China from 2009 to 2014. BMJ Open 6(10):e011038. https://doi.org/10.1136/bmjopen-2016-011038

Zhao J, Hu X (2019) The complex transmission seasonality of hand, foot, and mouth disease and its driving factors. BMC Infect Dis 19(1):521. https://doi.org/10.1186/s12879-019-4153-6

Zhao J, Li X (2016) Determinants of the transmission variation of hand, foot and mouth disease in China. PLoS ONE 11(10):e0163789. https://doi.org/10.1371/journal.pone.0163789
Zhao J, Jiang F, Zhong L, Sun J, Ding J (2016) Age patterns and transmission characteristics of hand, foot and mouth disease in China. BMC Infect Dis 16:691. https://doi.org/10.1186/s12879-016-2008-y
Zhong R, Wu Y, Cai Y, Wang R, Zheng J, Lin D et al (2018) Forecasting hand, foot, and mouth disease in Shenzhen based on daily level clinical data and multiple environmental factors. Biosci Trends 12(5):450–455. https://doi.org/10.5582/bst.2018.01126
Zou JJ, Jiang GF, Xie XX, Huang J, Yang XB (2019) Application of a combined model with seasonal autoregressive integrated moving average and support vector regression in forecasting hand-foot-mouth disease incidence in Wuhan. China Med (Baltimore) 98(6):e14195. https://doi.org/10.1097/MD.00000000000014195

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.