JAMIA Open, 5(3), 2022, 1–11
https://doi.org/10.1093/jamiaopen/ooac056
Research and Applications

Research and Applications

Predicting COVID-19 county-level case number trend by combining demographic characteristics and social distancing policies

Megan Mun Li[1], Anh Pham[2], and Tsung-Ting Kuo[2]

[1]Department of Biology, University of California San Diego, La Jolla, California, USA, [2]UCSD Health Department of Biomedical Informatics, University of California San Diego, La Jolla, California, USA

Corresponding Author: Tsung-Ting Kuo, PhD, UCSD Health Department of Biomedical Informatics, University of California San Diego, 9500 Gilman Dr, La Jolla, CA, USA; tskuo@health.ucsd.edu

Received 20 March 2022; Revised 9 June 2022; Editorial Decision 22 June 2022; Accepted 23 June 2022

ABSTRACT

Objective: Predicting daily trends in the Coronavirus Disease 2019 (COVID-19) case number is important to support individual decisions in taking preventative measures. This study aims to use COVID-19 case number history, demographic characteristics, and social distancing policies both independently/interdependently to predict the daily trend in the rise or fall of county-level cases.
Materials and Methods: We extracted 2093 features (5 from the US COVID-19 case number history, 1824 from the demographic characteristics independently/interdependently, and 264 from the social distancing policies independently/interdependently) for 3142 US counties. Using the top selected 200 features, we built 4 machine learning models: Logistic Regression, Naïve Bayes, Multi-Layer Perceptron, and Random Forest, along with 4 Ensemble methods: Average, Product, Minimum, and Maximum, and compared their performances.
Results: The Ensemble Average method had the highest area-under the receiver operator characteristic curve (AUC) of 0.692. The top ranked features were all interdependent features.
Conclusion: The findings of this study suggest the predictive power of diverse features, especially when combined, in predicting county-level trends of COVID-19 cases and can be helpful to individuals in making their daily decisions. Our results may guide future studies to consider more features interdependently from conventionally distinct data sources in county-level predictive models. Our code is available at: https://doi.org/10.5281/zenodo.6332944.

Key words: COVID-19, machine learning, county-level case number trend prediction, demographic characteristics, social distancing policies

LAY SUMMARY

Predicting the Coronavirus Disease 2019 (COVID-19) daily trend is important to support individual decisions in taking preventative measures. This study aims to utilize COVID-19 case number history, population demographic characteristics, and social distancing policies to predict the trend in the rise or fall of county-level cases in the United States, with a unique aspect of using predictors from data sources that are conventionally not seen to be combined with each other. Using the top 200 selected features among 2093 ones for 3142 US counties, we built 4 machine learning models, along with 4 ensemble methods, and compared their performances. We achieved relatively reasonable prediction and calibration results across all constructed models, with comparatively negligible runtimes. Our feature analysis showed the most impactful predictors to be features derived from combining independent ones. The findings of this study suggest the importance of diverse features in predicting county-level trends of COVID-19 cases within the United States when they are combined across traditionally distinct domains. Our results may guide future studies to consider more diverse features in predictive models.

INTRODUCTION

With the prevalence of the Coronavirus Disease 2019 (COVID-19), it is critical to understand the pandemic's pattern and characteristics to design effective prevention methods. Among various research tasks such as risk classification[1–4] and medical image analysis,[5–7] COVID-19 case prediction is crucial because it can impact how the government decides on mitigation methods and how medical workers plan for the distribution of healthcare resources. A recent review[8] showed that the state of the pandemic can worsen when precautions are undervalued; thus, case prediction can aid in locating the appropriate level of precautions. In practice, various global-, country-, and state-level COVID-19 case predictions and feature importance analyses have been executed.[3,9–11]

To account for more granular variations, *county-level* COVID-19 case number prediction is especially important for local mitigation of COVID-19. However, predicting case numbers accurately could be challenging. For example, a recent Least Absolute Shrinkage and Selection operator (LASSO) regression model provided moderate correlation (Pearson's correlation coefficient = 0.49) for cases by county.[12] Another spatio-temporal vector autoregressive model had mean absolute error (MAE) between 10% and 16% for most affected counties.[13] Besides, a study showed that linear regression and Multi-Layer Perceptron models resulted in MAE scores ranging from 0.35 to 0.58.[14]

To the best of our knowledge, available models either focus on predicting the *count* of infection (the number of reported cases) rather than the *trend* of infection (the net change of cases over some window of time) or use data at different levels of granularity. Among those using county-level data,[12–14] previously mentioned performance metrics are not necessarily strong. Hence, it is practical to consider relaxing the prediction task from case number to case trend, which is a directional forecast of whether the number of cases would rise or fall, to provide an intuitive guidance for people to make their daily decisions. Several county-level COVID-19 case trend studies use features such as demographic characteristic (eg, age, gender, and ethnicity),[15] government interventions (eg, social distancing policies that affect peoples' movements and behaviors),[16,17] and other features[12–14] *independently* in their models, without considering that infection may spread in a more comprehensive, community-oriented manner. On the other hand, the *combination* of these features (eg, male living in a county whose policy dictates that restaurant occupancy limit is up to 25) can take the relationships between different types of data that are previously precluded from being combined with each other into account. Therefore, a model that both *uses* features from a wide range of data sources not conventionally associated, and *combines* such features to quantify their possible intercorrelation, could potentially provide more insights into how those relationships may impact the trend in county-level COVID-19 case numbers.

OBJECTIVE

This study aims to use (1) daily case number history, (2) demographic characteristics, and (3) social distancing policies both independently (ie, originally collected data), and interdependently (ie, derived from combining independent features), to predict whether the next day would see an increase (positive classification) or decrease (negative classification) in the number of COVID-19 cases relative to the previous date.

MATERIALS AND METHOD

To construct such a predictive model, we first collected and preprocessed the 3 types of data (ie, daily case number history, demographic characteristics, and social distancing policies) into independent and interdependent features. Then, we used these features in 4 machine learning algorithms: Logistic Regression, Naïve Bayes, Multi-Layer Perceptron, and Random Forest, followed by the ensemble of these algorithms in 4 ways (Average, Product, Minimum, and Maximum of the predicted distributions for the positive class). The overall process is shown in Figure 1, and the details of our methodology are described in the following subsections.

Data

We collected data from 3142 counties in the United States.[18] The 3 types of publicly available data in our predictive models are as follows:

1. *County-level daily confirmed cases*. Intuitively, the history of county-level COVID-19 case numbers may contain patterns helpful to predict future trend.[10,13,19,20] We used the US county-level case data from the COVID-19 Data Repository[18] prepared by the Center for Systems Science and Engineering (CSSE) at John Hopkins University (JHU) for its completeness and trustworthiness. The data were collected from sources including the European Centre for Disease Prevention (ECDC), the United States Centers for Disease Control and Prevention (CDC), and the BNO News. We used the confirmed cases data for the 3142 counties from June 4, 2020, to May 17, 2021, for a total of 348 days.
2. *Demographic characteristics information*. Differences in demographic characteristics such as age, sex, and race can affect the likelihood of exposure to COVID-19.[9,12–14,19,20] The county-level demographic characteristics data were collected from the U.S. Census Bureau, latest available as of July 1, 2019.[21] We

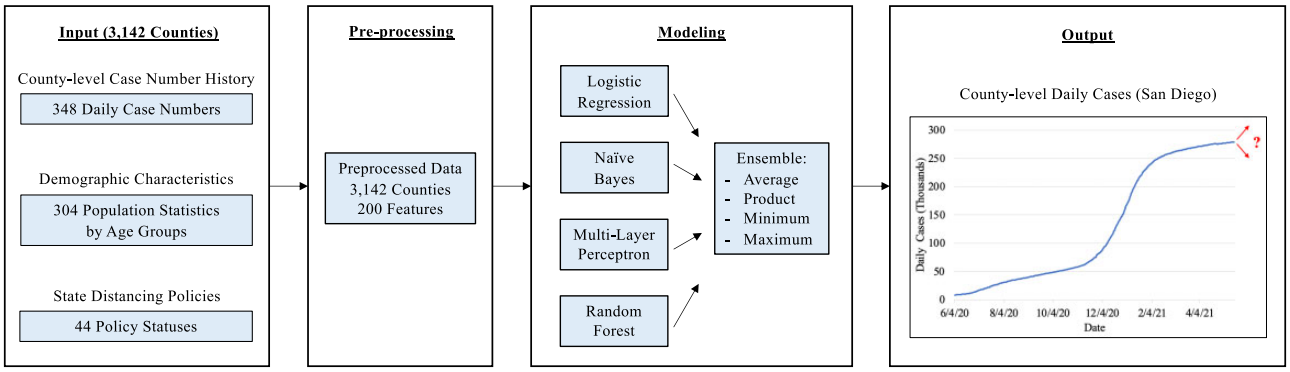


Figure 1. Overview of our predictive modeling pipeline. In this example, features created from case number history, demographic characteristics, and state distancing policies are input into predictive models to predict daily case change as increase or decrease for San Diego County. Our model will include all 3142 counties in the United States.

Table 1. County-level population statistics

No.	Official code	Definition	Example value
1	TOT_MALE	Total Male Population	103 970
2	TOT_FEMALE	Total Female Population	99 195
3	WA_MALE	White Male Population	77 429
4	WA_FEMALE	White Female Population	74 066
5	BA_MALE	African American Male Population	5656
6	BA_FEMALE	African American Female Population	5349
7	IA_MALE	American Indian and Alaska Native Male Population	1315
8	IA_FEMALE	American Indian and Alaska Native Female Population	1311
9	AA_MALE	Asian Male Population	9662
10	AA_FEMALE	Asian Female Population	8852
11	NA_MALE	Native Hawaiian and Other Pacific Islander Male Population	675
12	NA_FEMALE	Native Hawaiian and Other Pacific Islander Female Population	642
13	H_MALE	Hispanic Male Population	46 734
14	H_FEMALE	Hispanic Female Population	44 745
15	HWA_MALE	Hispanic, White Male Population	41 496
16	HWA_FEMALE	Hispanic, White Female Population	39 777

Note: There are 16 county-level population statistics extracted from the U.S. Census Bureau in 2019.[21] The 16 population statistics for San Diego County age group 0–4 are shown in this table.

used 304 variables, derived from 16 population characteristics and 19 distinct age groups, with the first age group being the total of all age groups (0–85+), and most of the other groups being in 4-year increments (eg, 0–4, 5–9, . . ., 85+), as demonstrated in Table 1. Of the 16 population characteristics, 2 were the total male and the total female population. The other 14 population characteristics were 7 male/female pairs of race/ethnicity characteristics. We chose these demographics because they have been found to be more susceptible to COVID-19 transmission.[22–24] Due to the fact that we chose 7 most relevant pairs of races/ethnicities out of 35 pairs available from the data source,[21] with some possible overlap between race and ethnicity, the total male/female population values (Nos 1 and 2 in Table 1) are not the sums of all male population and female population values (Nos 3–16 in Table 1).

3. *State social distancing policies*. Changes in social distancing policies such as gathering limits, or business/restaurant policies to restrict or enable people's movements can also impact COVID-19 incidences.[25,26] Therefore, we used the COVID-19 Data Repository from the Kaiser Family Foundation (KFF)[27] to include this information. This data set contains state-level and structured records, which can be mapped to county-level and includes the state social distancing policy actions for all 50 states, and therefore all 3142 counties in the United States as of a specific date. We obtained the state policy records from April 4, 2020 to May 17, 2021 to cover the whole period of our case number history (ie, June 4, 2020 to May 17, 2021). These records were updated during policy changes (which did not occur daily); therefore, we selected 6 policies that were most consistent/present throughout the period and merged policy statuses with the same meanings as demonstrated in Table 2.

Data preprocessing

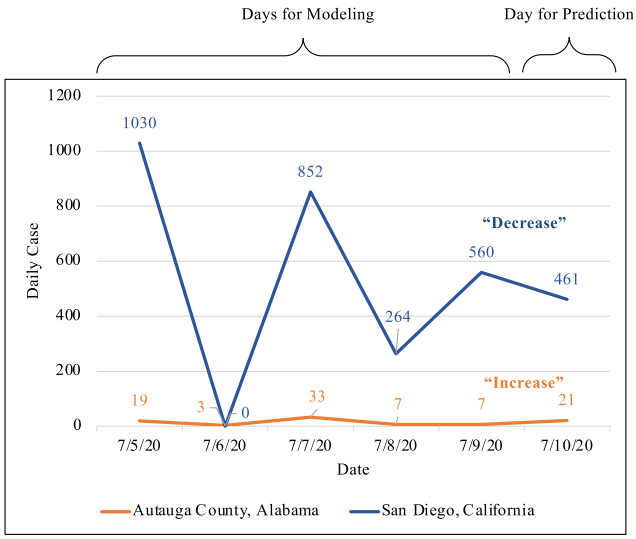
Our data preprocessing steps for the 3 types of data are summarized below. Each type was collected for the 3142 counties in the United States.

1. *Case summaries*. We defined the *Label* for each county as "0" if the value of daily case change was less than or equal to zero and defined the *Label* as "1" otherwise (Figure 2A). For instance, using July 10, 2020 as the label date, 1817 counties would be labeled as "0" and 1325 counties would be labeled as "1" (ie, the positive rate is 42.17%). Using these historical cumulative cases, we calculated the numbers of daily cases and daily case changes to extract 5 case summary features as defined in Table 3 and shown in Figure 2B.
2. *Demographic characteristics*. We used the 304 independent demographic characteristics features (Figure 3C) and created 304 * 5 (the case summary features shown in Table 3)=1520 interdependent demographic characteristic features (Figure 3D) to represent the relationship between case summaries and demographic characteristics.
3. *Social distancing policies*. From the 6 policies defined in Table 2, we cleaned the 54 policy statuses by manually merging statuses with the same meanings (eg, ">25 Prohibited" and "Limit≤25"), resulting in 44 distinct policy statuses. To fill in

Table 2. State social distancing policies

No.	Official code	Definition	Example value
1	RESTAURANT	Restaurant Limits	Open
2	STAY_HOME	Stay at Home Order	Statewide
3	GATHERINGS	Large Gatherings Ban	Limit>50
4	TRAVELER_QUARANTINE	Mandatory Quarantine for Travelers	All Air Travelers
5	BUSINESS_CLOSURES	Nonessential Business Closures	New Business Closures or Limits
6	EMERGENCY_DECLARATION	Emergency Declaration	Yes

Note: There are 6 state social distancing policies, each with different policy statuses (eg, "Open" for "RESTAURANT"), extracted from the Kaiser Family Foundation (KFF) COVID-19 Data Repository.[27]



(A) Label

Day for Prediction	7/10/2020
SD	"0", Decrease (i.e., i.e., Negative, Daily Case Change <= 0)
AC	"1", Increase (i.e., Positive, Daily Case Change > 0)
… up to 3,142 Counties	…

(B) Case Summary Features

Descriptive Name	Case Sum	Case Last Day	Change Positive Days	Change Negative Days	Change Last Day
SD	2,706	560	2	2	296
AC	69	7	1	2	0
… up to 3,142 Counties	…	…	…	…	…

Figure 2. Case summary features. In this example, the daily cases from July 5, 2020 to July 10, 2020 for San Diego, California (SD) and Autauga County, Alabama (AC) are displayed. (A) The label is computed using the daily case change on July 10, 2020. The label for SD is "0" (ie, "decrease") and the label for AC is "1" (ie, "increase"). (B) Case summary features are computed using the daily cases and daily case change from July 5, 2020 to July 9, 2020. Taking this time range for SD for example, the sum of cases is 2706, the number of cases on the last day (ie, July 9, 2020) before prediction day is 560, the number of positive daily case change is 2, the number of negative case change is 2, and the last daily case change (ie, between July 9, 2020 and July 8, 2020) is 296.

Table 3. Case summary features

No.	Feature name	Definition	Example value
1	CASE_SUM	Sum of daily cases	139
2	CASE_LAST_DAY	Case number on last day	83
3	CHG_POS_DAYS	Sum of positive daily case changes	3
4	CHG_NEG_DAYS	Sum of negative daily case changes	1
5	CHG_LAST_DAY	Daily case change on last day	62

Note: The "Case Sum" and "Case Last Day" features are defined using the numbers of daily cases, and the "Change Positive Days," "Change Negative Days," and "Change Last Day" ones are defined using the numbers of daily case changes.

policy statuses for dates without records in the data set, implying days without policy status changes, we used the most recent policy. Then, we used one-hot encoding[28] to encode categorical variables with categorical values into new features, whose numerical values can be "0" representing absent or "1" representing present. The "Emergency Declaration" policy was the only policy with 2 status options "Yes" and "No." Therefore, we use dummy coding[29] to extract only one feature, "Emergency Declaration is Yes," with value "1" if emergency is declared or "0" otherwise. We extracted a total of 44 policy status features (Figure 4E). We also created 44 * 5 (the case summary features shown in Table 3)=220 interdependent policy status features to represent the relationship between case summaries and policy statuses (Figure 4F).

In total, we extracted 5 (case summaries)+304 (independent demographic characteristics)+1520 (interdependent demographic characteristics)+44 (independent policy statuses)+220 (interdependent policy status)=2093 features. We then normalized all features, and selected the top 200 using Gain Ratio[30,31] (which can handle features with many distinct values) to focus on the most relevant features.

Classifiers

We adopted 4 individual classifiers as follows:

- *Logistic regression (LR)*. We used a multinomial logistic regression model with a ridge estimator to guard against overfitting by penalizing large coefficients.[31] To tune this ridge hyperparameter, our search space was [10^1, 10°, …, 10^{-10}].

(C) Demographic Characteristics Independent Features

Demographic	2019 Total Male Population for Ages 0-4	… up to 304 Features
SD	103,970	…
AC	1,713	…
… up to 3,142 Counties	…	…

(B) Case Summary Features

(D) Demographic Characteristics Interdependent Features

Descriptive Name	Case Sum per 2019 Total Male Population for Ages 0-4	… up to 304 Demographic Characteristics x 5 Case Summary Combinations = 1,520 Features
Formula	$\frac{(CASE\_SUM)}{(TOT\_MALE)}$	…
SD	$\frac{2{,}706}{103{,}970} = 0.0260$	…
AC	$\frac{69}{1{,}713} = 0.0403$	…
… up to 3,142 Counties	…	…

Figure 3. Demographic characteristics features. The total male population from 0 to 4 years old in 2019 for San Diego, California (SD) and Autauga County, Alabama (AC) is displayed. (C) There are 304 independent features for demographic characteristics (eg, "2019 Total Male Population for Ages 0–4"), which represent 16 population statistics for 19 age groups, summing to 304 demographic characteristics. (D) Interdependent features combine case summaries and demographic characteristics.

State Distancing Policies for 7/9/2020

Policy	Large Gatherings Ban	Stay at Home Order
SD	All Gatherings Prohibited	Statewide
AC	Lifted	Lifted
… up to 3,142 Counties	…	

(E) Policy Status Independent Features

Policy Status	Large Gatherings Ban is Lifted	Stay at Home Order is Statewide	… up to 44 Features
SD	0	1	…
AC	1	0	…
… up to 3,142 Counties	…	…	…

(B) Case Summary Features

(F) Policy Status Interdependent Features

Descriptive Name	Case Sum if Large Gatherings Ban is Lifted	Case Sum if Stay at Home Order is Statewide	… up to 44 Distancing Policies x 5 Case Summary Combinations = 220 Features
Formula	*CASE_SUM* x *RESTAURANT_OPEN*	*CASE_SUM* x *STAY_HOME_STATEWIDE*	…
SD	0	2,706	…
AC	69	0	…
… up to 3,142 Counties	…	…	…

Figure 4. Distancing policy status features. The state distancing policy statuses as of July 7, 2020 for San Diego, California and Autauga County, Alabama are displayed. (E) Independent features for policy status represent each policy status after one-hot encoding. (F) Interdependent features represent case summaries if a policy status is present.

- *Naïve Bayes (NB)*. We used a Bayesian probabilistic classifier.[32] We tuned hyperparameters for the use of the kernel density estimator or use of supervised discretization, which can both be used to handle numeric attributes, or use of neither.[31]
- *Multilayer perceptron (MLP)*. We used a feed-forward neural network that is trained using back propagation.[31] We tuned hyperparameters for the learning rate, momentum rate, number of epochs to train through, presence of learning rate decay,

number of nodes on each layer, and number of consecutive increases of error allowed before training terminates. Our search space consisted of learning rate=[0.1, 0.3, 0.5], momentum rate=[0.1, 0.2, 0.5], number of epochs=[100, 500, 1000], learning rate decay=[present, absent], and number of consecutive errors=[15, 20].

- *Random forest (RF)*. We used random forest, which is a combination of decision trees.[33] We tuned hyperparameters for the size of each bag, number of iterations, and number of attributes to randomly investigate. Our search space consisted of bag size=[50, 60, 70, 80, 90, 100], iterations=[10, 50, 100, 150, 200, 250, 500, 1000], and number of attributes=[0, 1, 5, 10, 15, 20].

Additionally, we used ensemble to combine the outputs of the 4 classifiers described above, because ensemble methods have been empirically shown to improve discrimination capability.[34,35] We adopted 4 ensemble methods with different combination rules for the predicted distributions for the positive class: Average, Product, Minimum, and Maximum. Average sums each input classifiers' predicted distribution, while Product multiplies the predicted distribution; both normalize the results at the end. Minimum computes the input classifiers' lowest predicted distribution, and Maximum computes the highest predicted distribution.[31,36]

Decision threshold

To estimate the "ideal" decision threshold, we started by assessing the relative harm and benefit for individuals[37] when predicting the next day change in case numbers. We first estimated the case change in all US counties (D) as the "onset of viral outbreak" ($D+$) or "pre-pandemic" ($D-$), and a typical individual's decision (A) to take preventative measures such as self-isolation or quarantine ($A+$) or not ($A-$).[37,38] Combinations of these states ($U[D+\ A+]$, $U[D-\ A+]$, $U[D+\ A-]$, and $U[D-\ A-]$) gives an estimation on the effect on a typical individual's well-being such as fear and anxiety in response to the case change. Following, we estimated net benefit $B = U[D+\ A+]—U[D+\ A-]$,[37] which is the value of self-isolating or quarantining (ie, given that the number of cases is predicted to increase) vs not doing so, in the presence of a positive net case change. We adopted the regression model coefficient of fear or anxiety predicting preventative behaviors (0.13) *during the onset of viral outbreak*,[38] and inverted it to estimate the net benefit of preventative measures on fear or anxiety when case number did increase (ie, onset of viral outbreak). That is, the net benefit $B = 1/0.13 = 7.69$. Similarly, we estimated the net harm $H = U[D-\ A-]—U[D-\ A+]$,[37] which is the value of not self-isolating or quarantining (ie, given that the number of cases is predicted to decrease) vs doing so, in a presence of a negative net case change, using the model coefficient of fear or anxiety predicting preventative (-0.06) during the *pre-pandemic* period.[38] Note that this coefficient of -0.06 compared "taking preventive measures" with "not doing so," and thus was the opposite of computing the net harm. Therefore, we used 0.06 instead, and calculated our estimated net harm $H = 1/0.06 = 16.67$. Finally, to estimate the "ideal" decision threshold $T = H/(H + B)$,[37] we used the estimated H and B values from above to obtain $T = 0.68$.

Validation and evaluation

We performed validation based on the COVID-19 historical case numbers to tune the hyperparameters for the classifiers (Figure 5). Because the transmissibility of COVID-19 in adults ceases after 10 days from symptom onset,[39] we selected 10 days for both the validation phase (to tune the hyperparameters of each classifier) and evaluation phase (to evaluate the models with the best-performed hyperparameters identified in the validation phase). We evaluated the discrimination using full Area-Under the receiver operator characteristic Curve (AUC), sensitivity, specificity, precision, and accuracy, the best-tuned hyperparameters, the training/test time, and the important features learned by the LR classifier. We calculated sensitivity, specificity, precision, and accuracy using our estimated "ideal" decision threshold of 0.68. For all ensemble methods, we used the best-tuned hyperparameters found from each of the 4 classifiers' search space. We implemented our algorithm using Java and the Waikato Environment for Knowledge Analysis (WEKA) library.[31,40] To conduct the experiments, we used a UCSD Campus Amazon Web Services (AWS) Virtual Machine (VM) with 2 vCPUs, 8 GB RAM, and 100 GB SSD hard disk.

RESULTS

Discrimination

We predicted the change in daily case numbers for all 3142 counties, with AUC results shown in Figure 6. No counties had missing features or missing labels of case trends. All single classifiers: LR, RF, MLP, and NB, had average AUC values ranging from 0.665 to 0.683. All ensemble methods had average AUC values ranging from 0.682 to 0.692, with the Ensemble Average having the highest average AUC of 0.692. The Ensemble Maximum had the highest average specificity of 0.735 and precision of 0.806. The Ensemble Product had the highest average sensitivity of 0.693 and accuracy of 0.640.

Important features

The top 10 features with the highest absolute learned coefficients for LR were all interdependent features, shown in Table 4A. All 200 features with their learned coefficients for LR along with the intercept are shown in Supplementary Appendix Table A1. The 6 features that combined case summary data and social distancing policy data included Traveler's Quarantine policy, Gathering limits, and Restaurant limits. The remaining 4 features that combined case summary data and demographic characteristics mostly included Total or White Alone Males, and one for Black Alone Females. In addition, these top predictors all feature populations of higher age groups, ranging from 50 to 79 years.

Meanwhile, among the top 10 important features in the RF model (Table 4B, with all 200 feature results shown in Supplementary Appendix Table A2), there is one interdependent feature of case summary/social distancing policy, which is the case last day value with emergency declaration. The other 9 are interdependent features of case summary/demographics. In particular, the populations of American Indian and Alaska Native males and females, and Native Hawaiian and Other Pacific Islander males and females are spread out among different age groups, with a higher concentration towards the upper range of 50+.

Execution time

As for the evaluation training times, MLP took the longest time of 441.788 s and NB took the least time of less than 1 s. With regards to the evaluation testing times, all classifiers each took less than 1 s, with MLP taking the longest time of 0.80 s and LR taking the least time of 0.018 s. The evaluation testing times for all ensemble methods were also negligible.

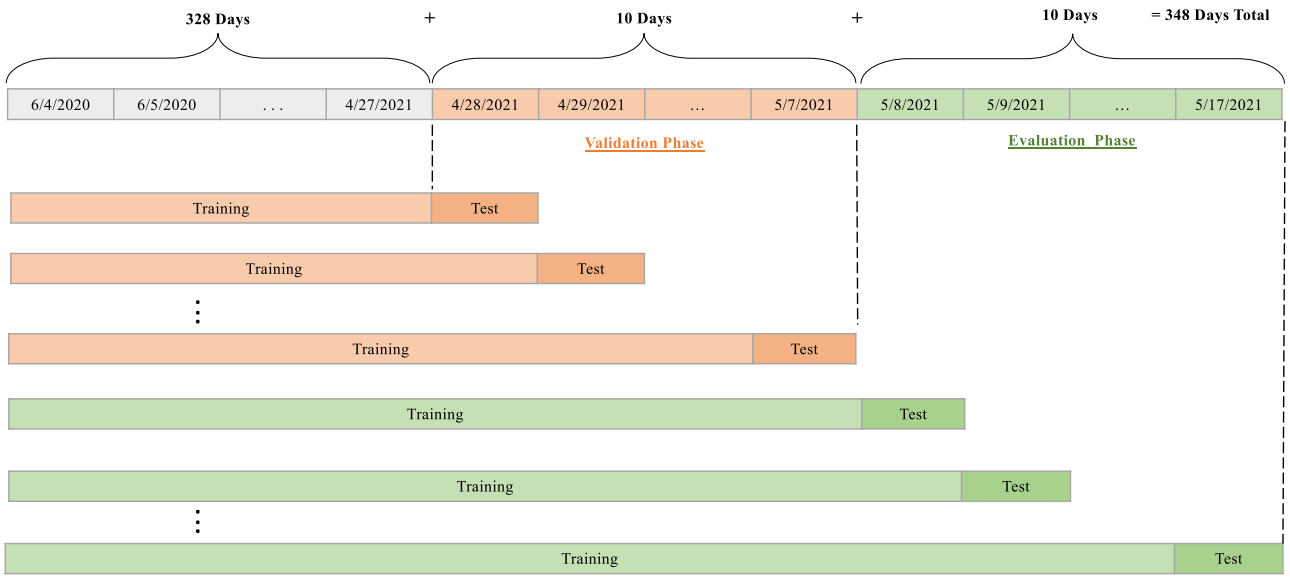


Figure 5. Data splitting for model validation and evaluation. In the validation phase using April 28, 2021 to May 7, 2021 test dates, we execute a grid search to find the best hyperparameters values, which are then used in the models during the evaluation phase using May 8, 2021 to May 17, 2021 test dates.

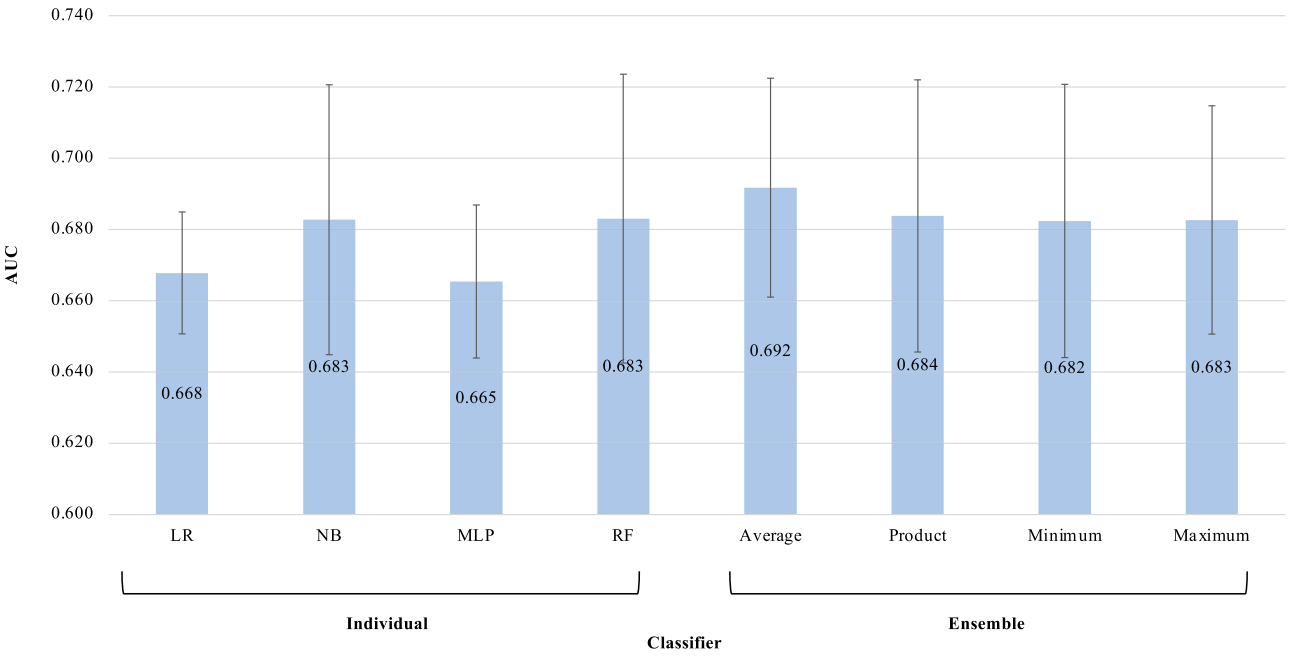


Figure 6. The average full area-under receiver operator characteristic curve (AUC) scores with 95% confidence interval (CI) for individual and ensemble classifiers. AUC scores are represented by the bars and CIs are displayed by the line ranges.

Hyperparameters

For LR, the best hyperparameter value found was "ridge = 10^1." For NB, the best hyperparameter values found were "presence of kernel density estimator=false" and "presence of supervised discretization=true." For MLP, the best hyperparameter values found were "learning rate = 0.1," "momentum rate = 0.2," "number of epochs = 500," "presence of learning rate decay=false," "number of nodes on each layer=(attributes+classes)/2," and "number of consecutive errors = 15." For RF, the best hyperparameter values found were "bag size = 100," "number of iterations = 10," and "number of attributes = 10."

Calibration

We also calibrated our best model (ie, the Ensemble Average) to provide individuals with a more precise probability of case changes, allowing them to make better decisions in taking preventative measures.[41] We applied the Isotonic Regression function[40,42] to the predicted scores from the Ensemble Average calculated from May 16, 2021, and evaluated our calibrated model using features calculated from May 17, 2021 (the last date of our data set). To understand the effectiveness of calibration, we computed the Hosmer and Lemeshow (H-L) test[43] from the calibrated prediction scores and the labels. Specifically, we used H-L H-statistic for equal intervals,

Table 4. (A) Feature analysis results using logistic regression (LR) and (B) feature analysis results using random forest (RF)

(A)	Feature description	Coefficient	Case summaries	Demographic characteristics	Social distancing policies
1	Change Last Day value if Mandatory Quarantine for Travelers applies to certain states	75.425	X		X
2	Case Last Day value if Mandatory Quarantine for Travelers applies to certain states	−32.003	X		X
3	Case Last Day value if Large Gatherings Ban is limited to less than or equal to 25 people	31.933	X		X
4	Change Last Day value if Large Gatherings Ban is limited to less than or equal to 25 people	27.137	X		X
5	Case Last Day value for total Male population ages 75–79 years	17.653	X	X	
6	Change Last Day value if Restaurant Limits Policy is Open with Service Limits	−8.276	X		X
7	Case Sum value for White alone Male population ages 50–54 years	8.067	X	X	
8	Case Sum value for total Male population ages 65–69 years	−7.407	X	X	
9	Change Last Day value for Black or African American alone Female population ages 65–69 years	6.770	X	X	
10	Case Last Day value if Large Gatherings Ban >50 Prohibited	6.127	X		X

(B)	Feature description	Importance index	Case summaries	Demographic characteristics	Social distancing policies
1	Case last day value for American Indian and Alaska Native alone Male population ages 85 years or older	0.534	X	X	
2	Case last day value for Native Hawaiian and Other Pacific Islander alone Female population ages 50–54 years	0.531	X	X	
3	Case last day value for Native Hawaiian and Other Pacific Islander alone Male population ages 50–54 years	0.528	X	X	
4	Case last day value for American Indian and Alaska Native alone Female population ages 65–69 years	0.518	X	X	
5	Case last day value for Native Hawaiian and Other Pacific Islander alone Male population ages 55–59 years	0.507	X	X	
6	Case last day value for Native Hawaiian and Other Pacific Islander alone Female population ages 0–4 years	0.491	X	X	
7	Case last day value for Native Hawaiian and Other Pacific Islander alone Female population ages 70–74 years	0.490	X	X	
8	Case last day value for American Indian and Alaska Native alone Male population ages 70–74 years	0.487	X	X	
9	Case last day value if Emergency Declaration is declared	0.485	X		X
10	Case last day value for American Indian and Alaska Native alone Male population ages 75–79 years	0.483	X	X	

Note: (A) The features, extracted from data on the last date in the evaluation phase, are ordered by the absolute values of their coefficients. The data type used to create each feature is marked with a "X." (B) The features are ordered by their importance indices.

bins = 10, ranging from 0.65 up to 0.70, and with an increment of 0.05. We chose the range (0.65, 0.70) to include neighboring prediction scores from our estimated "ideal" decision threshold of 0.68, calculated in "Decision threshold." The *P*-value of the calibrated model was 0.791, indicating that our best model Ensemble Average is well-calibrated ($P > 0.1$) after calibration.

DISCUSSION

Findings

Our overall AUCs averaging to approximately 0.68 indicate that our prediction task of county-level case trends is still nontrivial. While this AUC may not be sufficient to influence policy makers, it is helpful to individuals, as the use of a discrimination threshold based on the average net harm/benefit to a typical individual suggests that our predictions can aid residents of a county in assessing their motivation to take conservative measures. The top 10 LR-ranked features, as well as the top 10 RF-ranked features, revealed the benefits of integrating case data with demographic characteristics and social distancing policy, given that all 20 previously mentioned features are interdependent ones derived from conventionally distinct data sources. It is seen across 2 methods of identifying important features (coefficient values for Logistic Regression and feature importance indices for Random Forest) that interdependent factors may have a strong influence on COVID-19 trend. Furthermore, out of the selected 200 features, 16 used social distancing policies and 183 used demographic characteristics, while the last feature of "Case Last Day" used a case summary alone. This agrees with existing studies that policies and demographics can affect

COVID-19 transmissibility.[44–47] Demographic characteristic of specific subgroups such as White Alone Males, Black Alone Females, American Indian and Alaska Native, Native Hawaiian and Other Pacific Islander, and higher age groups, as well as social distancing policies involving quarantine rules, gathering sizes, and declaration of emergency are the most impactful features for our prediction task. This presence of minority groups in our top features may alert policy makers to investigate further the impact of COVID-19 on minority populations.

In terms of execution time, the average training time was less than 10 min (using MLP), and the average testing time was at most around 1 s (using MLP). Both training and testing times are reasonable, given that the frequency of our prediction is daily. We also tried to create features using case summaries, the percentage of positive/negative days over 10 days, to use population statistics such as population density[48,49] and location,[50,51] and to adopt demographic characteristics such as age,[46,52] which have been found to impact COVID-19 transmission. However, we found that including these features did not significantly improve prediction results.

Limitations

There are few limitations in our study:

a. *Policy suggestions*. In our models, we predicted the outcome as an increase or decrease of daily case number (ie, predicting for the next day) only. We have yet to consult with public-health policy makers to suggest policies based on our prediction model. For example, we could try to determine what policy a county should execute after *N* days from now. To address these questions, a change of model to predict county case trend *N* days ahead (instead of only 1 day ahead) has yet to be investigated. In addition, we have yet to consult with public health experts to perform a "blind assessment" of our prediction.
b. *Features*. From the census demographic characteristics data set, we selected 7 of 35 pairs of races/ethnicities. We have yet to use all pairs of races/ethnicities, such as "being two or more races" and "Asian alone." Other potentially useful features that encompass demographic details beyond race/ethnicities and age groups such as employment percentage and disadvantaged socioeconomic positions,[53] mobility status,[54] social connectedness data,[55] weather factors,[56] clinical features and pre-existing medical conditions,[57,58] have yet to be integrated into our current models.
c. *Dataset*. The social distancing-related features in our experiments were limited due to lack of consistent and thorough social distancing policy data sets. Only the 6 policies we chose were present from April 4, 2020 to May 17, 2021, which was the timeframe considered, prohibiting us from considering other policy measures like school/university closure, facemask/vaccination mandates, or measures related to travel that are not quarantine-based. Overall, we have yet to identify more public data sets containing consistent social distancing policy information with clear statuses.
d. *Class imbalance*. Given the highly interrelated nature of time series data, the task of handling prediction class imbalance is not trivial. We have yet to adopt techniques to handle the imbalanced distribution of the 2 predicted classes ("0" and "1") in our time series data, such as classic methods of oversampling or undersampling, weighted penalization, as well as other methods that are more specifically engineered towards time series.[59,60]
e. *Validation and feature selection*. As with the nature of time series data, the sequential order of sample days need to be considered, therefore, we adopted a validation scheme similar to the "evaluation on a rolling forecasting origin."[61] We have yet to adapt the classic methods such as single/nested k-fold cross validation in which data are assigned to random groups to validate our models. Furthermore, other feature selection methods such as Information Gain,[62] CfsSubsetEval,[63] and Correlation Attribute Evaluation[63] have yet to be added to our grid search to potentially locate better features.
f. *Model type*. We did not explore the possible presence of causal relationships using models such as Temporal Bayesian Networks.[64] We have yet to include time series forecasting models such as Autoregressive models (AR),[65] and hybrid models such as SeriesNet,[66] along with other complicated models such as bagging,[67] boosting,[68] and deep neural networks.[69]
g. *County stratification*. We have yet to consult with public health experts to create "risk groups" by stratifying the counties by their predicted change of case number, which could consider the varying degrees in county-level vulnerability to COVID-19 transmission.[20]

CONCLUSION

Although there are plenty of existing COVID-19 prediction models, the unique contributions of our study include the following. (1) The experiment results revealed that predicting the county-level trend of COVID-19 case numbers is an important yet nontrivial task. (2) By integrating demographic characteristics and state social distancing policies, we showed that methods such as Ensemble Average performed best. (3) These results can act as a premise for future studies to use other types of data, including the possibility to derive interdependent features from combining such data, to predict the change of pandemic case numbers for each county.

FUNDING

The authors MML, AP, and T-TK were funded by the U.S. National Institutes of Health (NIH) (R00HG009680, R01HL136835, R01GM118609, R01HG011066, U24LM013755, and T15LM011271). The content is solely the responsibility of the author and does not necessarily represent the official views of the NIH. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

MML contributed to conceptualization, methodology, software, validation, formal analysis, investigation, visualization, data curation, and writing (original draft). AP contributed to methodology, investigation, visualization, and writing (review and editing). T-TK contributed to conceptualization, methodology, software, validation, formal analysis, investigation, resources, visualization, supervision, project administration, funding acquisition, and writing (review and editing). AP contributed to writing (review and editing).

SUPPLEMENTARY MATERIAL

Supplementary material is available at *JAMIA Open* online.

ACKNOWLEDGMENTS

The use of the UCSD Campus AWS cloud network was supported by Michael Hogarth, MD.

CONFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILABILITY

The data underlying this article are available in Zenodo, at https://doi.org/10.5281/zenodo.6332944. The data sets were derived from sources in the public domain: https://github.com/CSSEGISandData/COVID-19, https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-detail.html, and https://github.com/KFFData/COVID-19-Data/tree/kff_master/State%20Policy%20Actions/State%20Social%20Distancing%20Actions.

REFERENCES

1. Bird JJ, Barnes CM, Premebida C, Ekárt A, Faria DR. Country-level pandemic risk and preparedness classification based on COVID-19 data: a machine learning approach. *PLoS One* 2020; 15 (10): e0241332.
2. Khan SS, Krefman AE, McCabe ME, *et al.* Association between county-level risk groups and COVID-19 outcomes in the United States: a socioecological study. *BMC Public Health* 2022; 22 (1): 81.
3. Huang J, Zhang L, Liu X, *et al.* Global prediction system for COVID-19 pandemic. *Sci Bull* 2020; 65 (22): 1884–7.
4. Edelson M, Kuo T-T. Generalizable prediction of COVID-19 mortality on worldwide patient data. *JAMIA Open* 2022; 5 (2): 1–9.
5. Li K, Fang Y, Li W, *et al.* CT image visual quantitative evaluation and clinical classification of coronavirus disease (COVID-19). *Eur Radiol* 2020; 30 (8): 4407–16.
6. Li K, Wu J, Wu F, *et al.* The clinical and chest CT features associated with severe and critical COVID-19 pneumonia. *Invest Radiol* 2020; 55 (6): 327–31.
7. Li MM, Kuo T-T. Previewable contract-based on-chain X-ray image sharing framework for clinical research. *Int J Med Inform* 2021; 156: 104599.
8. Cakir Z, Savas H. A mathematical modelling approach in the spread of the novel 2019 coronavirus SARS-CoV-2 (COVID-19) pandemic. *Electron J Gen Med* 2020; 17 (4): em205.
9. Hirschprung RS, Hajaj C. Prediction model for the spread of the COVID-19 outbreak in the global environment. *Heliyon* 2021; 7 (7): e07416.
10. Hamzah FB, Lau C, Nazri H, *et al.* CoronaTracker: worldwide COVID-19 outbreak data analysis and prediction. *Bull World Health Organ* 2020; 1 (32): 1–32.
11. Roy S, Ghosh P. Factors affecting COVID-19 infected and death rates inform lockdown-related policymaking. *PLoS One* 2020; 15 (10): e0241165.
12. Richmond HL, Tome J, Rochani H, Fung IC-H, Shah GH, Schwind JS. The use of penalized regression analysis to identify county-level demographic and socioeconomic variables predictive of increased COVID-19 cumulative case rates in the state of Georgia. *Int J Environ Res Public Health* 2020; 17 (21): 8036.
13. Zhu S, Bukharin A, Xie L, Santillana M, Yang S, Xie Y. High-resolution spatio-temporal model for county-level COVID-19 activity in the U.S. *ACM Trans Manage Inf Syst* 2021; 12 (4): 1–20.
14. Mollalo A, Rivera KM, Vahedi B. Artificial neural network modeling of novel coronavirus (COVID-19) incidence rates across the continental United States. *Int J Environ Res Public Health* 2020; 17 (12): 4204.
15. Karmakar M, Lantz PM, Tipirneni R. Association of social and demographic factors with COVID-19 incidence and death rates in the US. *JAMA Netw Open* 2021; 4 (1): e2036462.
16. Bhowmik T, Tirtha SD, Iraganaboina NC, Eluru N. A comprehensive analysis of COVID-19 transmission and mortality rates at the county level in the United States considering socio-demographics, health indicators, mobility trends and health care infrastructure attributes. *PLoS One* 2021; 16 (4): e0249133.
17. Engle S, Stromme J, Zhou A. *Staying at Home: Mobility Effects of Covid-19*. Available at SSRN 3565703; 2020.
18. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect Dis* 2020; 20 (5): 533–4.
19. Li Q, Yang Y, Wang W, *et al.* Unraveling the dynamic importance of county-level features in trajectory of COVID-19. *Sci Rep* 2021; 11 (1): 13058.
20. Mehta M, Julaiti J, Griffin P, Kumara S. Early stage machine learning–based prediction of US county vulnerability to the COVID-19 pandemic: machine learning approach. *JMIR Public Health Surveill* 2020; 6 (3): e19446.
21. United State Census Bureau. *County Population by Characteristics: 2010–2019*. Suitland, MD: U.S Department of Commerce; 2021.
22. Stokes EK, Zambrano LD, Anderson KN, *et al.* Coronavirus disease 2019 case surveillance—United States, January 22–May 30, 2020. *MMWR Morb Mortal Wkly Rep* 2020; 69 (24): 759–65.
23. Figueroa JF, Wadhera RK, Lee D, Yeh RW, Sommers BD. Community-level factors associated with racial and ethnic disparities in COVID-19 rates in Massachusetts: study examines community-level factors associated with racial and ethnic disparities in COVID-19 rates in Massachusetts. *Health Aff (Millwood)* 2020; 39 (11): 1984–92.
24. Boserup B, McKenney M, Elkbuli A. Disproportionate impact of COVID-19 pandemic on racial and ethnic minorities. *Am Surg* 2020; 86 (12): 1615–22.
25. VoPham T, Weaver MD, Hart JE, Ton M, White E, Newcomb PA. Effect of social distancing on COVID-19 incidence and mortality in the US. medRxiv 2020: 2020.06.10.20127589. doi: 10.1101/2020.06.10.20127589.
26. Thunström L, Newbold SC, Finnoff D, Ashworth M, Shogren JF. The benefits and costs of using social distancing to flatten the curve for COVID-19. *J Benefit Cost Anal* 2020; 11 (2): 179–95.
27. Foundation KF. KFF COVID-19 Data Repository. https://github.com/KFFData/COVID-19-Data. Accessed June 30, 2021.
28. Zheng A, Casari A. *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. Sebastopol, CA: O'Reilly Media, Inc.; 2018.
29. Hensher DA, Rose JM, Rose JM, Greene WH. *Applied Choice Analysis: A Primer*. New York, NY: Cambridge University Press; 2005.
30. Quinlan JR. Induction of decision trees. *Mach Learn* 1986; 1 (1): 81–106.
31. Ian HW, Eibe F. *Data Mining: Practical Machine Learning Tools and Techniques*. Burlington, MA: Morgan Kaufmann Publishers; 2005.
32. Rish I. An empirical study of the naive Bayes classifier. In: IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence. San Francisco, CA: Morgan Kaufmann Publishers Inc.; 2001.
33. Breiman L. Random forests. *Mach Learn* 2001; 45 (1): 5–32.
34. Kuo T-T, Kim J, Gabriel RA. Privacy-preserving model learning on a blockchain network-of-networks. *J Am Med Inform Assoc* 2020; 27 (3): 343–54.
35. Kuo T-T, Rao P, Maehara C, *et al.* Ensembles of nlp tools for data element extraction from clinical notes. In: AMIA Annual Symposium Proceedings; 2016. American Medical Informatics Association.
36. Kittler J, Hatef M, Duin RP, Matas J. On combining classifiers. *IEEE Trans Pattern Anal Mach Intell* 1998; 20 (3): 226–39.
37. Sox HC, Higgins MC, Owens DK. *Medical Decision Making*. 2nd ed. Germany: John Wiley & Sons, Ltd; 2013.
38. Li Y, Luan S, Li Y, Hertwig R. Changing emotions in the COVID-19 pandemic: a four-wave longitudinal study in the United States and China. *Soc Sci Med* 2021; 285: 114222.
39. Centers for Disease Control and Prevention. *Interim Guidance on Ending Isolation and Precautions for Adults with COVID-19*. Atlanta, GA: U.S. Department of Health & Human Services; 2021.
40. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. *SIGKDD Explor Newsl* 2009; 11 (1): 10–8.

41. Huang Y, Li W, Macheret F, Gabriel RA, Ohno-Machado L. A tutorial on calibration measurements and calibration models for clinical prediction models. *J Am Med Inform Assoc* 2020; 27 (4): 621–33.
42. De Leeuw J, Hornik K, Mair P. Isotone optimization in R: pool-adjacent-violators algorithm (PAVA) and active set methods. *J Stat Softw* 2010; 32: 1–24.
43. Hosmer DW, Lemesbow S. Goodness of fit tests for the multiple logistic regression model. *Commun Stat Theory Methods* 1980; 9 (10): 1043–69.
44. Pan D, Sze S, Minhas JS, *et al.* The impact of ethnicity on clinical outcomes in COVID-19: a systematic review. *EClinicalMedicine* 2020; 23: 100404.
45. Kakkar N, Dunphy J, Raza M. Ethnicity profiles of COVID-19 admissions and outcomes. *J Infect* 2020; 81 (2): e110–e111.
46. Oster AM, Caruso E, DeVies J, Hartnett KP, Boehmer TK. Transmission dynamics by age group in COVID-19 hotspot counties—United States, April–September 2020. *MMWR Morb Mortal Wkly Rep* 2020; 69 (41): 1494–6.
47. Li M, Zhang Z, Cao W, *et al.* Identifying novel factors associated with COVID-19 transmission and fatality using the machine learning approach. *Sci Total Environ* 2021; 764: 142810.
48. Santosh K. COVID-19 prediction models and unexploited data. *J Med Syst* 2020; 44 (9): 1–4.
49. Andersen LM, Harden SR, Sugg MM, Runkle JD, Lundquist TE. Analyzing the spatial determinants of local Covid-19 transmission in the United States. *Sci Total Environ* 2021; 754: 142396.
50. Team CC-R. Geographic differences in COVID-19 cases, deaths, and incidence – United States, February 12-April 7, 2020. *MMWR Morb Mortal Wkly Rep* 2020; 69 (15): 465–71.
51. Jen T-H, Chien T-W, Yeh Y-T, Lin J-CJ, Kuo S-C, Chou W. Geographic risk assessment of COVID-19 transmission using recent data: an observational study. *Medicine (Baltimore)* 2020; 99 (24): e20774.
52. Dowd JB, Andriano L, Brazel DM, *et al.* Demographic science aids in understanding the spread and fatality rates of COVID-19. *Proc Natl Acad Sci USA* 2020; 117 (18): 9696–8.
53. Cifuentes-Faura J. COVID-19 mortality rate and its incidence in Latin America: dependence on demographic and economic variables. *Int J Environ Res Public Health* 2021; 18 (13): 6900.
54. Wang J, Tang K, Feng K, *et al.* Impact of temperature and relative humidity on the transmission of COVID-19: a modelling study in China and the United States. *BMJ Open* 2021; 11 (2): e043863.
55. Kuchler T, Russel D, Stroebel J. JUE insight: the geographic spread of COVID-19 correlates with the structure of social networks as measured by Facebook. *J Urban Econ* 2022; 127: 103314.
56. Khalatbari-Soltani S, Cumming RC, Delpierre C, Kelly-Irving M. Importance of collecting data on socioeconomic determinants from the early stage of the COVID-19 outbreak onwards. *J Epidemiol Commun Health* 2020; 74 (8): 620–3.
57. Cecconi M, Piovani D, Brunetta E, *et al.* Early predictors of clinical deterioration in a cohort of 239 patients hospitalized for Covid-19 infection in Lombardy, Italy. *J Clin Med* 2020; 9 (5): 1548.
58. Huang C, Wang Y, Li X, *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* 2020; 395 (10223): 497–506.
59. Cao H, Li X-L, Woon Y-K. Ng S-K. SPO: structure preserving oversampling for imbalanced time series classification. In: 2011 IEEE 11th International Conference on Data Mining. IEEE, 2011.
60. Zhu T, Luo C, Zhang Z, Li J, Ren S, Zeng Y. Minority oversampling for imbalanced time series classification. *Knowl Based Syst* 2022; 247: 108764.
61. Hyndman RJ, Athanasopoulos G. *Forecasting: Principles and Practice*. Melbourne, Australia: OTexts; 2018.
62. Azhagusundari B, Thanamani AS. Feature selection based on information gain. *Int J Innovative Technol Explor Eng (IJITEE)* 2013; 2 (2): 18–21.
63. Gnanambal S, Thangaraj M, Meenatchi V, Gayathri V. Classification algorithms with attribute selection: an evaluation study using WEKA. *Int J Adv Netw Appl* 2018; 9 (6): 3640–4.
64. Arroyo-Figueroa G, Sucar LE. A temporal Bayesian network for diagnosis and prediction. In: Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence. Stockholm, Sweden: Morgan Kaufmann Publishers Inc.; 1999: 13–20.
65. Box GE, Pierce DA. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *J Am Stat Assoc* 1970; 65 (332): 1509–26.
66. Shen Z, Zhang Y, Lu J, Xu J, Xiao G. A novel time series forecasting model with deep learning. *Neurocomputing* 2020; 396: 302–13.
67. Chen T, Ren J. Bagging for Gaussian process regression. *Neurocomputing* 2009; 72 (7–9): 1605–10.
68. Bühlmann P, Hothorn T. Boosting algorithms: regularization, prediction and model fitting. *Stat Sci* 2007; 22 (4): 477–505.
69. Larochelle H, Bengio Y, Louradour J, Lamblin P. Exploring strategies for training deep neural networks. *J Mach Learn Res* 2009; 10 (1): 1–40.