A data-driven interpretable ensemble framework based on tree models for forecasting the occurrence of COVID-19 in the USA

Hu-Li Zheng1 · Shu-Yi An2 · Bao-Jun Qiao2 · Peng Guan1 · De-Sheng Huang3 · Wei Wu1

Received: 12 May 2022 / Accepted: 16 September 2022 / Published online: 22 September 2022
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

Abstract
This prevalence of coronavirus disease 2019 (COVID-19) has become one of the most serious public health crises. Tree-based machine learning methods, with the advantages of high efficiency, and strong interpretability, have been widely used in predicting diseases. A data-driven interpretable ensemble framework based on tree models was designed to forecast daily new cases of COVID-19 in the USA and to determine the important factors related to COVID-19. Based on a hyperparameter optimization technique, we developed three machine learning algorithms based on decision trees, including random forest (RF), eXtreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM), and three linear ensemble models were used to integrate these outcomes for better prediction accuracy. Finally, the SHapley Additive explanation (SHAP) value was used to obtain the feature importance ranking. Our outcomes demonstrated that, among the three basic machine learners, the prediction accuracy was the following in descending order: LightGBM, XGBoost, and RF. The optimized LAD ensemble was the most precise prediction model that reduced the prediction error of the best base learner (LightGBM) by approximately 3.111%, while vaccination, wearing masks, less mobility, and government interventions had positive effects on the control and prevention of COVID-19.

Keywords COVID-19 · Ensemble · Hyperopt · XGBoost · LightGBM · RF

Introduction
The novel virus that can cause severe acute respiratory disease (COVID-19) has become one of the most serious public health crises. Furthermore, the current COVID-19 outbreak remains a global pandemic. Globally, as of 5:40 p.m. Central European time, November 2, 2021, there have been 240 million cumulative confirmed cases of COVID-19, including 5 million deaths, according to the World Health Organization (https://covid19.who.int/). This new epidemic has attracted global attention and has become one of the most serious public health crises. The common symptoms of COVID-19 are coughing, fever, fatigue, anosmia, headache, rhinorrhea, and myalgia, and SARS-CoV-2 infection is believed to be transmitted through aerosols or droplets (Guan et al. 2020; Mao et al. 2020). Since the epidemic was proclaimed a pandemic, types of precautions have been taken to control the epidemic’s spread, covering the prevention and control of the disease during public transportation, distancing policies, population-wide movement control, wearing of masks, and vaccinations (Ng et al. 2020; Shen et al. 2020). Due to the measures undertaken, the daily confirmed cases in China...
have decreased drastically, but globally, the virus has not yet been completely stopped owing to its high infectious power and strong pathogenicity. Since the outbreak, the USA has always been one of areas most affected by the epidemic. Therefore, we decided to predict the trend of the epidemic in the USA and offered relevant prevention suggestions.

Various models have been used to predict COVID-19 transmission. Based on the traditional susceptible–infected–recovered compartment model, many novel dynamical models (Abbasi et al. 2020; Campillo-Funollet et al. 2021; Sun et al. 2020) have been proposed that consider many factors, such as deaths, daily admissions, discharges, and quarantine. There have also been time series models that have been used to predict COVID-19, such as the autoregressive integrated moving average method (Asfahan et al. 2020) and its variants (Ahmar and Del Val 2020; ArunKumar et al. 2021). However, the above methods cannot consider a broad range of factors affecting the development of the epidemic and cannot address the changing environment.

More importantly, with the continuous upgrading of computer information and software technology, artificial intelligence is progressively becoming widely used in medical systems to detect diseases and make clinical diagnoses. Machine learning, including deep learning, is supposed to be an indispensable part of artificial intelligence (Yu et al. 2021), which has also been widely used in predicting COVID-19. A spatial–temporal analysis framework was developed by combining random forest (RF) regression and a multiobjective optimization algorithm to predict the daily cases and death rate in Asia (Pan et al. 2021). The prediction effects of support vector regression and stacking-ensemble learning were better than those of comparison models in Brazil (Ribeiro et al. 2020). The Prophet algorithm was also considered to have reliable prediction ability in South Korea (Asfahan et al. 2020). Recently, special attention has been paid to deep learning methods, because of their excellent universality and superior nonlinear approximation in time series analysis. Abdelkader et al. confirmed that hybrid convolutional neural network–long short-term memory and hybrid gated recurrent unit–convolutional neural networks could efficiently predict COVID-19 cases (Dairi et al. 2021). Another study also showed that the bidirectional long short-term memory method could be used for pandemic prediction and better planning and management (Shahid et al. 2020). However, it is worth noting that the interpretability of the deep learning model has not been strong due to its black box problem, and these related studies could not well analyze the correlations between the influencing factors and the diseases.

Previous studies that only applied traditional epidemic models were subject to underfitting or overfitting problems and had poor generalization ability. Compared with other models, machine learning models have the advantages of approximation excellent universality, superior nonlinear approximation, interpretability, and not easy overfitting in time series analysis, and can analyze a large number of features simultaneously. Therefore, we believe that the machine learning model is most suitable for predicting COVID-19 trends and analyzing influencing factors at the same time. Before us, RF was used in research on COVID-19 and other various diseases (Sarica et al. 2017; Wu et al. 2021; Yang et al. 2020). The eXtreme Gradient Boosting (XGBoost) method was used to predict the mortality of COVID-19 in Wuhan, China (K. Wang et al. 2020). XGBoost has also been used in other diseases for disease prediction and risk factor analysis, such as smoking-induced noncommunicable disease (Davagdorj et al. 2020) and kidney disease (Chen et al. 2019). The Light Gradient Boosting Machine (LightGBM) model showed better discrimination ability than the traditional model in predicting the all-cause mortality of patients (Zheng et al. 2021a, b), but it has not been used in COVID-19 prediction. To date, none of these three methods has been used to forecast daily cases of COVID-19 in the USA, which is one of the innovations in our research.

Furthermore, considering one situation in which a single machine learner is inferior to an ensemble model that can reduce deviation and improve robustness (L. Wang et al. 2021; Ye et al. 2021), in this study, we used three linear ensemble methods — simple averaging (SA), ordinary least square (OLS), and least absolute deviation (LAD) ensembles — to integrate three tree-based models, including RF, XGBoost, and LightGBM, to predict the prevalence of COVID-19 in the USA and analyze its influencing factors for COVID-19 prevention and control.

**Methods**

**Framework of ensemble methods**

To achieve the two objectives of prediction and prevention, three main steps were implemented, including data preparation, prediction of single machine learning models, and ensemble methods as shown in Fig. 1. First, all of our data come from public datasets. We have tried our best to gather more comprehensive data for ensemble machine learning models. In addition to the number of daily new cases, other data were divided into four categories, namely, personal protection, social policy indicators, community mobility and time indices. Second, based on the hyperparametric
optimization technique (Hyperopt) to tune the parameters automatically, we developed three machine learning algorithms based on decision trees, including RF, XGBoost, and LightGBM, to forecast daily new cases. Finally, three linear ensemble methods — SA, OLS, and LAD ensembles — were adopted to repredict the daily new cases by combining the results of the three basic models for better prediction accuracy and robustness. At the same time, for the sake of interpretability, we determined the impact of the included variables on the outcomes using SHapley Additive explanation (SHAP) values, to identify the important factors for COVID-19 transmission.

Data collection and preprocessing

The data were collected from the following four public data sources: (1) cases of COVID-19 and the number of vaccinations in the USA were obtained from the official website of the Centers for Disease Control and Prevention of the USA (https://covid.cdc.gov); (2) the usage rate of masks was collected from the Institute for Health Metrics and Evaluation (https://covid19.healthdata.org/united-states-of-america?view=mask-use&tab=trend), which could reflect people’s awareness of self-protection; (3) social policy indicators were obtained via the Oxford Covid Government Response Tracker (Hale et al. 2021), which could quantify the extent of government responses; (4) the travel popularity data were obtained through Google Community Mobility Reports (https://www.google.com/covid19/mobility/), reflecting the movement trend of citizens over time through geographical location.

In addition to the above information sources, we added week and festival information, time trend items and lags as input variables. There were no missing data in this study.

In the early stage of the epidemic, the number of cases was small and unstable, so our research dates were from 1 April 2020 to 31 August 2021. We deleted some variables that did not change significantly during the study period and some repeated variables. Considering the seasonality, autocorrelation and partial autocorrelation, we included the trend item and 1, 2, and 7 time-lagged variables as the input features. Finally, we took the daily new cases as the outcome variable and included a total of 35 input features, as shown in Table 1, and preliminarily explored the correlation between the target and input features through Spearman’s correlation, as shown in Fig. 2. Although some features had a statistically weak association with daily cases, they were still included to ensure the integrity of the features and avoid the meaningless features in reality from being ignored. More importantly, that weakly correlated features were retained hardly affected the prediction results of the three machine learners. These models based on decision trees in our research could assign appropriate weights to each feature by self-learning using the training set data.

Data analysis of three machine learning models, including RF, XGBoost, and LightGBM, was conducted using Python software, version 3.8.8. We adopted sklearn.metrics, sklearn.model_selection, and matplotlib.pyplot modules in Python and some main Python packages, including shap, hyperopt, RandomForestRegressor, xgboost, and lightgbm. Other data analysis of the three linear ensemble methods was conducted by using R software, version 4.0.5. The ForecastComb, forecast, ggplot2, graphics, and tseries packages were used. Methods were performed in accordance with relevant guidelines and regulations.

Random forest (RF)

Random forest (RF) based on bagging integration is one of the most common and powerful supervised learning algorithms that can solve regression and classification problems (Breiman 2001). Its technique is to create multiple samples from the same set of data, readjust them through bootstrap technology, and randomly select predictors to form each node of the decision tree. The randomness of time series can also be well handled by RF (Casiraghi et al. 2020). The random forest model can be described as Eq. (1), where P represents the number of decision trees.

$$Y = \frac{1}{P} \sum_{i=1}^{P} f_i(x)$$

(1)

LightGBM and XGBoost models

Gradient boosting is a tree-based machine learning ensemble method (Kim et al. 2021) that can improve the accuracy and robustness of overall training and prediction by integrating multiple weak learners. In this study, we utilized two relatively advanced and quick gradient-lifting algorithms: XGBoost and LightGBM. The most important feature of XGBoost is that it can automatically use the multithreading of the CPU in parallel and improve the algorithm to improve the accuracy. The XGBoost algorithm can be summarized by Eq. (2), where l denotes the loss function, f denotes a weak learner, and Ω denotes the regularization term (Nishio et al. 2018).

$$L^f = \sum_{i=1}^{n} l\left( y_i, y_i^{pred} + f_i(x_i) \right) + \Omega(f)$$

(2)

LightGBM is a decision tree algorithm based on histogram that has two novel techniques to improve performance and reduce computing time: Gradient-based One-Side Sampling and Exclusive Feature Bundling (Ke et al. 2017). The first technique retains instances with large gradients, and only randomly omits instances with small gradients to retain the accuracy of the information gain estimation. The second makes it possible
to design a nearly lossless method to reduce the number of features in sparse high-dimensional data (Yu 2019).

Hyperopt (a hyperparametric optimization technique)

Hyperopt is a distributed asynchronous hyperparameter optimization method (Bergstra et al. 2013) based on Bayesian optimization, which has been used in the parameter optimization of end products (Shahriari et al. 2015), such as recommendation systems, medical analysis tools, and speech recognizers. Grid searching is often used for this purpose. However, when the number of parameters increases, grid searches are not feasible due to the large amount of calculation, making it less efficient than Hyperopt.

The Tree of Parzen Estimators (TPE) was selected as the search algorithm. Therefore, we sought the best parameter combinations of three machine learners by this optimization method, which was an innovation of our research. Combined with tenfold cross validation, we took the mean absolute percentage error (MAPE) as the objective function, which was the goal that we wanted to optimize in the parameter space that we defined. The following parameter spaces were used for parameter optimization.

i. For random forest, the parameters and their ranges were as follows: n_estimators, 160–190; and max_depth,

| Category                  | Feature code | Feature Code | Feature                        |
|---------------------------|--------------|--------------|--------------------------------|
| Self-protection           | ×1           | Mask (%)     |                                |
|                           | ×2           | People receiving 1 or more doses cumulative |                                |
|                           | ×3           | People fully vaccinated cumulative |                                |
| Social policy indicators  | ×4           | School closing |                                |
|                           | ×5           | Workplace closing |                                |
|                           | ×6           | Cancel public events |                                |
|                           | ×7           | Restrictions on gatherings |                                |
|                           | ×8           | Close public transport |                                |
|                           | ×9           | Requirements to stay at home |                                |
|                           | ×10          | Restrictions on internal movement |                                |
|                           | ×11          | International travel controls |                                |
|                           | ×12          | Income support |                                |
|                           | ×13          | Debt/contract relief |                                |
|                           | ×14          | Public information campaigns |                                |
|                           | ×15          | Testing policy |                                |
|                           | ×16          | Contact tracing |                                |
|                           | ×17          | Protection of elderly people |                                |
|                           | ×18          | Government response index |                                |
|                           | ×19          | Containment health index |                                |
|                           | ×20          | Economic support index |                                |
| Community mobility        | ×21          | Retail and recreation change from baseline (%) |                                |
|                           | ×22          | Grocery and pharmacy percent change from baseline (%) |                                |
|                           | ×23          | Parks percent change from baseline (%) |                                |
|                           | ×24          | Transit stations percent change from baseline (%) |                                |
|                           | ×25          | Workplaces percent change from baseline (%) |                                |
|                           | ×26          | Residential percent change from baseline (%) |                                |
| Time index                | ×27          | Holiday |                                |
|                           | ×28          | Previous a day is holiday |                                |
|                           | ×29          | Previous 2 days is holiday |                                |
|                           | ×30          | Previous 3 days is holiday |                                |
|                           | ×31          | Day of week |                                |
|                           | ×32          | Trend |                                |
|                           | ×33          | Lag1 |                                |
|                           | ×34          | Lag2 |                                |
|                           | ×35          | Lag7 |                                |
8–15. More parameter explanations and introductions of random forest parameters are available on the following website: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html.

ii. For XGBoost, the parameters and their ranges were as follows: n_estimators, 130–190; learning_rate, 0.070–0.095; max_depth, 5–9; min_child_weight, 8–16; subsample, 0.85–0.95; colsample_bytree, 0.8–0.9; reg_lambda, 1–8; and alpha, 10–20. More parameter explanations and introductions of XGBoost parameters are available on the following website: https://xgboost.ai.

iii. For LightGBM, the parameters and their ranges were as follows: n_estimators, 110–140; learning_rate, 0.09–0.1; max_bin, 130–150; max_depth, 4–6; num_leaves, 8–12; bagging_freq, 30–40; bagging_fraction, 0.85–0.95; feature_fraction, 0.85–0.95; lambda_1, 170–200; and lambda_2, 0.0000035–0.0000045. More parameter explanations and introductions of LightGBM parameters can be available on the following website: https://lightgbm.readthedocs.io/en/latest/.

**Ensemble methods**

In this study, the three ensemble models used the outputs of the three basic models as input variables for secondary prediction, which could increase the forecast accuracy and robustness. Obviously, the SA method gives each instance the same weight. The principle of OLS is that the sum of the squares of the errors between the estimated value and the actual value is the smallest. The weight of the OLS ensemble is generally learned from the training data, but the weighted average method is not necessarily better than the simple average method (Shahhosseini et al. 2020), especially for large-scale integration and situations in which the performance of individual learners is similar. The LAD ensemble computes forecast combination weights using the principle of minimum absolute deviation. One characteristic of LAD is that it does not minimize the squared error loss as OLS and constrained least squares, but the absolute values of the errors. We hoped to improve the accuracy of prediction by considering the outputs of the three machine learning methods as the inputs of the three ensembles.

**Model evaluation**

In our study, three accuracy metrics were applied to evaluate the performance of the models: mean absolute error (MAE), root mean square error (RMSE), and mean absolute percentage error (MAPE) as follows:

\[
MAE = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - y_i| \tag{3}
\]

\[
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2} \tag{4}
\]
where \( y_i \) denotes the observed values, \( \hat{y}_i \) is the prediction, and \( n \) denotes the number of data points.

MAE is the average of absolute prediction error and represents the arithmetic mean of the absolute error between the predicted value and the actual value. RMSE is the square root of the average squared deviation of predictions from real values. MAPE quantifies the accuracy as a percentage, which can be calculated as the cumulative absolute percentage error of each time frame.

**Feature importance measurement**

Understanding data from machine learning models was also one of our research goals. Our chosen models — RF, XGBoost, and LightGBM — all have natural methods to quantify the importance of input features. However, for the interpretability of ensembles and the consistency of all models, Tree Explainer (Lundberg et al. 2020), an explanation method for trees, was adopted to measure and rank the feature importance. This method could easily calculate the optimal local interpretation according to the expected properties in game theory. Using the SHapley Additive explanation (SHAP) values of the whole dataset, the ability of local interpretation could be calculated efficiently and accurately, and a series of tools could be developed to explain the global behavior of models and to directly capture feature interactions (Lundberg et al. 2020). We calculated SHAP values for all variables for three basic machine learners and then combined these importance values. More principles and calculation details can be seen in this study (Mangalathu et al. 2020).

**Results and discussions**

**Characteristics of cases of COVID-19 in the USA**

This study focused on the number of daily new cases in the USA from April 1, 2020, to August 31, 2021. First, we decomposed the data. The data on daily cases, seasonality, trends and remainders are displayed from top to bottom in Fig. 3. In the figure, the development of daily new cases in the USA is fully shown. There is also a seasonal pattern and a trend in our data. Second, to understand the seasonality of the data more clearly, we drew a seasonal subseries plot (Fig. 4). The horizontal lines indicate the means for one day of all weeks. This figure clearly depicts the underlying periodicity and shows the regular pattern in a cycle. In a week, the numbers of cases on Monday, Tuesday and Wednesday are higher than those on other days.
Prediction effects of all models

Our entire time series data include the daily new cases as the outcome variable and 35 input variables. The daily data in the USA from 1 April 2020 to 31 August 2021 were spilt into two parts: a training set (from 1 April 2020 to 31 July 2021) to construct three basic models: RF, XGBoost, and LightGBM; and a test set (from August 1 to 31, 2021) to validate the predictive performance of each model. Then, three ensemble methods were used to integrate the results of the three basic models. Figures 5 and 6 illustrate the relationship between real COVID-19 cases and predicted values achieved by three basic learners and three ensembles, respectively. It can be seen from the figures that the prediction effect of the ensemble methods is better than those of the single methods. In addition, we have found that the LightGBM and ensemble models perform better at the data inflection points in Figs. 5 and 6, showing that these models could have good prediction ability for complex and changeable data and situations.

Performance measures for all models

We set two identical parameters for the three basic machine learning models to render them comparable: tenfold cross validation and seed number 2021. The details of the model evaluation criteria are shown in Table 2. Obviously, among the three basic machine learning methods, considering all of the criteria, the accuracy of prediction is as follows in descending order: LightGBM, XGBoost, and RF. LightGBM could more accurately forecast the COVID-19 trend in the USA. The SA ensemble could not greatly improve the prediction accuracy. Compared with the base learners, the remaining two ensembles provided better accuracy. From Table 2, the optimized LAD ensemble is the most precise prediction model, with an MAE of 8540.411, reducing the prediction error of the best base learner (LightGBM) by approximately 3.111%. Moreover, other metrics are lower than those of basic learners. This outcome might occur because there are many categorical variables in our data and LightGBM can offer good accuracy with integer-encoded categorical features. Moreover, its leafwise algorithms tend to achieve smaller losses than level-wise algorithms, such as XGBoost. Interestingly, the optimized LAD ensemble could further improve the prediction accuracy for the COVID outbreak in the USA. This improvement could be explained by the LAD method being able resist outliers in the data, while the OLS method gives more weight to outliers.

Feature importance

Based on the SHAP-based method, we obtained the feature SHAP values of the three basic machine learning methods and then determined the feature importance according to the weight achieved by the optimal ensemble method, i.e., LAD ensemble. These artificially controllable external factors are the focus of our attention, as shown in Fig. 6. The mean SHAP absolute value indicates the average impact on the model output magnitude. Figure 6 presents the top...
10 most important features for the outcomes of the RF, XGBoost, LightGBM, and LAD ensemble models. Community mobility features account for a large proportion of the top 10 features. The four subgraphs show that the workplace and residential percent changes from baseline have been ranked first and second. There are five community mobility indicators in Fig. 7d. Social policy indicators, such as restrictions on gatherings, government response index, and requirements to stay at home, also appear frequently in Fig. 7. The feature, i.e., the cumulative number of people receiving 1 dose or more, also ranks high. Wearing a mask is also an important feature. The currently developed vaccine has remained effective for moderate and severe COVID-19, even in the face of virus variants (Thiruvengadam et al. 2021). The effectiveness of social isolation and face covering in epidemic control has also been confirmed in other countries (Trauer et al. 2021).

The great innovation of our study was that a large number of input variables, such as self-protection, social policy, community mobility, and time index, were entered into the models. The study was the first to use LightGBM to predict the epidemic situation in the USA and to use a hyperparametric optimization technique (Hyperopt) to turn parameters. The use of ensemble methods is another of our strengths, and ensemble methods could improve prediction accuracy. The SHapley Additive explanation (SHAP) value was used to improve the interpretability of the model.

However, in our study, there were still many limitations that must to be improved upon by future studies. This research was based on the USA, a whole country with a very large geographical area and various weather environments and landforms. Thus, it is difficult for us to find a suitable meteorological index or air pollution index to exactly describe the characteristics of this geographical environment, although meteorological factors...
and air quality can affect the spread of COVID-19 (Copat et al. 2020; Zheng et al. 2021a, b). In fact, the occurrence of COVID-19 should be impacted by the spatiotemporal variations especially in large areas. Spatial factors were not considered in this study. Future research could conduct make prediction analysis over a larger range and include more environmental and spatiotemporal variables. The number of cases worldwide is still increasing rapidly. New variants of coronavirus have been found in many countries. In the USA, the delta variant has increased the risk of hospitalization and death (Bast et al. 2021). To resist COVID-19, it is crucial to formulate a clear reporting policy for potential global health emergencies (Chams et al. 2020), also making it possible for the world to jointly build a global COVID-19 database, including virus variants, government policies, population mobility and other relevant data. Then, a larger and more comprehensive dataset can be created to better serve the predictive model and government decision-making, aiming to penetrate COVID-19 evolution in more countries. On the foundation of big data, future research could build a more comprehensive and practical prediction model from the aspects of space–time geography, medical resources, economic support and feature interaction.

### Conclusions

In this study, data related to the COVID-19 epidemic were collected as much as possible, and a total of 35 variables were entered into the machine learning models. Based on the authenticity and validity of our data source, we confirmed...
that, among the three basic machine learning models, LightGBM had the best prediction performance. Moreover, the ensemble models, especially the LAD ensemble, could further improve the prediction accuracy. At the same time, the results of importance ranking illustrated that vaccination, wearing a mask, less mobility, and appropriate government intervention measures could effectively slow the incidence rate, providing a professional basis for the government to formulate relevant policies on the prevention of and response to COVID-19. Our models can be applied to many other countries in which similar data are available.

Author contribution The conceptualization was performed from by ZHL and WW. ZHL designed and drafted the manuscript. ASY, QBJ, and GP participated in the data collection. ZHL, HDS, and GP participated in the data analysis. ASY and WW critically revised the manuscript.

Funding This study 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 (LJKQZ2021027).

Data availability All data came 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

Abbasi Z, Zamani I, Mehr AHA, Shafeirad M, Ibeas A (2020) Optimal control design of impulsive SIRSIAR epidemic models with application to COVID-19. Chaos, Solitons Fractals 139:110054–110054. https://doi.org/10.1016/j.chaos.2020.110054

Ahmar AS, Del Val EB (2020) SutteARIMA: Short-term forecasting method, a case: Covid-19 and stock market in Spain. Sci Total Environ 729:138883. https://doi.org/10.1016/j.scitotenv.2020.138883

ArunKumar KE, Kalaga DV, Sai Kumar CM, Chilkoor G, Kawaji M, Brenza TM (2021) Forecasting the dynamics of cumulative COVID-19 cases (confirmed, recovered and deaths) for top-16 countries using statistical machine learning models: auto-regressive integrated moving average (ARIMA) and seasonal auto-regressive integrated moving average (SARIMA). Appl Soft Comput 103:107161–107161. https://doi.org/10.1016/j.asoc.2021.107161

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/ARM.a2020.0156

Bast E, Tang F, Dahn J, Palacio A (2021) Increased risk of hospitalisation and death with the delta variant in the USA. Lancet Infect Dis 21(12):1629–1630. https://doi.org/10.1016/S1473-3099(21)00685-X

Bergstra J, Yamins D, Cox DD (2013) Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures. Paper presented at the Proceedings of the 30th International Conference on International Conference on Machine Learning

Breiman L (2001) Random forests. Mach Learn 45(1):5–32. https://doi.org/10.1023/A:101093404324

Campillo-Funollet E, Van Yperen J, Allman P, Bell M, Beresford W, Clay J et al (2021) Predicting and forecasting the impact of local outbreaks of COVID-19: use of SEIR-D quantitative epidemiological modelling for healthcare demand and capacity. Int J Epidemiol 50(4):1103–1113. https://doi.org/10.1093/ije/dyab106

Castrighi E, Malchiodi D, Trucco G, Frasca M, Cappelletti L, Fontana T et al (2020) Explainable machine learning for early assessment of COVID-19 risk prediction in emergency departments. IEEE Access 8:196299–196325. https://doi.org/10.1109/access.2020.3034032

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

Chams N, Chams S, Badran R, Shams A, Araji A, Raad M et al (2020) COVID-19: a multidisciplinary review. Front Public Health 8:383. https://doi.org/10.3389/fpubh.2020.00383

Chen T, Li X, Li Y, Xia E, Qin Y, Liang S et al (2019) Prediction and risk stratification of kidney outcomes in IgA nephropathy. Am J Kidney Dis 74(3):300–309. https://doi.org/10.1053/j.ajkd.2019.02.016

Copat C, Cristaldi A, Fiore M, Grasso A, Zuccarello P, Signorelli SS et al (2020) The role of air pollution (PM and NO(2)) in COVID-19 spread and lethality: a systematic review. Environ Res 191:110129. https://doi.org/10.1016/j.envres.2020.110129

Dairi A, Harrou F, Zeroulou A, Hittawe MM, Sun Y (2021) Comparative study of machine learning methods for COVID-19 transmission forecasting. J Biomed Inform 118:103791–103791. https://doi.org/10.1016/j.jbi.2021.103791

Davhdorj K, Pham VH, Theria-Umpom N, Ryu KH (2020) XGBoost-based framework for smoking-induced noncommunicable disease prediction. International journal of environmental research and public health, 17(18). https://doi.org/10.3390/ijerph17186513.

Guo W, Ji X, Wu CQ, He JX et al (2020) Clinical characteristics of coronavirus disease 2019 in China. N Engl J Med 382(18):1708–1720. https://doi.org/10.1056/NEJMoa2002302

Hale T, Angrist N, Goldszmidt R, Kira B, Petherick A, Phillips T, Webster S, et al (2021) A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). Nat Human Behav. Retrieved from. https://doi.org/10.1038/s41562-021-01079-8

Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W et al (2017) LightGBM: a highly efficient gradient boosting decision tree. Paper presented at the Proceedings of the 31st International Conference on Neural Information Processing Systems

Kim BW, Choi MC, Kim MK, Lee J-W, Kim MT, Noh JJ, et al (2021) Machine learning for recurrence prediction of gynecologic cancers using Lynch syndrome-related screening markers. Cancers 13(22). https://doi.org/10.3390/cancers13225670.

Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B et al (2020) From local explanations to global understanding with SHapley Additive exPlanations (SHAP) approach. Eng Struct 219:110927. https://doi.org/10.1016/j.engstruct.2020.110927
