scientific reports

Check for updates

OPEN

Analysis of the effectiveness of face-coverings on the death ratio of COVID-19 using machine learning

Ali Lafzi[1]✉, Miad Boodaghi[2], Siavash Zamani[2], Niyousha Mohammadshafie[3] & Veeraraghava Raju Hasti[2]

The recent outbreak of the COVID-19 led to death of millions of people worldwide. To stave off the spread of the virus, the authorities in the US employed different strategies, including the mask mandate order issued by the states' governors. In the current work, we defined a parameter called average death ratio as the monthly average of the number of daily deaths to the monthly average number of daily cases. We utilized survey data to quantify people's abidance by the mask mandate order. Additionally, we implicitly addressed the extent to which people abide by the mask mandate order, which may depend on some parameters such as population, income, and education level. Using different machine learning classification algorithms, we investigated how the decrease or increase in death ratio for the counties in the US West Coast correlates with the input parameters. The results showed that for the majority of counties, the mask mandate order decreased the death ratio, reflecting the effectiveness of such a preventive measure on the West Coast. Additionally, the changes in the death ratio demonstrated a noticeable correlation with the socio-economic condition of each county. Moreover, the results showed a promising classification accuracy score as high as 90%.

The recent COVID-19 pandemic has affected millions of people worldwide and led to the tragic death of many innocent lives. The lack of a certain treatment at the beginning of the pandemic traumatized the populace. The only solutions were limited to preventive actions such as wearing face coverings, maintaining social distancing, washing hands, and self-quarantine. Owing to the high transmission rate, only in the US, the number of new daily cases increased from 6 to 22,562 during March 2020 according to CDC (Center for Disease Control and Prevention)[1]. There is still extensive ongoing research about the possible factors being effective in the pace of this spread; as of now, scientists have declared that meteorological factors such as temperature, wind speed, precipitation, and humidity are some of the critical environmental parameters in this regard[2]. However, controlling environmental factors involved in the spread of COVID-19 are very challenging and sometimes impossible. As a result, state officials began to impose legislative guidelines, including mandatory use of masks and closure of businesses such as bars and restaurants. Shutting down different businesses has been sporadic due to its adverse economic impact, but obligatory face coverings order is still in effect across the US. In this respect, the effectiveness of facial masks gains further importance and requires scientific studies.

Presenting a model that can measure the effectiveness of the mask mandate orders can pave the way for governments to take decisive actions during pandemics. The experimental data in tandem with mathematical modelings can be utilized to study the effects of facial coverings on the spread of viral infections. Many previous publications have tried to address the effectiveness of nonpharmaceutical interventions (NPIs) during pandemics, particularly for the spread of influenza[3,4]. Deterministic models have been widely used to study the effects of facial masks on the reproduction number R_0. Indeed, the face mask is taken into account by its role in reducing the transmission per contact[5]. The results of the deterministic model indicated that public use of face masks delays the influenza pandemic. On the other hand, some studies suggest that the use of a face mask does not substantially affect influenza transmission and there is little evidence in favor of the effectiveness of facial masks[6,7]. As for the COVID-19, the efficacy of the facial mask in impeding the infectivity of the SARS-CoV-2 remains unclear. Considering the effects of mask in reproduction number R_0, Li et al.[8] claimed that wearing face masks alongside

[1]Department of Agricultural and Biological Engineering, Purdue University, Indiana 47907, USA. [2]School of Mechanical Engineering, Purdue University, Indiana 47907, USA. [3]Department of Civil and Environmental Engineering, University of Pittsburgh, Pennsylvania 15261, USA. ✉email: alafzi@purdue.edu

the social distancing can flatten the epidemic curve. Other studies also pinpointed that public use of a facial mask may reduce the spread of COVID-19[9]. Despite these findings, the efficacy of face masks remains controversial.

The cardinal point that has not garnered enough attention is the relationship between the degree of exposure to the virus and its mortality rate. Some researchers presented the idea that the severity of the symptoms correlates with the extent of exposure to justify the high death rate in healthcare workers[10]. Unfortunately, there is no universal trend that can predict the relationship between the dose of the virus and the severity of the resulting symptoms. A study performed on the relationship between influenza and rhinovirus viral load and the severity in the upper respiratory tract infections reported a different behavior for those viruses[11]. In fact, the results indicated that for influenza A and the rhinovirus, viral loads were not associated with hospitalization/ICU. On the other hand, for influenza B, viral load was higher in hospitalized/ICU patients. Furthermore, for respiratory syncytial virus (RSV), the viral load seems to correlate with the severity of symptoms as many studies in the literature suggest that a correlation exists[12–14]. The same controversy holds for the COVID-19. Recently, some studies have tried to investigate the severity of COVID-19 with its load, where they found that the load tightly correlates with the severity[15,16]. However, another study suggests that no such a correlation exists[17].

To unveil whether COVID-19 viral load is related to disease severity requires an in-depth study, which involves infecting volunteers with controlled doses of virus and monitoring their symptoms. However, experimental challenges, in addition to the ethicality of these experiments, make this type of research very challenging[10]. Although studies have not been convergent in whether nose[18] or mouth[19] is the primary site for COVID-19 infection, they underscored the importance of wearing a facial mask as a barrier to the virus spread. Additionally, although the protection level of different types of mask is different, wearing any mask, even a cloth mask, is better than wearing nothing at all, which can play a role in protection from the exposure to COVID-19[20,21]. As mentioned, conducting experimental studies to reveal the relationship between the extent of exposure and severity of COVID-19 is very challenging. One way to circumvent theses challenges is to conduct an indirect study by introducing a model that can capture changes in the mortality rate due to wearing a facial mask. Indeed, if the ratio of the number of deaths to the number of cases decreases, this can support the hypothesis that there is a correlation between the viral load and the severity of symptoms. Thus, studying the effects of Mask Mandate order on the mortality rate gains extra importance.

A Machine Learning (ML) analysis can be instrumental in shedding light on the possible correlation between the public use of masks and changes in the mortality rate. The success of implementing ML and Artificial Intelligence (AI) techniques in the previous pandemics has convinced researchers to use them as precious tools in fighting against the current outbreak[22]. ML and AI can be used for prediction and forecasting in different regions so that the corresponding health officials can take necessary actions in advance[22]. In addition, this technology is capable of enhancing the prediction accuracy for screening both infectious and non-infectious diseases[23]. Six ML methods have been carried out to predict 1, 3, and 6 days ahead the total number of confirmed COVID-19 cases with errors in the ranges of 0.87%–3.51%, 1.02%–5.63%, and 0.95%–6.90%, respectively, in 10 Brazilian states[24]. Moreover, an ML method like XGBoost model was capable of identifying 3 important biomarkers from 485 blood samples in Wuhan, China as the key mortality parameters[25]. ML algorithms also have been used to capture the correlation between the weather data and COVID-19 mortality and transmission rates[26,27]. Additionally, ML has been utilized to study the effects of mask mandate (MM) order on the number of daily cases, where no significant statistical difference was observed in the number of daily cases in the state-wise analysis[28]. These studies confirm the strength of ML as a great tool to investigate the effects of MM order on mortality rates of COVID-19.

Another important factor regarding the effectiveness of MM order is society's adherence to the regulations. One study that tried to quantify the public compliance with COVID-19 public health recommendations found notable regional differences in intent to follow health guidelines[29]. In addition, some studies noticed a correlation between the level of education and intent to voluntarily adhere to social distancing guidelines[29,30]. However, not only the level of education but also the level of income and race can play a role in the adherence to the regulations[31]. Based on these findings, it is important to take into account the features that might be correlated with people's compliance with the MM order. Additionally, we will use data based on the survey provided by the New York Times (NYT) available on GitHub, which quantifies people's adherence to the MM order[32]. As a result, in this study, we will include factors that might play a role in people's adherence to the MM order as our input features.

In the proposed work, utilizing different ML classification algorithms, we aim to unveil how the change in the mortality rate correlates with certain features. The features will be chosen in a way that they can reflect abidance by MM order in different counties. We will use the data provided by CDC to find the average monthly number of COVID-19 cases. Additionally, the exact dates of the executive orders signed by the state officials are available for each state—California: June 18th 2020, Oregon: June 19th 2020, Washington: June 26th 2020. To have appropriate unbiased data, similar to what Maloney et al.[28] has done in his study of the effect of mask mandate, we will be using the data for one month after and before the executive orders for each preventive measure for the three states on US West Coast. With this data selection method, we limit the geographical region of the study to ensure that changes in the cases are highly attributed to the public use of masks rather than other factors such as environmental changes.

The rest of the paper is organized as follows. First, we will represent how our data was collected and arranged. Then we will explain the ML methods we have used for our prediction. Finally, we will describe and compare the results obtained from different ML methods.

Methodology

In this section, we will explain the collected data and the ML algorithms used for the training and prediction.

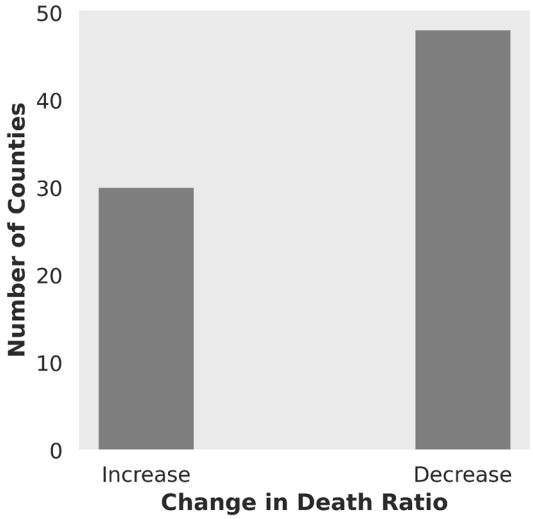

Figure 1. Histogram of change in death ratio for the three states.

Data. We defined the parameter of interest as the ratio of the monthly average number of deaths to the monthly average number of cases, referred to as the death ratio, which can be interpreted as a measure of the severity of the disease. The effective date of the executive orders by the governors, requiring mask mandate at all the counties in the three West Coast states of California, Oregon, and Washington, has been identified, which is publicly available[33]. We used the average death ratio one month before and after the order to study the mortality rate. The rationale behind this selection is to minimize the effects of other factors that might play a role in changing the COVID-19 data. The raw dataset for the daily cases and deaths for all the US counties over time is extracted from the USAFACTS website[34], where county-level data is confirmed by the state and local agencies directly. After obtaining the daily values of death and case numbers for a month before and after the MM order, we divided the monthly average number of deaths by the monthly average number of cases for each county. Then we found the difference between the death ratio for one month before and after the MM order. Finally, we categorized the variation based on its sign to quantify whether the death ratio increases, decreases, or no change occurs. Out of the 130 samples, 47, 30, and 53 of them belong to the "decrease", "increase", and "no change" classes, respectively. We dropped the "no change" data as they all correspond to small counties, where there were zero reported COVID-19 cases and deaths, leaving 77 counties in total. Consequently, the two categories of increase and decrease in the death ratio remain for the prediction task. Figure 1 illustrates the number of samples in each category, which expresses that the available data for classification is not biased.

It is a hard task to directly determine the exact percentage of the population that follows the MM order and uses face coverings. As a result, it is necessary to come up with features that can indirectly capture how likely is an individual to follow the recommended practice. For bridging this gap, four main features are chosen as primary indicators, which are listed below:

1. County population.
2. Median household income.
3. Education level.
4. Mask usage based on New York Times survey.

Population in each county is obtained from the most recent surveys for the year 2019. The income level is the median household income in US dollars and the education level is the percentage of people who have completed high school in each county in the years 2015–2019. The raw data for these features is obtained from the US Census website[35]. The US Census measures the median income as the regular income received, excluding other payments like tax, etc[36]. Furthermore, we used survey data provided by the New York Times that quantifies the mask usage from 7/2/2020 to 7/14/2020[32]. Since the survey timeline lies within the month after the MM order for all three studied states, it is valid to use its data for our purpose. Finally, we will try to establish an AI-based relationship between the features and the sign of the change in the death ratios of the Pacific Coast states at the county level using nine different classification algorithms, provided in "Methods".

Methods. In this study, we have developed machine learning models to employ the specified features mentioned in "Data" to shed light on the relationship between adherence to mask mandate and mortality rate.

Classic ML methods of Logistic Regression[37] and Naive Bayes classifier[38] are used. In addition, ensemble learning-based models, Random Forest and Extra Trees, are also analyzed[39]. Moreover, the extreme boosting method, XGBoost is explored[40]. Other methods such as Support Vector Machine, K-Nearest Neighbors (KNN)[41], Decision Trees[42], and Neural Network[43] are additionally used for prediction of effect of Mask Mandate on the mortality rate.

				Mask usage					
	P	MI ($)	EL	N	R	S	F	A	DR (%)
Count	77	77	77	77	77	77	77	77	77
Mean	630,413.5	66,494.23	0.85	0.03	0.03	0.06	0.17	0.71	− 0.47
Std	1,297,275	18,484.92	0.07	0.02	0.03	0.03	0.05	0.09	2.83
Min	7208	43,313	0.67	0.001	0	0.004	0.07	0.31	− 12.9
25%	86,085	53,105	0.81	0.02	0.01	0.04	0.14	0.67	− 1.4
50%	219,186	62,077	0.88	0.02	0.02	0.06	0.16	0.72	− 0.44
75%	601,592	74,624	0.91	0.04	0.04	0.08	0.2	0.77	0.77
Max	10,039,110	124,055	0.96	0.11	0.21	0.21	0.3	0.87	7.69

Table 1. Statistical summary of the final dataset before scaling. Columns are P: population, MI: median income, EL: education level. Mask usage—N: never, R: rarely, S: sometimes, F: frequently, A: always. DR: change in death ratio between one month before and after the corresponding MM order date.

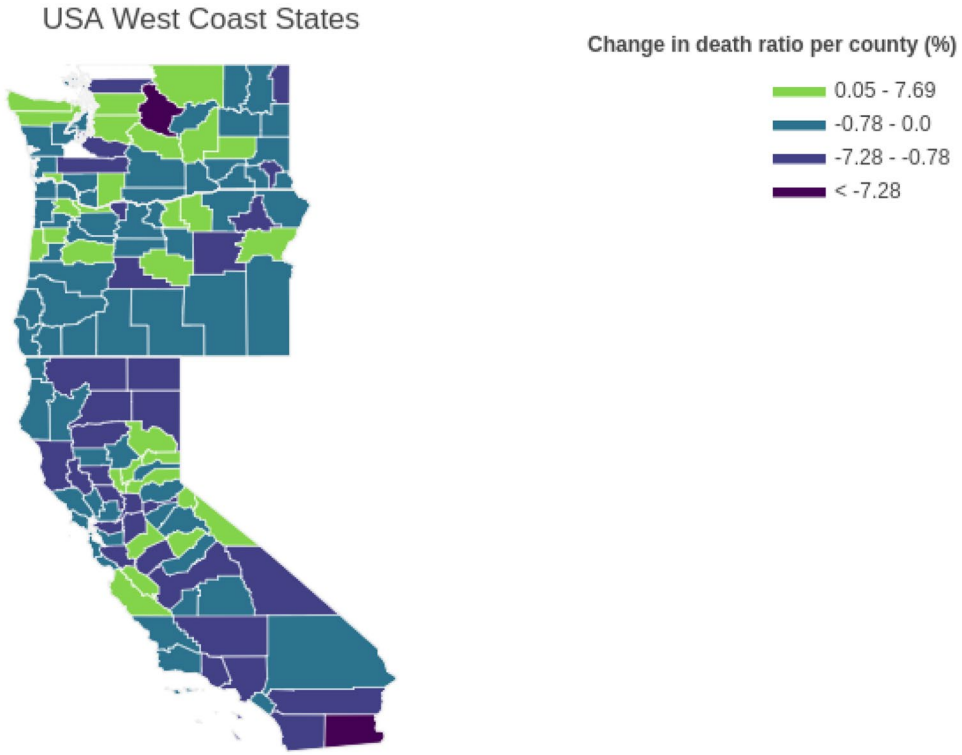

Figure 2. Change in death ratio in US West Coast states counties.

It should be noted that for carrying out the analysis, the data is split into training and test sets, with a test size of 20%[44]. A k-fold cross-validation scheme with five folds has been used to evaluate the performance of each method on the validation set and tune its hyper-parameters with the classification accuracy as the metric accordingly. The hyper-parameter tuning is done using either grid search or random search for all the methods. A statistical summary of the final dataset for binary classification is outlined in Table 1, which indicates a significant difference between the orders of magnitudes of the features. Therefore, min-max and max-abs scaling have been used to transform the input features and output, respectively, before passing the data to the ML algorithms for training. It should be noted that the data used in this article was accessed through publicly available sources as listed, and we confirm that all methods were performed in accordance with the relevant guidelines and regulations.

Results and discussions

The change in death ratio from one month before to one month after the date of mandating face-covering in the three states is visualized for each county in Fig. 2. Two clusters of increase in death ratio can be seen, one near northern Washington and one near central California. Our first intuition was that by increasing the population, the chance of viral spread would increase. Therefore, we expected to see a positive change in the death ratio for

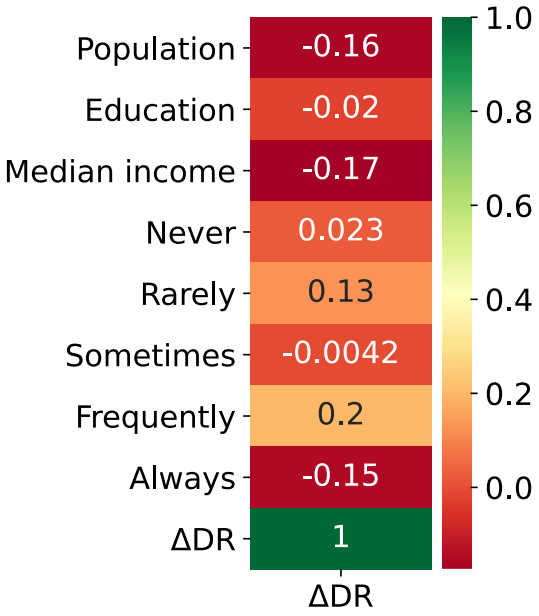

Figure 3. Correlations between the features and the output.

more populated counties. However, as it can be seen from the map, there is inherent randomness that defies our initial intuition about the spread mechanism. Further, it is shown that more counties experienced a decrease in death ratio one month after the usage of face-covering was mandated by each state, as shown in Fig. 1. Therefore, usage of face-covering is chosen as the main factor affecting the decrease of the change in death ratio. As explained previously, to quantify adherence to the mask mandate, other auxiliary features are chosen, namely, median income, and education level for each county.

The combined effect of features is analyzed on the death ratio. Then the performance of each algorithm is evaluated for test and train sets. The effect of each feature on the change of death ratio is visualized by the correlation heatmap provided in Fig. 3. Here, we have not presented the cross-correlations between features for a simpler visualization. Last element in each row of the complete correlation matrix is an appropriate indicator of how correlated the corresponding feature is with the change in death ratio, which is what Fig. 3 illustrates. A more negative value implies that the increase of that specific feature is positively correlated with a decrease in the change of death ratio. For instance, an increase in population, median income, and education level would result in a decrease in the change of death ratio. An interesting observation is the disordered correlation pattern for mask usage. As one expects, increasing the number of never and rarely mask users is positively correlated with a change in the death ratio. However, the data associated with frequently mask users have resulted in a positive correlation value. Such erratic correlation behavior necessitates the inclusion of other features in the analysis.

As a preliminary analysis, the relationship between the average values of the three auxiliary features and the change in death ratio has been visualized for each category separately in Fig. 4. Figure 4a expresses that the average percentages of people who have completed high school education in both categories of counties that have experienced an increase or decrease in their death ratios are almost the same. This could indicate why the correlation between this feature and output is very close to zero, as represented in Fig. 3. Further, a noticeable correlation is observed between average median income and the change of death ratio, presented in Fig. 4b. On average, the communities with less median income experienced a positive change in death ratio, meaning more mortality rate, which is in agreement with what is reported in[31]. However, the strongest correlation is observed by considering county population, shown in Fig. 4c. The counties with fewer residents were affected more adversely by the pandemic compared to high-population counties. The counterintuitive relation between population and change in death ratio further corroborates the necessity of inclusion of the two other supplementary features.

To have an initial assessment of the variation of the percentage change in the death ratio, we plotted the change in the death ratio as functions of population, median income, and portion of the population that frequently uses mask, which has a relatively high correlation coefficient according to Fig. 3. Figures 5a–c show no detectable pattern between parameters of interest and change in the death ratio. As a result, it is not possible to predict the value of change in the death ratio using regression. On the other hand, as we will show, converting changes to categories of increase and decrease would pave the way for capturing the status of the change.

A summary of the overall death ratios in the months before and after the mask mandate order for the three states is presented in Table 2. It can be observed that death ratio significantly decreases in California and Washington but slightly increases in Oregon. This result suggests an intrinsically complex pattern between the death ratio as the output and the selected inputs. Table 3 shows the changes in the average number of deaths and cases between the months before and after the MM order within the entire states. It can be seen that while the average number of deaths has decreased in Washington and increased in Oregon and California, the average number of

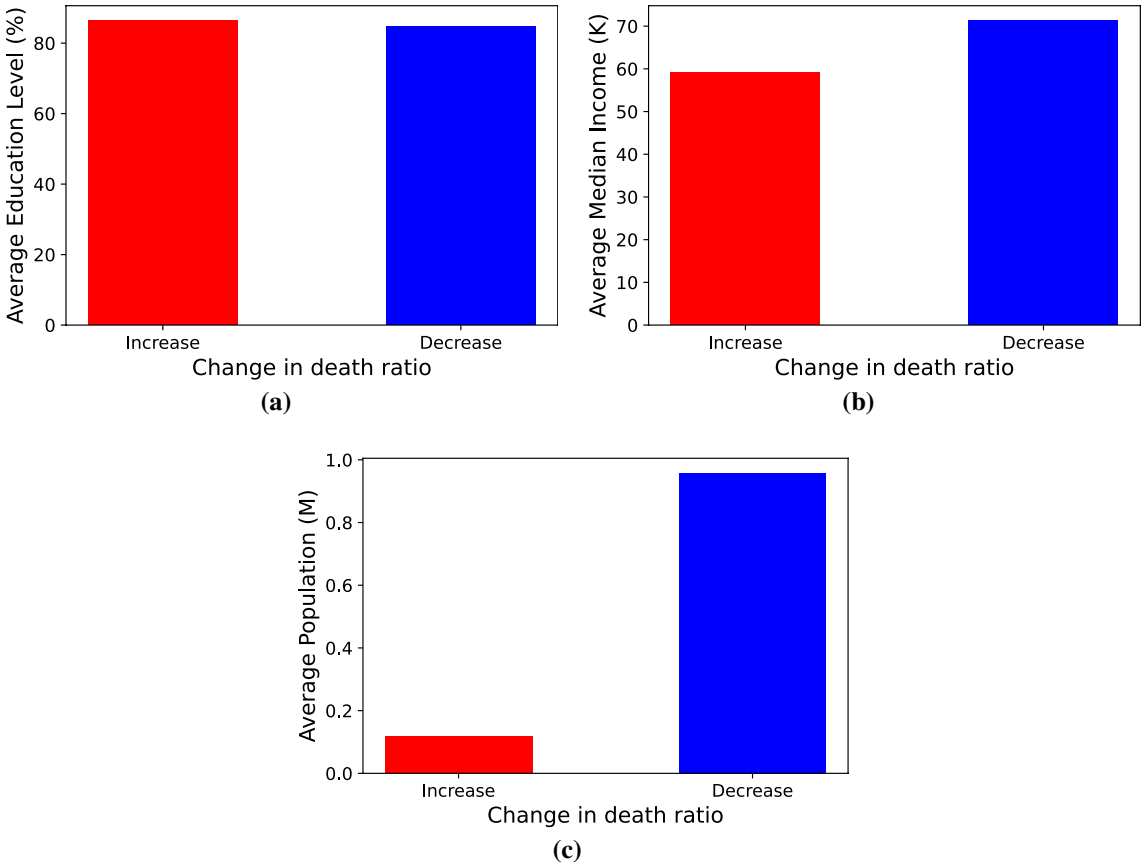

Figure 4. Visualization of the combined data for California, Oregon and Washington. Change in death ratio and average of (**a**) education level (**b**) median income and (**c**) population.

cases has increased in all of them. This implies that the observed decrease in death ratios, as reported in Table 2, can be because of the effect of face coverings in reducing the severity of COVID-19 infection.

Furthermore, to get some insight about the observed pattern in Table 2, the average percentage of people who use masks with different frequencies across all the counties experiencing both increase and decrease in their death ratios is illustrated in Fig. 6. As expected, the average percentage of people who always wear a mask (Fig. 6a) is slightly higher for the decreasing category, but in both categories, the values are the smallest for Oregon. Moreover, the average percentage of people who never use a mask (Fig. 6e) is lower for the decreasing category in all 3 states, which is also intuitively sensible. Although there are no prominent and consistent patterns for the remaining mask usage frequencies, these observations could implicitly and partially describe why there is a small increase in the overall death ratio in Oregon. However, since the mask usage data is from an NYT survey over a limited period (12 days), the observations in Fig. 6 cannot explain the entire underlying phenomenon. According to a recent study, several factors are attributing to the possibility of a person following or not following the health guidelines set by the state officials[31]. Therefore, three features among these parameters plus the aforementioned mask usage as the fourth feature have been used to conduct the current study.

All implemented algorithms in this study are capable of providing us with high classification accuracy, i.e, of predicting whether a county has experienced a decrease in its death ratio after the MM order or an increase. As provided in Table 4, it can be seen that, in general, most of the algorithms have relatively high accuracy scores for the test set. Despite the lack of sufficient training data set, Naive Bayes has an accuracy of 94%, and Random Forest, XGBoost, and Decision Tree have an accuracy of 88%. The selected hyper-parameters for XGBoost, Decision Tree, and Random Forest classifiers are shown in Table 5. The random search method has been done to tune these hyper-parameters for XGBoost, and grid search is used for Random Forest and Decicion Tree. Naive Bayes does not have any important hyper-parameter because of which, it has the capability of being generalized well. Besides, Random Forest, as a bagging method, and XGBoost, as a boosting method, have the popularity of rarely over-fitting the data. Moreover, the final hyper-parameters after tuning for the rest of the implemented algorithms are given in Table 6. Except for KNN where the elbow method is used to find the optimum number of nearest neighbors, the grid search method is applied for hyper-parameter tuning of the other methods in this table. Additionally, Table 4 includes the 95% confidence interval [denoted by CI (%)] for all algorithms. The interval is calculated based on the following:

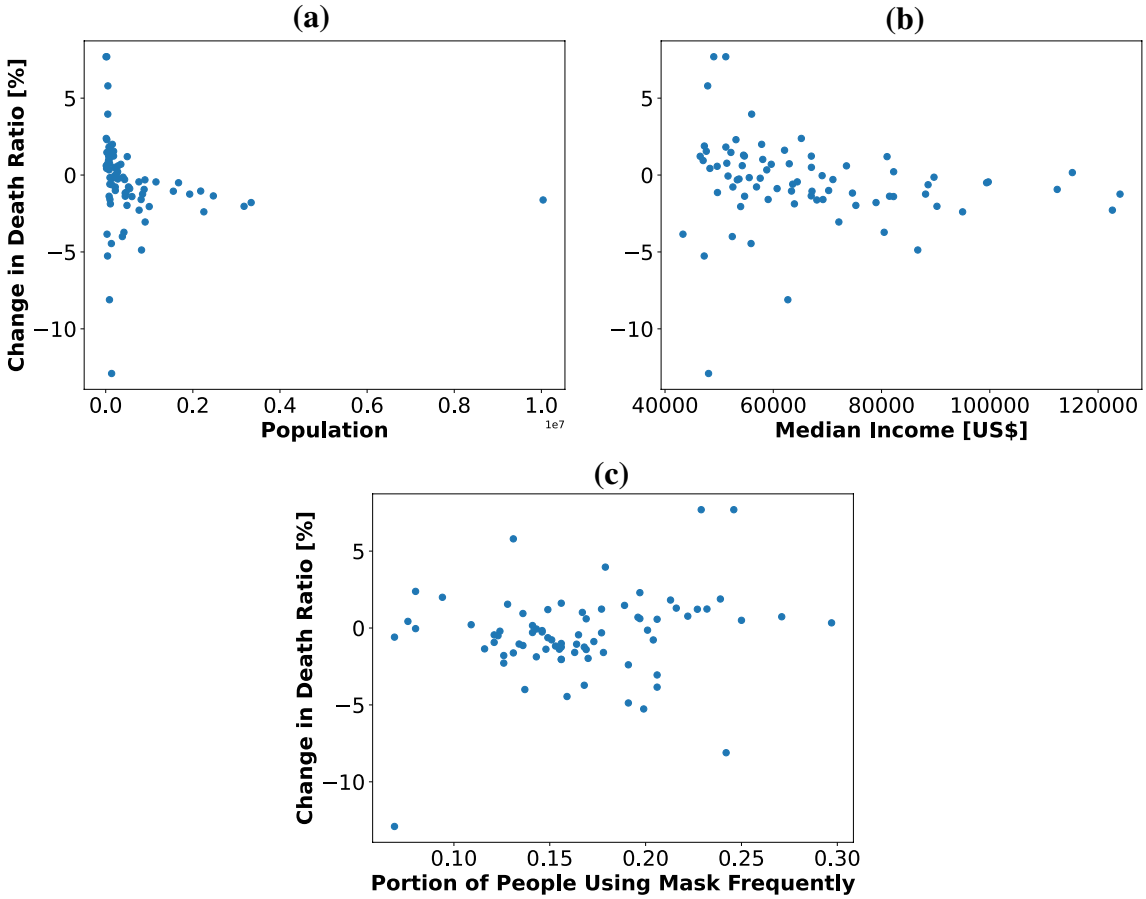

Figure 5. Scatter plot of the percentage change in the death ratio as a function of (**a**) population (**b**) median income (**c**) portion of people frequently using masks.

State	1 month before MM order	1 month after MM order	Change (%)
California	63.13	32.67	− 48
Washington	28.16	21.15	− 25
Oregon	38.03	39.14	+ 3

Table 2. Total death ratios in the month before and after the corresponding date of the mandatory face coverings executive order in each state.

State	Changes in the average cases	Changes in the average deaths
California	4210.74	9.76
Washington	379.56	− 0.82
Oregon	172.85	0.75

Table 3. Changes in the average cases and deaths between one month before and after the MM order across the entire states.

$$CI = z\sqrt{\frac{score(1-score)}{n_{test}}}, \tag{1}$$

where z is the number of standard deviations from the Gaussian distribution and equals to 1.96 for 95% CI, score is the classification accuracy of the algorithm, and n_{test} is the number of test points in our dataset. As expected, we see this interval becomes narrower as the test set accuracy increases. By taking a closer look at the accuracy

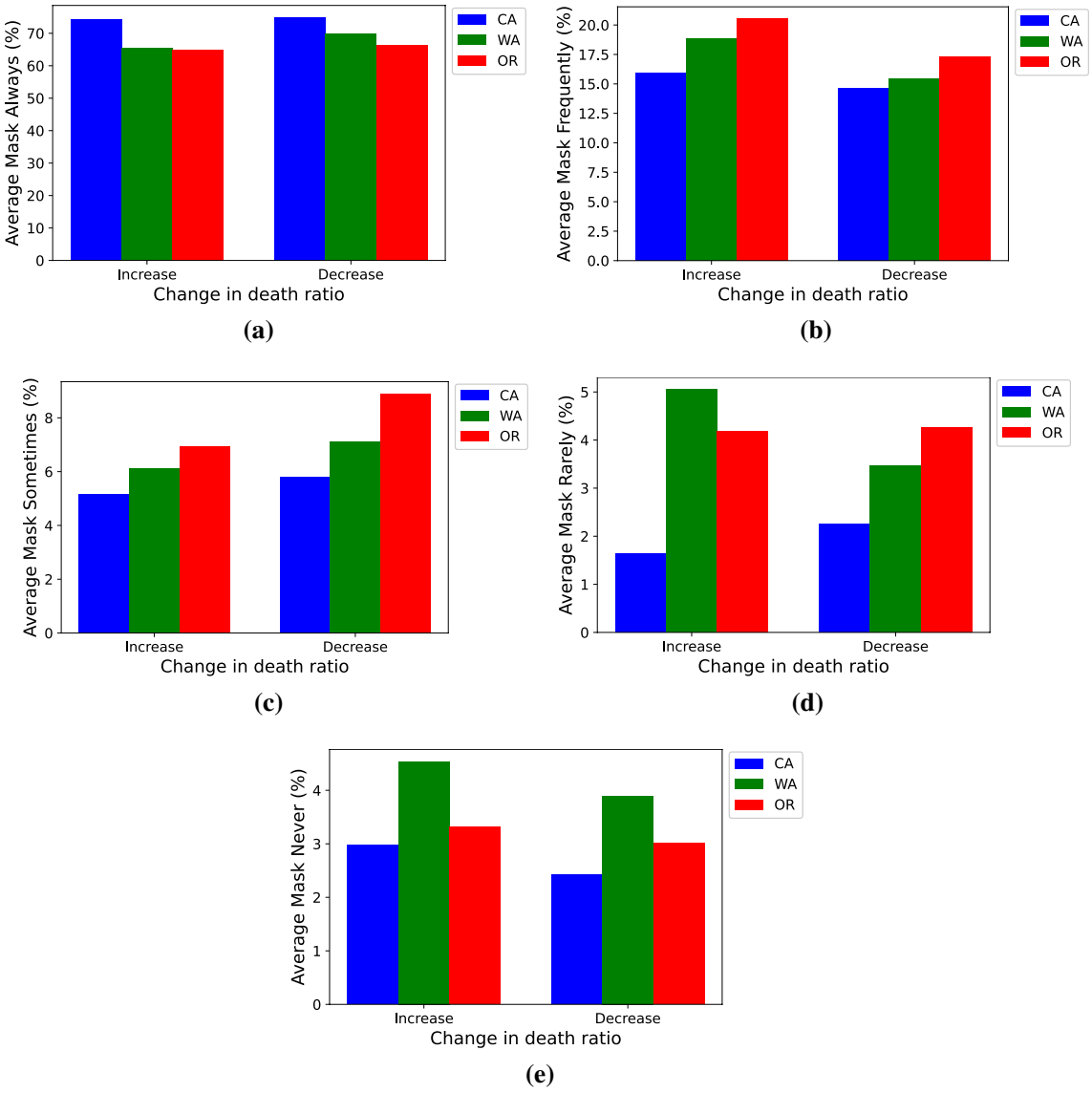

Figure 6. Average percentage of people who (**a**) always, (**b**) frequently, (**c**) sometimes, (**d**) rarely, and (**e**) never use mask across all the counties experiencing an increase and decrease in their death ratios.

Algorithm	Test (%)	Train (%)	CI (%)
Support Vector Machine	69	74	23
Extra Trees	75	93	21
KNN	81	75	19
Logistic Regression	81	79	19
Neural Net	81	80	19
Decision Tree	88	93	16
Random Forest	88	85	16
XGBoost	88	95	16
Naive Bayes	94	70	12

Table 4. Performance metrics for all the studied algorithms.

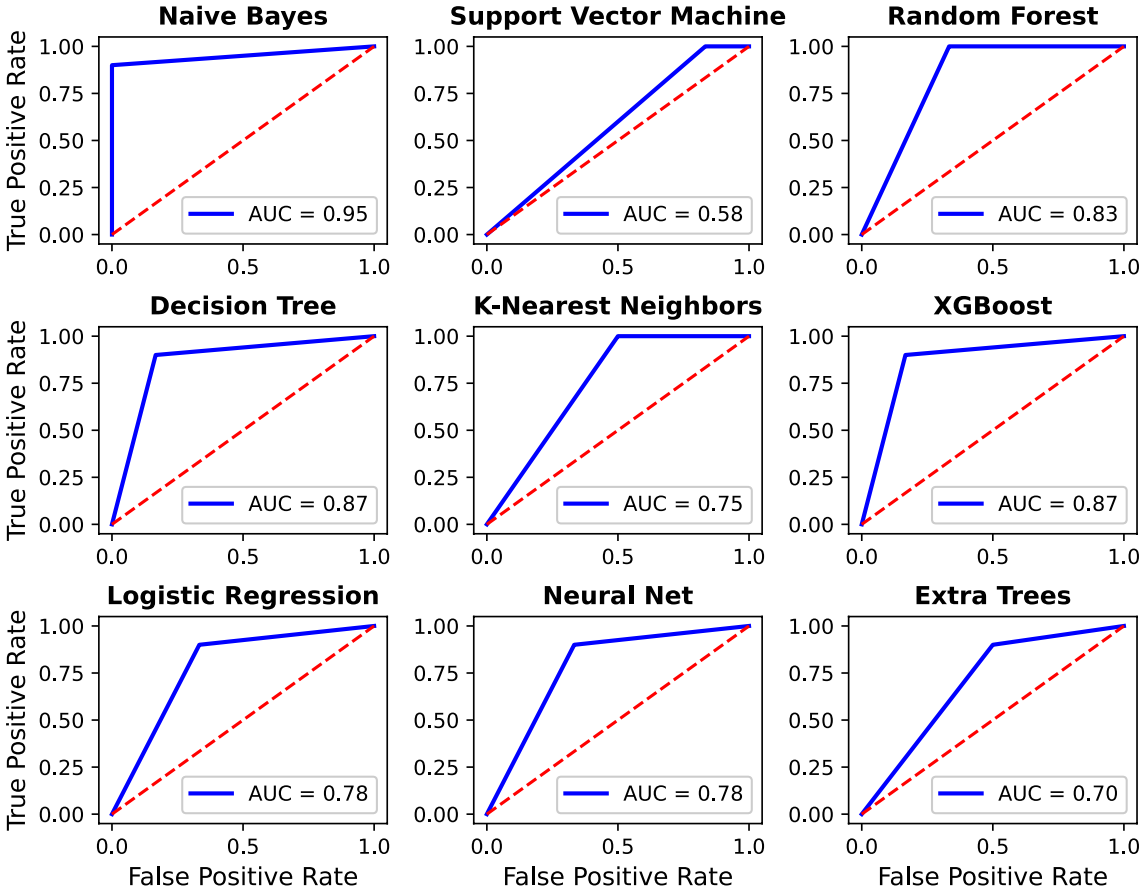

Figure 7. ROC plots of all the algorithms.

Extreme Gradient Boosting(XGBoost)					
CSbT	G	LR	MD	NE	S
0.9605	0.4735	0.0975	4	119	0.6232
Random Forest					
MD	MF	MSS	NE		
7	2	2	10		
Decision Tree					
MD	CR	MSS			
4	Gini	2			

Table 5. Model Parameters for XGBoost, Random Forest, and Decision Tree. Columns of XGBoost—CSbT: column sample by tree ratio, G: gamma, LR: learning rate, MD: maximum depth of each tree, NE: number of estimators, S: subsamples ratio. Columns of Random Forest—MD: maximum depth of each tree, MF: maximum number of features for best split, MSS: minimum number of samples to split an internal node, NE: number of estimators. Columns of Decision Tree—MD: maximum depth of each tree, CR: criterion to measure the quality of a split, MSS: minimum number of samples to split an internal node.

scores on the train set and comparing them with those of the test set, we note that none of the implemented methods overfits the data.

Additionally, to further compare the predictive power of the algorithms, the receiver operating characteristic ROC curve is shown in Fig. 7. The ROC curve shows the performance of a classification model at all classification thresholds. The diagonal red dashed lines demonstrate the no-discrimination line, which corresponds to the values of a random guess. As evident, for all algorithms, the ROC curve is above the line of no-discrimination. The area under ROC Curve (AUC) values demonstrate an aggregate measure of performance across all possible classification thresholds. In other words, the AUC values measure how accurate the model predictions are, and the values close to 1 are desirable. The measured AUC values, as shown in Fig. 7, suggest that the Naive Bayes

Neural Network				
Activation	**Learning rate**	**Neurons**	**Layers**	**Epochs**
ReLU	0.01	32	1	50
KNN				
Nearest Neighbors Number				
8				
Logistic Regression				
C (penalty term)		**Regularization Norm**		
1000		L2		
Extra Trees				
Criterion	**Min. Samples Split**	**Min. Samples Leaf**	**Estimator Count**	
Entropy	2	2	200	
Support Vector Machine				
C (penalty coefficient)	**Kernel**			
1	Radial Basis Function			

Table 6. Model parameters for the remaining methods.

algorithm leads to the best prediction. The AUC values are also in agreement with the testing accuracy, as shown in Table 4, where also the Naive Bayes algorithm has the highest testing accuracy.

The trend of high accuracy on test data signifies the existence of a pattern between the chosen features and the change in death ratio in the proposed model. Moreover, against the common belief that highly populated areas might experience harsher effects of COVID-19, on the west coast of the United States, the areas with lower populations endured worse conditions. Additionally, such a modeling approach could be used to optimize the distribution of services and media coverage for possible future adversities. A possible solution for decreasing the effect of future pandemics such as COVID-19 would be improving media coverage and public knowledge, especially in more vulnerable areas.

Conclusion

In this body of work, we have analyzed the effect of mask covering on the intensity of spread of the COVID-19 virus by considering the death ratio at the county level to be the primary indicator. To bridge the gap between level of adherence to mask mandate, we use four main features as input data: population, income, education level, and the survey results on mask usage from the New York Times. The change in the death ratio is used as the metric to quantify the effectiveness of face-coverings on the COVID-19 spread. After extracting and refining the data-set from reliable sources, we analyzed the information using nine different algorithms. Among all the methods used, Random Forest, XGBoost, Decision Tree, and Naive Bayes had the best performance with a classification accuracy of around 90%. Such a high accuracy shows the legibility of chosen features as influential criteria for modeling purposes. The obtained hyper-parameters for these models, along with the selected features, can now be used to predict future conditions of the spread of the virus.

The results show a connection between adherence to the mask mandate and change in death ratio in the majority of studied counties. The findings of this work emphasize the potential role that the immediate legislative action can play in improving the society's well-being during pandemics. It is hoped that the results of this work could further clarify the importance of preventive measures such as mask mandate order and highlight the importance of socioeconomic conditions on the behavior of different communities, which could be complex and counter-intuitive. However, it is important to note that the results we presented here are valid only for a specific geographic location, which in this study was the West Coast of the United States. Any generalization of our findings should be interpreted according to the overarching guidelines and applicable studies.

Received: 22 May 2021; Accepted: 18 October 2021
Published online: 04 November 2021

References

1. Centers for Disease Control and Prevention. Previous U.S. covid-19 case data. https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/previouscases.html.
2. Shakil, M. H., Munim, Z. H., Tasnia, M. & Sarowar, S. Covid-19 and the environment: A critical review and research agenda. *Sci. Total Environ.* **1**, 141022 (2020).
3. Aiello, A. E. *et al.* Research findings from nonpharmaceutical intervention studies for pandemic influenza and current gaps in the research. *Am. J. Infect. Control* **38**(4), 251–258 (2010).
4. Saunders-Hastings, P., Crispo, J. A., Sikora, L. & Krewski, D. Effectiveness of personal protective measures in reducing pandemic influenza transmission: A systematic review and meta-analysis. *Epidemics* **20**, 1–20 (2017).
5. Brienen, N. C., Timen, A., Wallinga, J., Van Steenbergen, J. E. & Teunis, P. F. The effect of mask use on the spread of influenza during a pandemic. *Risk Anal.* **30**(8), 1210–1218 (2010).
6. Xiao, J. *et al.* Nonpharmaceutical measures for pandemic influenza in nonhealthcare settings: Personal protective and environmental measures. *Emerg. Infect. Dis.* **26**(5), 967 (2020).
7. Cowling, B., Zhou, Y., Ip, D., Leung, G. & Aiello, A. E. Face masks to prevent transmission of influenza virus: A systematic review. *Epidemiol. Infect.* **138**(4), 449–456 (2010).

8. Li, T., Liu, Y., Li, M., Qian, X. & Dai, S. Y. Mask or no mask for covid-19: A public health and market study. *PLoS ONE* **15**(8), e0237691 (2020).
9. Cheng, V. C. *et al.* The role of community-wide wearing of face mask for control of coronavirus disease 2019 (covid-19) epidemic due to sars-cov-2. *J. Infect.* **1**, 1–10 (2020).
10. Caddy, S. Coronavirus: does the amount of virus you are exposed to determine how sick you'll get?. https://theconversation.com/coronavirus-does-the-amount-of-virus-you-are-exposed-to-determine-how-sick-youll-get-135119.
11. Granados, A., Peci, A., McGeer, A. & Gubbay, J. B. Influenza and rhinovirus viral load and disease severity in upper respiratory tract infections. *J. Clin. Virol.* **86**, 14–19 (2017).
12. Martin, E. T., Kuypers, J., Heugel, J. & Englund, J. A. Clinical disease and viral load in children infected with respiratory syncytial virus or human metapneumovirus. *Diagn. Microbiol. Infect. Dis.* **62**(4), 382–388 (2008).
13. Houben, M. *et al.* Disease severity and viral load are correlated in infants with primary respiratory syncytial virus infection in the community. *J. Med. Virol.* **82**(7), 1266–1271 (2010).
14. DeVincenzo, J. P., El Saleeby, C. M. & Bush, A. J. Respiratory syncytial virus load predicts disease severity in previously healthy infants. *J. Infect. Dis.* **191**(11), 1861–1868 (2005).
15. Liu, Y., Liao, W., Wan, L., Xiang, T. & Zhang, W. Correlation between relative nasopharyngeal virus rna load and lymphocyte count disease severity in patients with covid-19. *Viral Immunol.* **34**, 330–335 (2020).
16. Fajnzylber, J. *et al.* Sars-cov-2 viral load is associated with increased disease severity and mortality. *Nat. Commun.* **11**(1), 1–9 (2020).
17. He, X. *et al.* Temporal dynamics in viral shedding and transmissibility of covid-19. *Nat. Med.* **26**(5), 672–675 (2020).
18. Hou, Y. J. *et al.* Sars-cov-2 reverse genetics reveals a variable infection gradient in the respiratory tract. *Cell* **182**(2), 429–446 (2020).
19. Huang, H. *et al.* Integrated single-cell atlases reveal an oral sars-cov-2 infection and transmission axis. *MedRxiv* (2020).
20. Goh, Y., Tan, B. Y., Bhartendu, C., Ong, J. J. & Sharma, V. K. The face mask how a real protection becomes a psychological symbol during covid-19?. *Brain Behav. Immunity* **88**, 1–5 (2020).
21. Sharma, S. K., Mishra, M. & Mudgal, S. K. Efficacy of cloth face mask in prevention of novel coronavirus infection transmission: A systematic review and meta-analysis. *J. Educ. Health Promot.* **9**, 1–10 (2020).
22. Lalmuanawma, S., Hussain, J. & Chhakchhuak, L. Applications of machine learning and artificial intelligence for covid-19 (sars-cov-2) pandemic: A review. *Chaos Solitons Fract.* **1**, 110059 (2020).
23. Agrebi, S. & Larbi, A. Use of artificial intelligence in infectious diseases. in *Artificial Intelligence in Precision Health*, pp. 415–438 (Elsevier, 2020).
24. Ribeiro, M. H. D. M., da Silva, R. G., Mariani, V. C. & dos Santos Coelho, L. Short-term forecasting covid-19 cumulative confirmed cases: Perspectives for brazil. *Chaos Solitons Fract.* **1**, 109853 (2020).
25. Yan, L. *et al.* An interpretable mortality prediction model for covid-19 patients. *Nat. Mach. Intell.* **1**, 1–6 (2020).
26. Malki, Z. *et al.* Association between weather data and covid-19 pandemic predicting mortality rate: Machine learning approaches. *Chaos Solitons Fract.* **138**, 110137 (2020).
27. Shrivastav, L. K. & Jha, S. K. A gradient boosting machine learning approach in modeling the impact of temperature and humidity on the transmission rate of covid-19 in India. *Appl. Intell.* **1**, 1–13 (2020).
28. Maloney, M. J., Rhodes, M. J. & Yarnold, P. R. Mask mandates can limit covid spread: Quantitative assessment of month-over-month effectiveness of governmental policies in reducing the number of new covid-19 cases in 37 us states and the district of columbia. *MedRxiv* (2020).
29. Lennon, R. P. *et al.* Public intent to comply with covid-19 public health recommendations. *HLRP: Health Literacy Res. Pract.* **4**(3), e161–e165 (2020).
30. Sathianathan, S. *et al.* Knowledge, perceptions, and preferred information sources related to covid-19 among healthcare workers: Results of a cross sectional survey. *Am. J. Health Promot.* **1**, 982416 (2020).
31. Weiss, B. D. *et al.* Disparities in adherence to covid-19 public health recommendations. *HLRP: Health Literacy Res. Pract.* **4**(3), e171–e173 (2020).
32. N. Y. Times. Mask-wearing survey data. https://github.com/nytimes/covid-19-data/tree/master/mask-use.
33. Markowitz, A. State-by-state guide to face mask requirements. https://www.aarp.org/health/healthy-living/info-2020/states-mask-mandates-coronavirus.html.
34. USAFACTS. Usa coronavirus cases and deaths. https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/state/oregon.
35. USCensus. United states census bureau. https://www.census.gov.
36. USCensus. United states census bureau. https://www.census.gov/library/visualizations/interactive/2014-2018-median-household-income-by-county.html.
37. Ayyadevara, V. K. *Pro Machine Learning Algorithms* (Apress, 2018).
38. Richert, W. *Building Machine Learning Systems with Python* (Packt Publishing Ltd, 2013).
39. Steinki, O. & Mohammad, Z. Introduction to ensemble learning. SSRN 2634092 (2015).
40. Liang, W., Luo, S., Zhao, G. & Wu, H. Predicting hard rock pillar stability using gbdt, xgboost, and lightgbm algorithms. *Mathematics* **8**(5), 765 (2020).
41. Gad, I. & Hosahalli, D. A comparative study of prediction and classification models on ncdc weather data. *Int. J. Comput. Appl.* **1**, 1–12 (2020).
42. Priyanka, & Kumar, D. Decision tree classifier: A detailed survey. *Int. J. Inf. Decis. Sci.* **12**(3), 246–269 (2020).
43. Deng, L. & Liu, Y. *Deep Learning in Natural Language Processing* (Springer, 2018).
44. Khuzani, A. Z., Heidari, M. & Shariati, S. A. Covid-classifier: An automated machine learning model to assist in the diagnosis of covid-19 infection in chest x-ray images. *Sci. Rep.* **11**(1), 1–6 (2021).

Author contributions

A.L., M.B., S.Z., and V.R.H. designed the project. A.L., M.B., and S.Z. collected the data, and wrote the original version of the manuscript. A.L., M.B., S.Z., and N.M. wrote the classifiers and implemented the machine learning code. All authors contributed to the writing of the manuscript and discussion of results. V.R.H. supervised the project.

Funding

Publication of this article was funded in part by Purdue University Libraries Open Access Publishing Fund.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.L.