mathematics

MDPI

Article

COVID-19 Mortality Prediction Using Machine Learning-Integrated Random Forest Algorithm under Varying Patient Frailty

Erwin Cornelius [1], Olcay Akman [1,*] and Dan Hrozencik [2]

[1] Department of Mathematics, Illinois State University, Normal, IL 61701, USA; edcorne@ilstu.edu
[2] Department of Mathematics, Chicago State University, Chicago, IL 60628, USA; dhrozenc@csu.edu
* Correspondence: oakman@ilstu.edu

Abstract: The abundance of type and quantity of available data in the healthcare field has led many to utilize machine learning approaches to keep up with this influx of data. Data pertaining to COVID-19 is an area of recent interest. The widespread influence of the virus across the United States creates an obvious need to identify groups of individuals that are at an increased risk of mortality from the virus. We propose a so-called clustered random forest approach to predict COVID-19 patient mortality. We use this approach to examine the hidden heterogeneity of patient frailty by examining demographic information for COVID-19 patients. We find that our clustered random forest approach attains predictive performance comparable to other published methods. We also find that follow-up analysis with neural network modeling and k-means clustering provide insight into the type and magnitude of mortality risks associated with COVID-19.

Keywords: machine learning; random forest; neural network

check for updates

Citation: Cornelius, E.; Akman, O.; Hrozencik, D. COVID-19 Mortality Prediction Using Machine Learning-Integrated Random Forest Algorithm under Varying Patient Frailty. *Mathematics* **2021**, *9*, 2043. https://doi.org/10.3390/math9172043

Academic Editor: Alfonso Mateos Caballero

Received: 19 July 2021
Accepted: 16 August 2021
Published: 25 August 2021

1. Introduction

Healthcare in the United States is a constantly and rapidly evolving industry. The extensive branches of the industry make the collection and integration of data an important endeavor. From development and testing of medical treatments or pharmaceuticals to insurance risk assessment and premium pricing, a wide variety of data is necessary for use by industry professionals.

A general trend in the current world of information is an increasing abundance and availability of data. This creates a need for the implementation of methods that can better process this abundance of data and turn it into useful information. This ever-evolving challenge requires continual development and integration of data analytics tools. Consequently, machine learning (ML) approaches have become a topic of interest [1–7].

Unquestionably, the most significant challenge to face the healthcare industry in the United States over the past year has been COVID-19. Since the first confirmed case on 21 January 2020, the United States has seen over 32 million COVID-19 cases that have led to over 578,000 deaths [8]. Predicting the spread and severity of COVID-19 is an obvious area of interest.

Important in the study of COVID-19 are those individuals at particular risk. According to Aizenman [9], treatment of individuals with compromised immune systems may shape the way the virus runs its course:

> "There's mounting research to suggest that protecting people who are immuno-compromised from getting COVID is important not just for their sake—it could be critical in the effort to end the pandemic for everyone" [9].

For example, immuno-compromised individuals may not react to treatments in the same manner. In examining HIV-positive individuals who had been vaccinated against COVID-19, Dr. Laura McCoy, an infectious disease researcher at University College in

London, remarked: "we couldn't actually see any measurable levels of anti-coronavirus antibody in the blood" [9]. These immuno-compromised individuals may also create avenues of risk for the general population. Dr. Salim Abdool Karim of South Africa's Centre for the AIDS Programme of Research noted that the increased time immuno-compromised individuals may be infected with COVID-19 creates added risk, noting that a particular HIV-positive woman "became a cauldron for the creation of a whole lot of new variants" [9]. It is apparent that identifying certain groups at an increased risk from COVID-19 could be important in curtailing the spread of the virus.

To aid in evaluating and modeling the course of COVID-19, ML methods have been employed. Wang and Wong [10] utilized a neural network approach on chest X-ray images to identify COVID-19 cases. Pal et al. [11] developed country-specific neural network models to determine a country's COVID-19 risk. In Liu et al. [12], the authors use ML approaches to predict the spread of COVID-19 in Chinese provinces. The authors in Beck et al. [13] propose a list of drugs to possibly combat COVID-19 identified from a deep learning model. A generative adversarial network was built by Khalifa et al. [14] to identify COVID-19 linked pneumonia from chest X-ray images. Sujath et al. [15] attempt to forecast the spread of COVID-19 in India using multilayer perceptron and vector autoregression methods.

Recent works include Pourhomayoun et al. [16], who utilized several ML methods to identify priority COVID-19 patients. Karthikeyan et al. [17] used ML methods to predict COVID-19 patient mortality based on blood samples. Kar et al. [18] collected information on admitted patients and used ML methods to predict COVID-19 patient mortality.

The random forest (RF) is an approach in ML that has shown merit in accurately modeling COVID-19 outcomes. The RF approach utilizes an ensemble of weak decision tree classifiers to make predictions. The RF approach has been used somewhat commonly to model the COVID-19 pandemic. Tang et al. [19] built a RF model to predict the severity of a COVID-19 diagnosis from chest CT images. Barbosa et al. [20] utilized a RF classifier to diagnose COVID-19 from blood samples. Gupta et al. [21] predicted cases of COVID-19 in India with a RF model. The RF model was also employed by Yesilkanat [22] to predict COVID-19 case counts at the country level. An effort to predict COVID-19 health outcomes from demographic and medical information in Korea was made by the authors in An et al. [23]. A similar effort to predict COVID-19 health outcomes for patients in Wuhan using RF methods was made by Wang et al. [24]. Majhi et al. [25] also utilized a RF model to predict COVID-19 case counts. Iwendi et al. [26] set out with the goal of identifying a modeling approach that worked best for COVID-19 patient health outcome prediction. The authors utilized the RF method with the AdaBoost boosting algorithm to predict COVID-19 mortality at the individual level. Demographic information for individuals that contracted the virus was used to predict if a patient would die. The authors compared several common modeling approaches and determined that the boosted RF model was the best.

While the RF method is generally praiseworthy for its prediction performance, it does have limits in explaining why certain predictions were made. The RF method does not assume a structure for the data, so it can be difficult to assess how and why certain features of the data affect predictive performance.

One challenge with COVID-19 is the somewhat varied effects that it can have on an individual. Whether it be individuals in different age groups or races, it is apparent that COVID-19 affects certain groups quite differently [27]. To this end, it could be beneficial to explore the differing risks that individuals have in the context of exposure to COVID-19. In the aforementioned work on COVID-19 patient health prediction, the models were developed using the full body of selected data. This could be in part due to the lack of available data. Given the differing effects of COVID-19 on certain demographic groups, an exploration into modeling approaches that cater to these different effects seems worthy of investigation.

This idea of using individual patient characteristics to predict health outcomes underscores a current research trend in the healthcare industry: Personalized medicine [28–31]. The interest here is to use information about an individual to predict health outcomes.

To better predict a patient's health outcome, the main hook with an individualized model is the discernment with which a predictive model is built. Rather than using a predictive model that has been built using all available patients, a personalized model would be built using only those patients who are similar to the individual that is being assessed.

Building a model using only individuals similar to a patient in question obviously necessitates a way to define patient similarity. As the success of a personalized model may be heavily dependent on selecting appropriate individuals from which to build the model, appropriately defining a similarity measure is an important step.

Some different approaches have been utilized to measure patient similarity. Park et al. [32] used a so-called statistical case-based reasoning approach to identify similar patients. A clustering approach was used to identify similar patients by Panahiazar et al. [33]. A propensity score approach was utilized by Brookhart et al. [34] to define patient similarity.

Interestingly, a RF methodology can be employed to define a similarity measure. Lee [35] used RF methods to develop a patient similarity measure for ICU patients. Running a RF model in an unsupervised manner (with no prediction goal in mind) gives a similarity score between individuals. This similarity measure allowed a RF model to be developed for individual patients using only similar patients in the model-building process. The author found that the RF methods employed performed better than other common models to predict ICU patient mortality.

This paper intends to explain and implement a method to predict COVID-19 patient health outcomes using RF methods while concurrently giving insight into the different risks that different demographic groups face from COVID-19. We will use the boosted RF model shown by Iwendi et al. [26] to be a reasonable standard for addressing prediction performance for COVID-19 patient mortality as a benchmark for predictive performance. We will also incorporate the RF patient-similarity measure utilized by Lee [35] as a tool to examine the qualities that affect COVID-19 risk. A neural network model and k-means clustering will both be used as methods for follow-up work to better understand the conclusions drawn from the RF modeling. This work will show that a RF model that utilizes preliminary clustering by patient similarity can achieve prediction performance competitive with other methods while also providing qualitative information about feature effects that are typically hard to get at with RF methods.

2. Materials and Methods

2.1. Software

Modeling for this project was performed in R version 3.6.1. The randomForest [36] R package was used for building RF models. The adabag [37] and cluster [38] packages were used for implementing the AdaBoost algorithm and cluster analysis, respectively. The neuralnet [39] package was used for building the neural network model. The R code used for these analyses may be found in the Supplementary Materials section.

2.2. Data

For this project, the data were sourced from the CDC's case surveillance data [40]. This dataset contains observations of over 22 million occurrences of COVID-19 in the United States with 32 features. The observations range in time from as early as 1 January 2020 to as late as 30 March 2021. A summary of these data tells us that only about 5% of individuals observed died from COVID-19. This unbalanced distribution of surviving and dying observations is important to note for putting future evaluation metrics into context. To ease the computational burden for this project, a sample of 10,000 individuals was randomly selected from the CDC data to be used in this project. Additionally, due to the presence of many missing values for some data features, a selection of 10 features were used for our analysis. The inability of the RF approach to handle missing values forced this choice. Utilizing only observations with complete information would limit the scope of our study to only individuals in regions with very good reporting. We have attempted to limit this effect by including the features that we have a particular interest in studying.

The utilized features can be seen in Table 1. Additionally, date features were converted to numeric features to accommodate formatting for the utilized R packages.

Table 1. Features selected for analysis and their descriptions.

Feature	Description	Levels
race_ethnicity_combined	Race and Ethnicity (combined)	8 levels
cdc_case_earliest_dt	Earliest available date for record	NA
cdc_report_dt	Initial date case reported to CDC	NA
sex	Patient sex	Male, Female
onset_date	Date of symptom onset	NA
hosp_yn	Patient hospitalized	Yes, No
death_yn	Patient died	Yes, No
medcond_yn	Patient had pre-existing condition	Yes, No
res_state	State of residence	50 levels
age_group	Patient age group	8 levels

2.3. The Random Forest

Our method for predicting COVID-19 patient mortality in this project relies heavily on the random forest (RF) classifier from Breiman [41]. Consequently, a brief description of this method is appropriate.

The RF classifier is itself made up of many decision trees. A decision tree classifier is made by successively splitting our data at decision nodes according to feature values. Our initial decision node splits the data into two groups according to a cutoff value for one of the data features. Then these groups are again split by a decision node, and this process continues, building out the "branches" of the decisions tree. When the splitting stops, the last remaining groups, or "leaves" of the decision tree provide the designation for which class individuals in that group belong to. The feature used at each decision node to split the data is typically chosen so that error at that step is minimized. The RF classifier uses an ensemble "forest" of these decision trees to make its classifications.

Each tree in the RF ensemble is built using a bootstrapped random sample of the available data and considering only a random selection of available features when each splitting node is made. To classify an observation with the RF model, each decision tree in the ensemble "votes" for the class it predicts and the majority vote of the decision trees in the ensemble is the class that the RF classifier predicts. This reliance on the majority vote to classify an observation provides for better performance than a single decision tree classifier.

2.4. Boosting

The authors in Iwendi et al. [26] elect to use boosting to improve the RF prediction results. We elect not to utilize the boosting approach for our main method; however, we will utilize it for comparison and wish to give a general idea of how boosting works.

The AdaBoost method the authors use works by building a sequence of RF models with misclassified observations given more weight for each RF model in the sequence. These misclassified observations essentially train the model to give better results. Overall classifications are given with an aggregate of the model iterations.

2.5. Similarity Measure

We take a different approach to improve the RF model. Rather than utilizing a boosting approach, we employ a similarity measure between observations that will be used to cluster the data. The development of this similarity measure is given now.

As Lee [35] did, the RF classifier can create a similarity measure if used in an unsupervised manner. This is done by building fake observations from the available data. The RF classifier is then used to classify the data as real or false. A similarity measure between any two observations from the real dataset is taken to be the proportion of the time those two observations ended up in the same leaf of a decision tree in the RF ensemble. From Lee [35], we may then denote the similarity between a patient i and a patient j, call it $SIM(i,j)$, by

$$SIM(i,j) = \frac{Prox_{i,j}}{\Sigma_{k=1}^{N} Prox_{i,k}}, i = 1, \ldots, N, j = 1, \ldots, N, i \neq j, i \neq k.$$

Here $Prox_{i,j}$ is the number of trees in the RF ensemble that have both i and j in the same terminal node, or "leaf". If calculated for every pair of patients, we can develop a similarity measure for all considered observations. This similarity measure may be thought of as the proportion of the time that two observations end up in the same place on a decision tree.

2.6. Clustering

This similarity measure can be used to cluster the data into similar groups. Here we describe our clustering approach as well as the method for tuning the appropriate number of clusters.

We will be using the PAM algorithm implemented with the clara function in the cluster [38] R package to group our data. The appropriate number of clusters can be tuned by looking at the average silhouette for observations in a cluster. For each observation, the silhouette width is defined as the minimum average dissimilarity of that observation compared to each other cluster. Thus a high average silhouette width indicates that a cluster is different from the other clusters. That is, a high average silhouette width indicates that clusters are distinct, as desired.

2.7. The Clustered Random Forest

We will now discuss the methodology that will be used for the analysis in this paper, which will be referred to as the "Clustered RF" method.

The clustering described earlier will be used as a preliminary step to building our model. The clustering will be performed to partition the observations according to the highest average silhouette width, and then on each cluster, a RF model will be built to predict patient mortality. The classifications for the overall model will simply be the classifications given by the RF classifier built on each cluster. In this way, classifications for patients in a cluster will be based on information from the patients similar to them also in that cluster.

2.8. Performance Measures

Given the skew of the data toward surviving observations, Iwendi et al. [26] chose to use several metrics to evaluate predictive performance. We will use those same metrics in our evaluation. These measures are based on True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) and are given now.

$$\begin{aligned}
\text{Accuracy} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \\
\text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}} \\
\text{Recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}} \\
\text{F1 Score} &= 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\end{aligned}$$

Given the small proportion of individuals who did die from COVID-19 in the available data, these measures give a better picture of how powerful a model is in correctly predicting mortality.

2.9. The Neural Network

As we will see during our discussion of the results of the clustered RF method, the added risk that we are able to derive is a somewhat crude measure that does not take into account the risk associated with membership into multiple groups simultaneously. To get at the interaction between levels of different factors, a neural network model is used on the data for follow-up analyses. A brief overview of this method is given here.

The neural network approach is described in [42]. The neural network is built by deriving new features (these features are thought of as a hidden layer) that are linear combinations of the original features that have been then passed through an activation function. Then the variable we wish to predict is modeled as a function of linear combinations of the derived features.

Fitting such a neural network model then requires the selection of the number of derived features in each hidden layer and the weights used to create the aforementioned linear combinations. Additionally, an appropriate activation function must be specified. As the neural network requires numeric inputs, the categorical features in our data were converted to numeric for the neural network model. We elected to use a model with a single hidden layer. The number of nodes in this layer was tuned from using values from 1 to 10 with logistic, tanh, and softplus activation functions considered. Using a test set of data, the combination of 5 nodes in the hidden layer with a logistic activation function yielded the best results. Consequently, this was the neural network model we employed. The takeaway here is that we will be using the neural network model to get an estimated probability that an individual with certain characteristics that we specify will die from COVID-19.

2.10. K-Means Clustering

After observing the clusters that we obtain throughout the rest of this paper, a look at how discriminatory certain features of our data are in determining which individuals are clustered together will be a nice addition. To this end, we incorporate a k-means clustering of our data. This method clusters observations so as to minimize the sum-squared distance measure of each observation to its assigned cluster [43].

3. Results

3.1. AdaBoost Data

As a test of efficacy, the clustered RF method was performed on the data used by Iwendi et al. [26] using a partition of three clusters. The target feature to predict was patient mortality. To build each cluster's RF model, the randomForest command in the same-named R package was used with ntree = 1000 and mtry = 3. For our RF methods, the ntree value was chosen to be 1000, as this is a large number for this type of method that was not a computational burden. The RF method does not have an issue with overfitting, so we do not need to be concerned about that here. The mtry value is a standard value of roughly the square root of the number of features. The evaluation metrics for the clustered RF method compared to the metrics obtained by the boosted RF method may be seen in Table 2.

Table 2. Accuracy, Precision, Recall, and F1 Score for the prediction of patient mortality on the data used by Iwendi et al. [26].

Model	Accuracy	Precision	Recall	F1-Score
Boosted RF	0.94	1	0.75	0.86
Clustered RF	0.92	0.93	0.93	0.93

3.2. CDC Data

The clustered RF and the AdaBoost RF method were then performed on the CDC case surveillance data. Four clusters were used, as this number of clusters gave the highest average silhouette width of the considered values of 2 to 8. The target feature to predict was death_yn. To build each cluster's RF model, the randomForest command in the same-named R package was used with ntree = 1000 and mtry = 3. The evaluation metrics for these models may be seen in Table 3.

Table 3. Accuracy, Precision, Recall, and F1 Score for the prediction of patient mortality on CDC case surveillance data.

Model	Accuracy	Precision	Recall	F1-Score
Boosted RF	0.95	0.59	0.40	0.48
Clustered RF	0.95	0.61	0.38	0.47

3.3. Cluster Statistics

For the four clusters that our data was partitioned into, we have broken down the proportion of individuals that fall into each level of the features and compared them to the overall makeup of the sample of 10,000. We see these in Tables 4–7.

Table 4. Percentage of individuals who fall into health categories for overall sample of 10,000 and for each of the four clusters. Given as percentages. For the first two features, the number is not a percent, but the mean value.

Cluster	cdc_report_dt	onset_dt	Pre-Existing	Hospital	Died
Overall	281.77	267.58	46.34%	16.48%	5.62%
1	282.99	269.72	50.03	15.63	4.37
2	248.60	234.94	2.34	0	0
3	145.63	99.43	100	99.38	47.84
4	400.46	399.21	0	0	0

Table 5. Percentage of individuals who fall into racial categories for overall sample of 10,000 and for each of the four clusters. AI/AN is American Indian/Alaska Native, and NH/PI is Native Hawaiian/other Pacific Islander. Given as percentages.

Cluster	AI/AN	Asian	Black	Hispanic	Multiple	NH/PI	White
Overall	0.24%	2.88%	9.14%	20.4%	3.62%	0.42%	63.3%
1	0.24	2.74	9.45	21.46	3.52	0.46	62.13
2	0	0.78	1.17	8.20	2.73	0	87.11
3	0	4.94	24.69	23.46	4.01	0	42.90
4	0	2.26	3.39	14.25	2.94	0	77.15

Table 6. Percentage of individuals who fall into age categories for overall sample of 10,000 and for each of the four clusters. Given as percentages.

Cluster	0–9	10–19	20–29	30–39	40–49	50–59	60–69	70–79	80+
Overall	3.18%	9.8%	16.48%	15.08%	14.4%	15.2%	12.6%	8.08%	5.18%
1	2.92	7.61	14.95	15.24	15.72	16.72	13.02	8.63	5.18
2	2.15	28.71	46.48	10.74	6.25	3.71	1.76	0.20	0
3	0	0.31	3.09	4.94	6.48	21	26.54	15.74	21.91
4	11.54	30.32	18.55	22.62	12.67	4.07	0.23	0	0

Table 7. Percentage of individuals who fall into sex and geographic categories for overall sample of 10,000 and for each of the four clusters. Given as percentages.

Cluster	Female	Male	Northeast	Midwest	South	West
Overall	54.12%	45.88%	17.54%	53.86%	5.34%	23.26%
1	54.53	45.47	15.85	52.68	5.81	25.66
2	54.88	45.12	9.38	89.06	0	1.56
3	42.28	57.72	77.78	12.35	8.33	1.54
4	51.36	48.64	16.29	66.74	0	16.97

3.4. Neural Network Follow-Up

A neural network model was built for the CDC case surveillance data. Using the neuralnet command in R with hidden = 3, threshold = 0.1, and all others default. A logistic activation function was used with death_yn as the prediction target, coded 0 for no death and 1 for death. As a note, the goal of this follow-up was to compare mortality rates for a variety of different individual characteristics. During the initial run, the hosp_yn feature proved to be too strong of a factor in predicting death to make the results easily interpreted. Since we want to compare relative risk of individuals with different characteristics, this feature was removed for the neural network model results we will be discussing. Using the neural network model, a selection of mortality rates were calculated for individuals with characteristics of interest. These results can be seen in Tables 8–11.

Table 8. Neural network-estimated probabilities of death for a white male with no pre-existing conditions by region and age group. Given as percentages.

Age	10–19	20–29	30–39	40–49	50–59	60–69	70–79	80+
Northeast	0.29	0.33	0.38	0.36	0.94	2.48	5.19	9.41
Midwest	0.31	0.33	0.41	0.34	0.93	2.04	3.26	5.36
South	0.02	0.36	0	0.56	3.62	21.09	8.51	28.46
West	0.32	0.33	0.48	0.33	1.21	2.49	3.05	4.32

Table 9. Neural network-estimated probabilities of death for a black male with no pre-existing conditions by region and age group. Given as percentages.

Age	10–19	20–29	30–39	40–49	50–59	60–69	70–79	80+
Northeast	0.32	0.33	0.7	0.35	2.35	5.11	6.18	8.7
Midwest	0.32	0.33	0.53	0.34	1.44	2.97	3.46	4.73
South	0.26	0.39	1.7	0.5	7.5	16.27	23.32	35.47
West	0.32	0.33	0.5	0.33	1.25	2.48	2.72	3.55

Table 10. Neural network-estimated probabilities of death for a hispanic male with no pre-existing conditions by region and age group. Given as percentages.

Age	10–19	20–29	30–39	40–49	50–59	60–69	70–79	80+
Northeast	0.3	0.34	0.59	0.36	1.92	4.44	6.59	10.44
Midwest	0.31	0.33	0.5	0.34	1.35	2.91	3.88	5.81
South	0.12	0.38	0.1	0.56	0.95	4.62	17.5	35.5
West	0.32	0.33	0.52	0.33	1.39	2.85	3.31	4.52

Table 11. Neural network-estimated probabilities of death for an asian male with no pre-existing conditions by region and age group. Given as percentages.

Age	10–19	20–29	30–39	40–49	50–59	60–69	70–79	80+
Northeast	0.29	0.34	0.67	0.38	2.51	6.14	9.07	14.68
Midwest	0.31	0.33	0.57	0.35	1.76	3.93	5.42	8.28
South	0	0.41	0	0.68	48.68	51.87	63.73	42.9
West	0.32	0.33	0.61	0.34	1.87	3.98	4.68	6.47

3.5. K-Means Follow-Up

The kmeans R command was used to cluster our data into four clusters. The number of clusters was chosen both because it is the same number used in the RF clustering and because the useful distinction amongst clusters in regards to death rate seemingly began to deteriorate when more clusters were used. We project our data onto the two-dimensional axis, seen in Figure 1. The same data plotted with the four cluster labels is seen in Figure 2. Note that the projection onto the two-dimensional plane comes from coding our categorical features as numeric (i.e., 0 for death and 1 for survivor). Consequently, the values along the axes do not have a discernible biological meaning and have been excluded. As such, these figures should be used only to get a general sense of how closely the death observations and survivor observations are clustered according to the features used in our analysis. This gives us a sense of how well the available features can discern between a death and a surviving case. Our interest in this clustering application was looking at which features of the data were most discriminatory in determining membership in a certain higher risk group. To this end, we plotted the proportion of certain feature levels by membership in the four clusters, which we see in Figures 11–14.

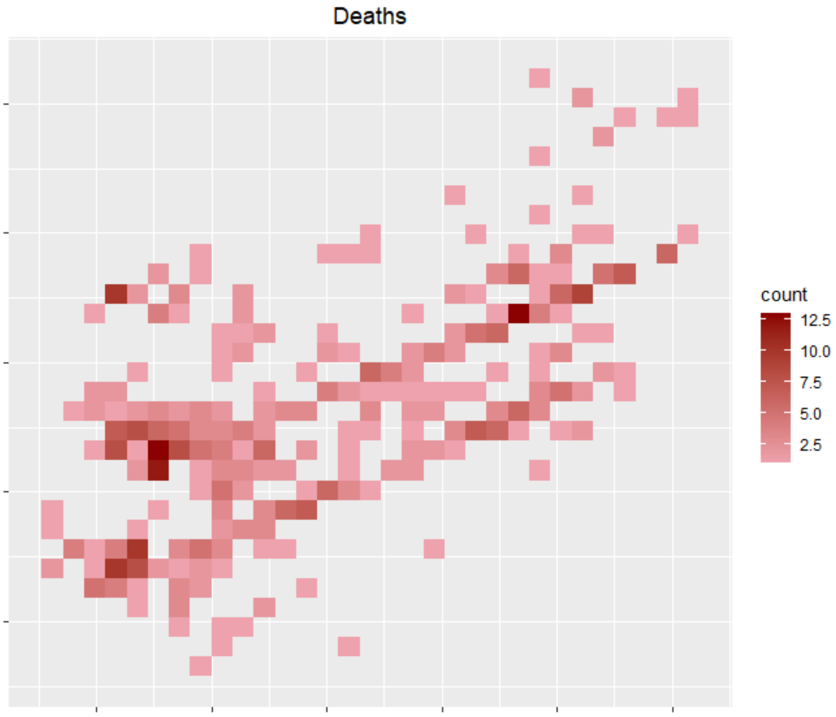

Figure 1. Cont.

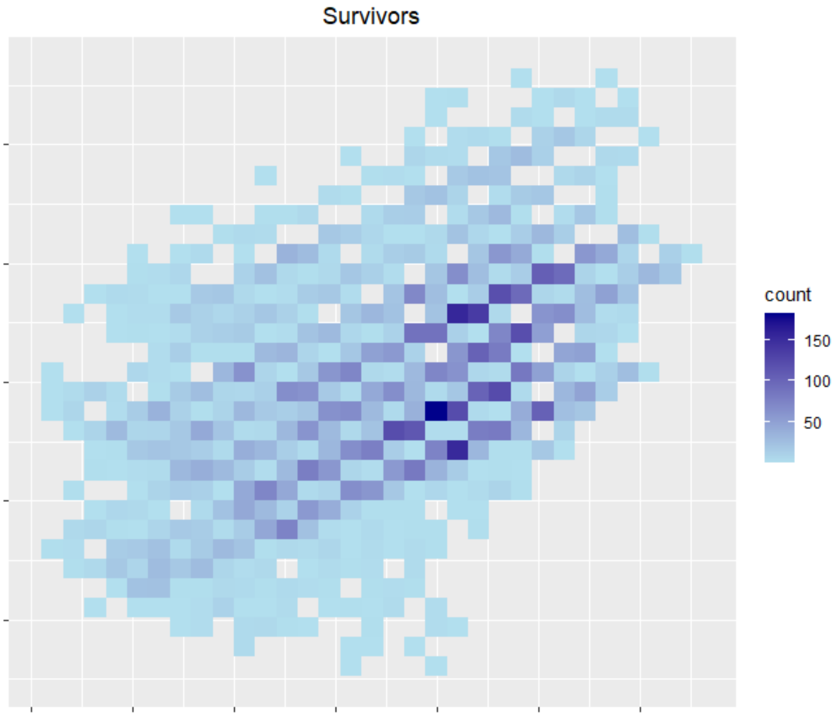

Figure 1. Density plot of observations projected onto the two-dimensional plane. Top in red are deaths, and bottom in blue are survivors. The axis values have been excluded, as they come from projecting categorical features coded as numeric onto the two-dimensional plane and do not have discernible biological meaning.

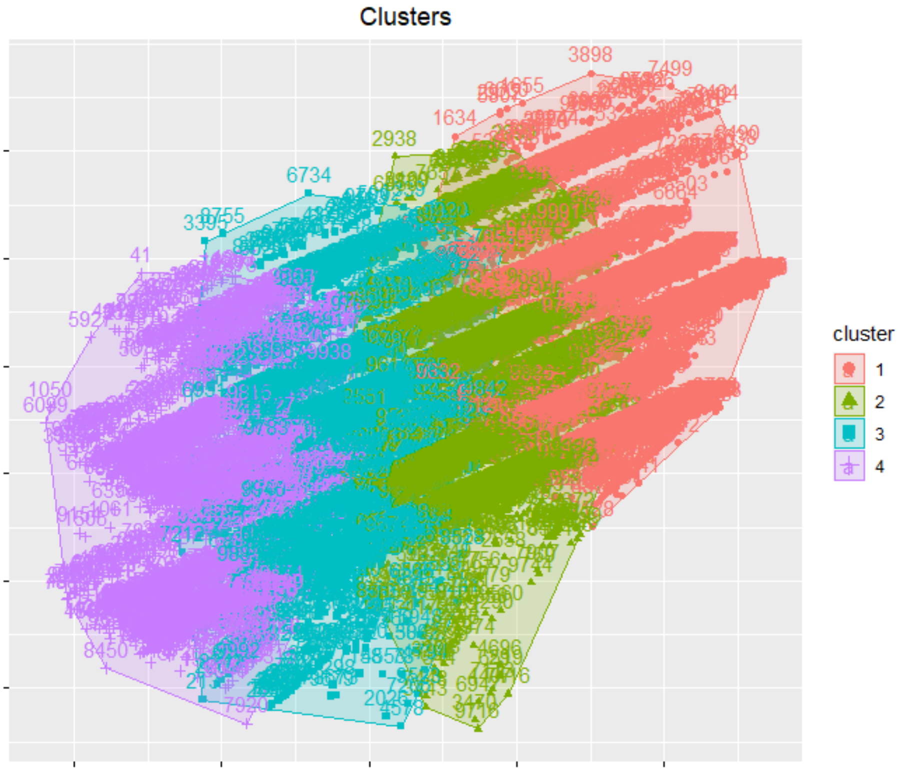

Figure 2. Density plot of observations projected onto the two-dimensional plane. Cluster 4 is the high death rate cluster. The axis values have been excluded, as they come from projecting categorical features coded as numeric onto the two-dimensional plane and do not have discernible biological meaning.

4. Discussion

4.1. Prediction Performance

Our aim in the implementation of the clustered RF is not to sacrifice predictive power. Consequently, it is important that our clustered RF approach performs comparably to the AdaBoost RF method determined by Iwendi et al. [26] to be a benchmark for COVID patient predictive models. We see from the comparison metrics in Table 2 that the clustered RF method does seem to perform comparably to the boosted RF. The Boosted RF model is superior in Accuracy and Precision, but the Clustered RF performs better in Recall and F1-score. It is fair to say that the clustered RF method is not an obvious step backwards. This gives a positive indication that the preliminary clustering can replace the boosting algorithm and give similar prediction performance. The comparison metrics for the clustered RF and the boosted RF on the CDC data seen in Table 3 yield similar results. The clustered RF seems to perform comparably to the AdaBoost method, with all compared metrics being within 0.02 of each other.

We do note that the performance of both methods on the CDC data is obviously much worse than the performance on the data from Iwendi et al. [26]. This is a bit troubling, but a deeper dive into the data used by those authors provides a possible explanation. Upon inspection of the data used, it is noteworthy that the deaths in these data have strong correlation with the patient being from a certain location. For example, 39 of the 63 reported death in these data came from China, and those 39 deaths were out of the total of 42 observations from China. It is then possible that this and other similar trends in the data created an unrealistic picture of correlation between country of origin and mortality. This has possibly created the capacity for the RF method to provide a prediction accuracy beyond what might be expected from a larger dataset, such as that provided by the CDC.

4.2. Cluster Analysis

Having established that the clustered RF method meets a reasonable standard of predictive performance, it becomes reasonable to now look at the benefit of the clustering approach. While RF methods typically provide a lesser ability to look at how certain predictors influence predictions compared to parametric methods, this clustering approach provides an avenue to gain some insights in this regard. The cluster analysis grouped our 10,000 observations into four groups. Naturally, some of these clusters will have mortality rates higher than the overall mortality rate for the 10,000 observations, and some of these clusters will have mortality rates lower than the overall. By examining how clusters vary by death rate, we can attempt to gain some information about the characteristics of individuals at a higher risk of death from COVID. These trends may be compared against trends that have been noted by the CDC to see if the insights given by the clustering RF model could be reasonably trusted.

Note that in Table 4, cluster 1 has a death rate of 4.37%, which is the closest of the four clusters to the overall death rate of 5.62%. This fits a general trend we see for cluster 1. For every feature in this project, cluster 1 stays close to the makeup of the overall sample. In no obvious way does cluster 1 distinguish itself from the whole. Consequently, we will be focusing on clusters 2–4 to gain insight into the variable risk of COVID-19. The mortality rates for individuals in clusters 2–4 are drastically different from the overall mortality rate of 5.62%. Clusters 2 and 4 did not have any deaths, while cluster 3 had a death rate of 47.84%. The natural inclination then is to view clusters 2 and 4 as low-risk clusters, and cluster 3 as a high-risk cluster with the intention of discerning characteristics of the individuals who reside in these clusters.

First, let us turn our attention to the features from Figure 3. We note that clusters 2 and 4 both had no patients that were hospitalized and a very low percentage of patients with pre-existing conditions. Cluster 3 had close to 100% of individuals hospitalized and having pre-existing conditions. This is in line with the risk of pre-existing conditions outlined by the CDC [44].

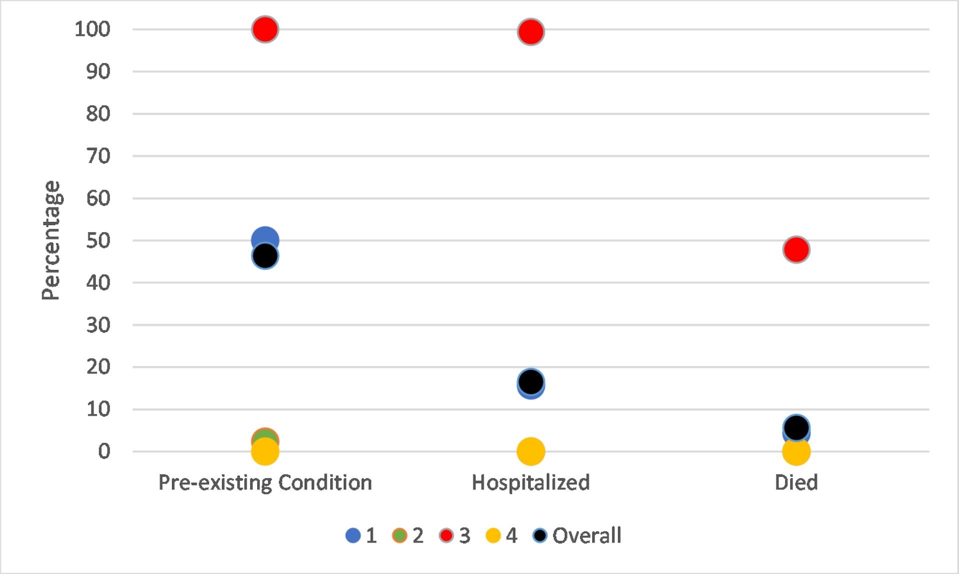

Figure 3. Percentage of individuals who fall into health categories for overall sample of 10,000 and for each of the four clusters. Same information as in last three columns of Table 4.

In Figure 4, we find the racial breakdowns for the four clusters. We can find a clear picture here. The high-risk cluster 3 is over-represented with individuals of Asian, Black, Hispanic, and multiple ethnicity and under-represented in White individuals. The low-risk clusters 2 and 4 show the opposite trend, both having more White people and fewer of all other groups. It is clear that minority groups are much more abundant in these high-risk clusters and obviously less abundant in the low-risk clusters. This is a compatible conclusion with some of the risk factors also outlined by the CDC [44].

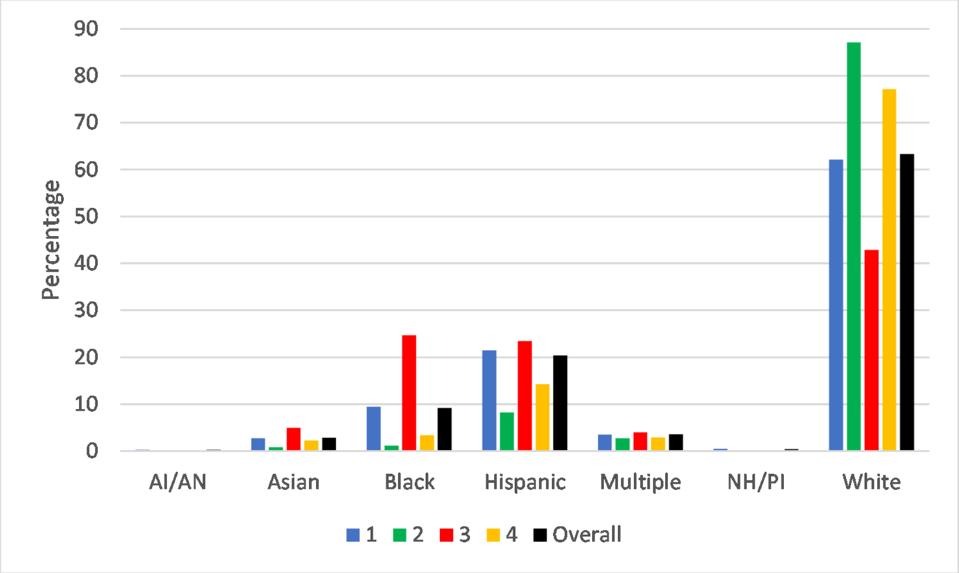

Figure 4. Percentage of individuals who fall into racial categories for overall sample of 10,000 and for each of the four clusters. AI/AN is American Indian/Alaska Native, and NH/PI is Native Hawaiian/other Pacific Islander. Same information as in Table 5.

We can also look at the age demographics in Figure 5. A general trend to note is that clusters 2 and 4 have a lower percentage of individuals in age groups above 39 than the overall, while cluster 3 has more individuals in the age categories above 49. Cluster 3 also has fewer individuals from all age groups 49 and younger. The trends for clusters 2 and

4 are less obvious in the younger age groups, but the general trend of older individuals being at greater risk follows some of the notions put forth by the CDC [44].

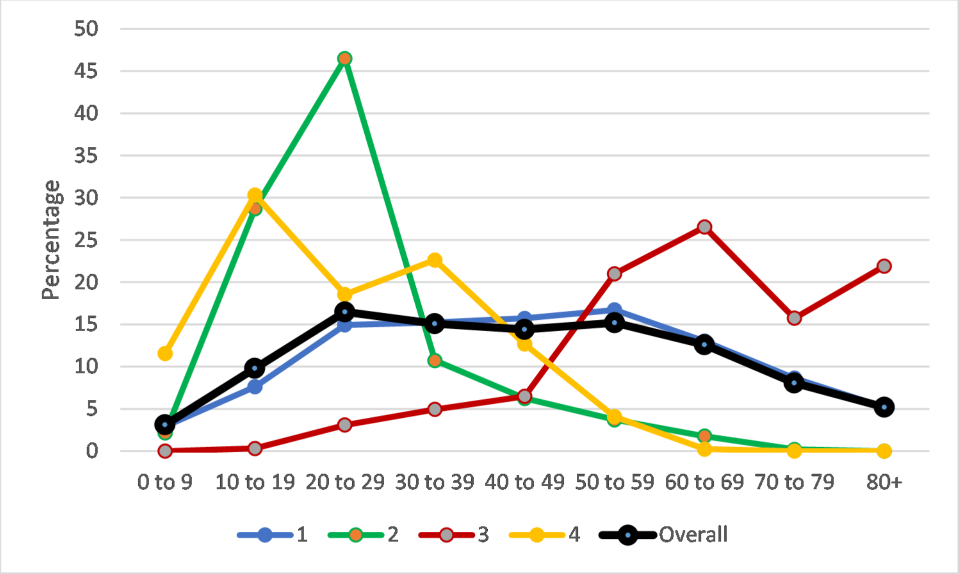

Figure 5. Percentage of individuals who fall into age categories for overall sample of 10,000 and for each of the four clusters. Same information as in Table 6.

Finally, in Figure 6, we have sex and geographic region breakdowns. Cluster 3 has relatively more males than overall and more people from the Northeast and South. There are also fewer people from the Midwest in cluster 3, while the percentage of people from the West is close to 0. Clusters 2 and 4 have sex proportions similar to the overall but noticeably have more people from the Midwest and no people from the South. Here we seem to have highlighted both being Male and being from the Northeast or South as possible added risk factors for COVID-19 mortality.

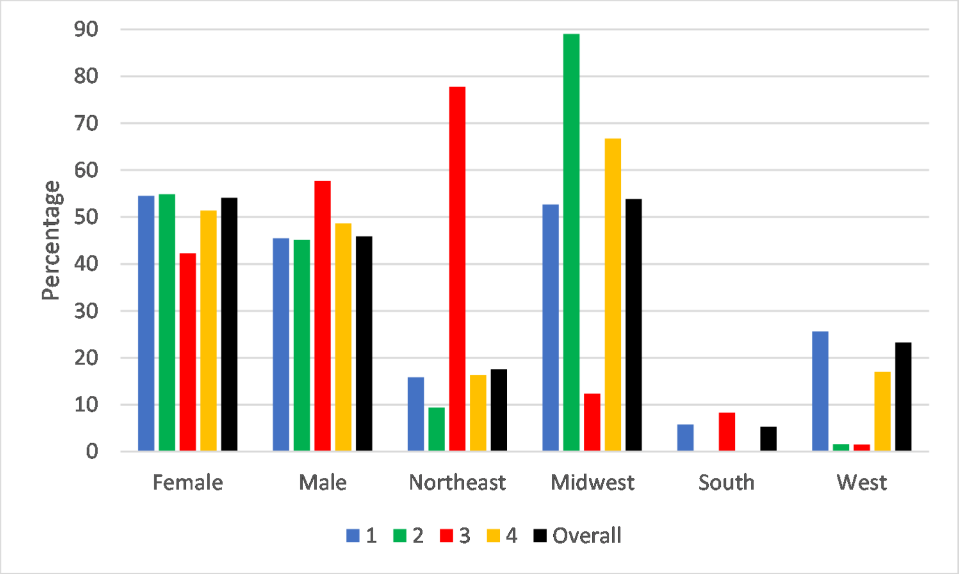

Figure 6. Percentage of individuals who fall into sex and geographic categories for overall sample of 10,000 and for each of the four clusters. Same information as in Table 7.

Given the absence of an assumed structure placed on our data by the RF method, calculating added risk factor by membership is not a straightforward task. However, we can obtain a somewhat crude measure by calculating how likely an individual is to find themselves in one of our clusters. For example, to calculate the added risk associated with being male, we look at how males are distributed among the clusters and calculate how likely a male is to fall into each cluster. We then multiply by the likelihood that an individual dies in each cluster to get how likely a male is to die. While this calculation ignores the interaction of other factors, it can give a general idea of what is happening. We look at a calculation of added risk for a selection of factors in Table 12. We see similar trends to what we might expect with minority groups and older individuals having added risk, with males and the Northeast and South also being identified as risk factors to COVID-19 mortality.

Table 12. Average added risk of mortality from COVID-19 for membership in a certain feature level, as compared to the overall mortality rate of 5.62%. Given as percentages.

Factor Level	Added Risk
Asian	+1.11%
Black	+2.41%
Hispanic	+0.13%
White	−0.84%
Female	−0.56%
Male	+0.09%
Pre-existing Condition	+1.73%
50–59 yrs	+0.54%
60–69 yrs	+1.75%
70–79 yrs	+1.50%
80+ yrs	+4.65%
Northeast	+4.70%
Midwest	−1.54%
South	+0.95%
West	−1.31%

4.3. Added Risk Follow-Up

The method we used to estimate added risk from the RF clusters has the drawback of not providing an assessment of added risk for more than a single factor level at a time. For example, the RF calculation only provided an added risk of 4.7% for living in the northeast. It would be worth investigating if this risk affected all individuals from the northeast equally or if certain groups felt the detrimental effect of living in the northeast more strongly. To this end, we performed mortality prediction using the neural network model. Specifically, we looked at risk for individuals from several racial groups by region and age group. We looked at male individuals with no pre-existing conditions to focus on the possible different effect that region would have on race and age group mortality. We have the mortality estimates for white, black, Hispanic, and Asian individuals in Figures 7–10, respectively.

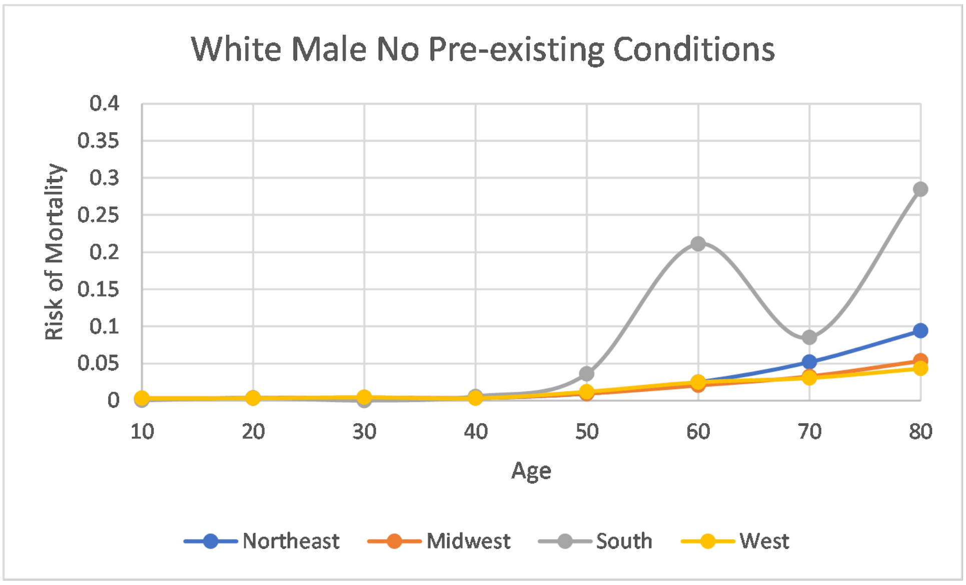

Figure 7. Probability of mortality for a white male with no pre-existing conditions for certain age groups. Given as a decimal. Same information as in Table 8.

One overarching trend is that the effect of region seems to be hidden in the earlier age groups. For all four races, the mortality rates for earlier age groups are indistinguishable by region. However, at the later age groups, we see that the mortality seems to increase for the south and northeast regions, compared to the midwest and west. We also note that for the younger age groups the higher risk for the black, Hispanic, and Asian groups that was detected earlier does not show up either. A reasonable explanation for this phenomenon would be that health outcomes for younger individuals are not dependent on medical intervention, and, consequently, any disparity in mortality due to differing quality or access to healthcare caused by membership in a racial or regional group does not show up for those younger individuals.

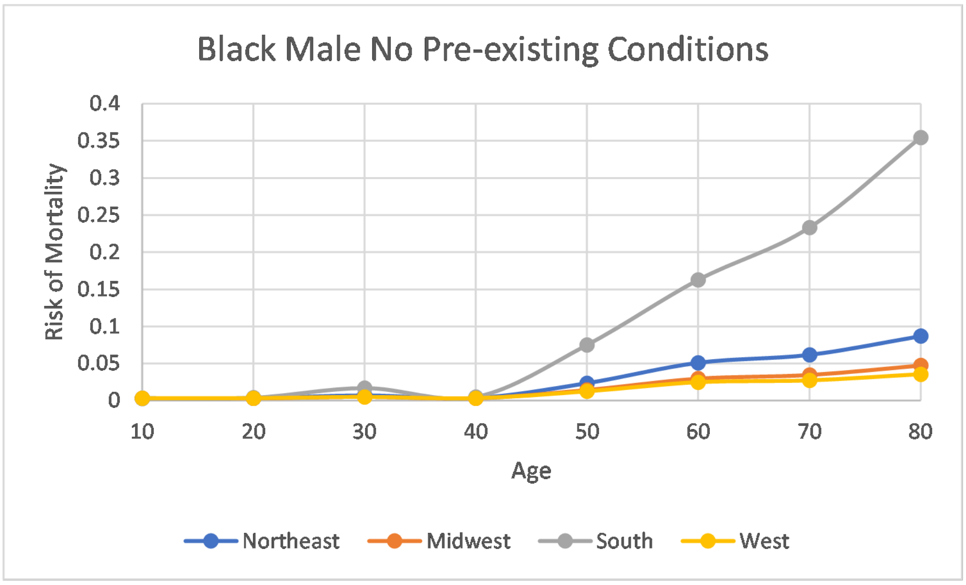

Figure 8. Probability of mortality for a black male with no pre-existing conditions for certain age groups. Given as a decimal. Same information as in Table 9.

Interestingly, we do see the influence of racial groups appear differently for the later age groups. The mortality rates for non-white individuals in the older age groups do tend to be higher than the mortality rates for white individuals in the same age groups.

Additionally, the detriment associated with living in the northeast or south takes effect earlier in the non-white race groups. For example, a white individual does not see the northeast detriment until the age 70–79 age group, while the other racial groups see this departure start to occur in the 50–59 or 60–69 age groups. We also see that Hispanic individuals do not have especially worse mortality risk for living in the south until the 70–79 age group, contrary to the other age groups.

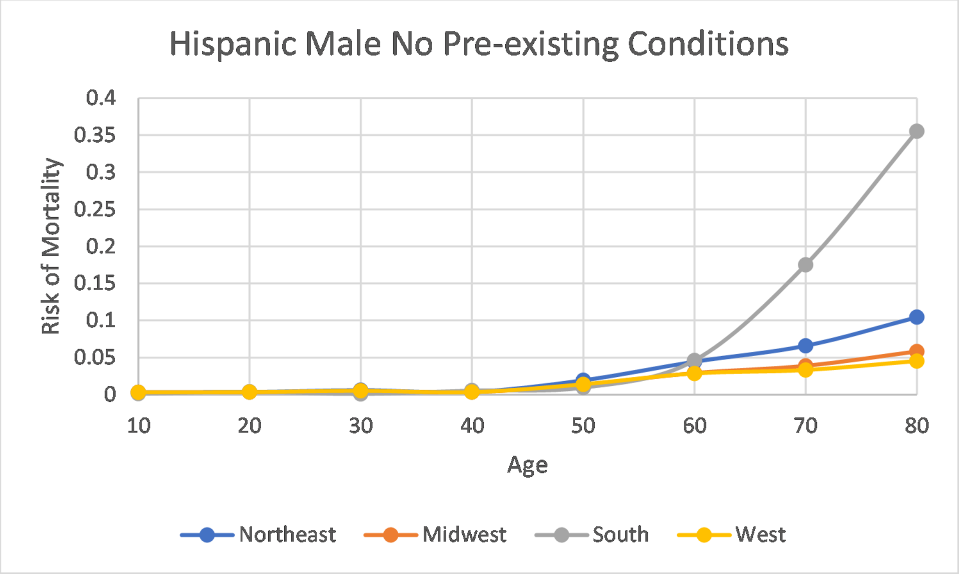

Figure 9. Probability of mortality for a Hispanic male with no pre-existing conditions for certain age groups. Given as a decimal. Same information as in Table 10.

Another obvious trend is the unrealistically high mortality risk for the south for many of the older age groups. The south was one of the lesser reported regions in our data, and it is quite possible that we have some reporting bias, with possibly only the most severe cases being reported and contributing to these extreme predicted mortality rates. It should be kept in mind that the focus for these predicted mortality rates is to look at the heterogeneity of risk factors, rather than getting the most accurate mortality prediction.

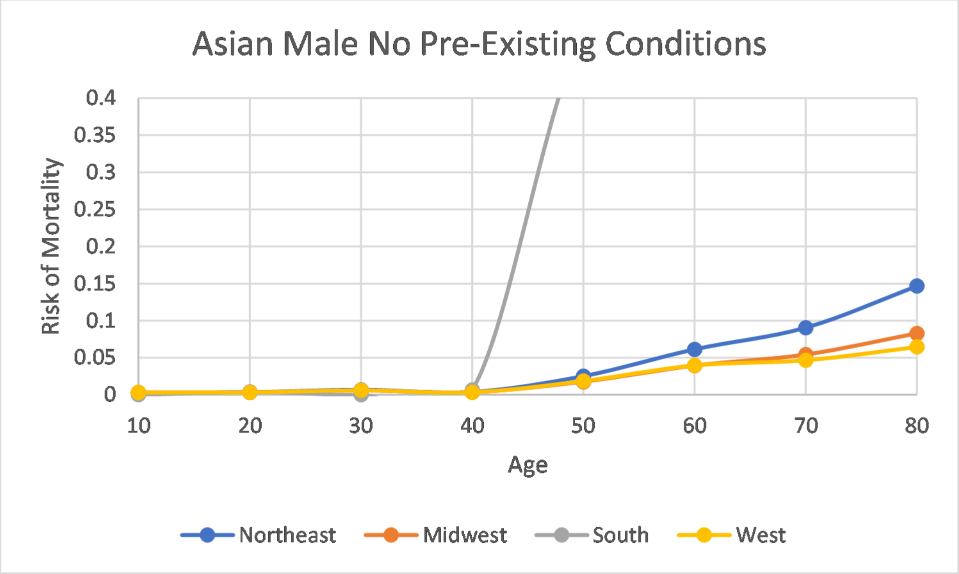

Figure 10. Probability of mortality for an Asian male with no pre-existing conditions for certain age groups. Given as a decimal. Same information as in Table 11.

4.4. K-Means Follow-Up

The general trend we see when plotting the data in a two-dimensional manner, as in Figure 1, is that the deaths are more heavily populated on the bottom left side of our plane. However, we do note that both deaths and survivors are found throughout the covered area of the plane with no real way to separate them into two neat clusters. Consequently, the features available to us are not discriminating factors in determining whether or not an individual dies from COVID-19. Despite this, clustering into four clusters, as seen in Figure 2, does get us somewhere. Cluster 4, seen toward the left of the plane in purple, is quite distinct from the other three clusters in terms of mortality. This cluster has a death rate of about 17%, while the other three clusters have death rates from around 2% to 4%. We can then look at how certain features are distributed amongst the four clusters to get a sense of what features determine membership in a cluster. We see these breakdowns in Figures 11–14. The general trend we see is unsurprising. We see younger individuals under-allocated and older individuals over-allocated to the high-risk cluster. Moreover, males, non-white individuals, and individuals with pre-existing conditions are over-allocated to the high-risk cluster. This makes sense, as our previous work has identified these as areas of risk for increased mortality.

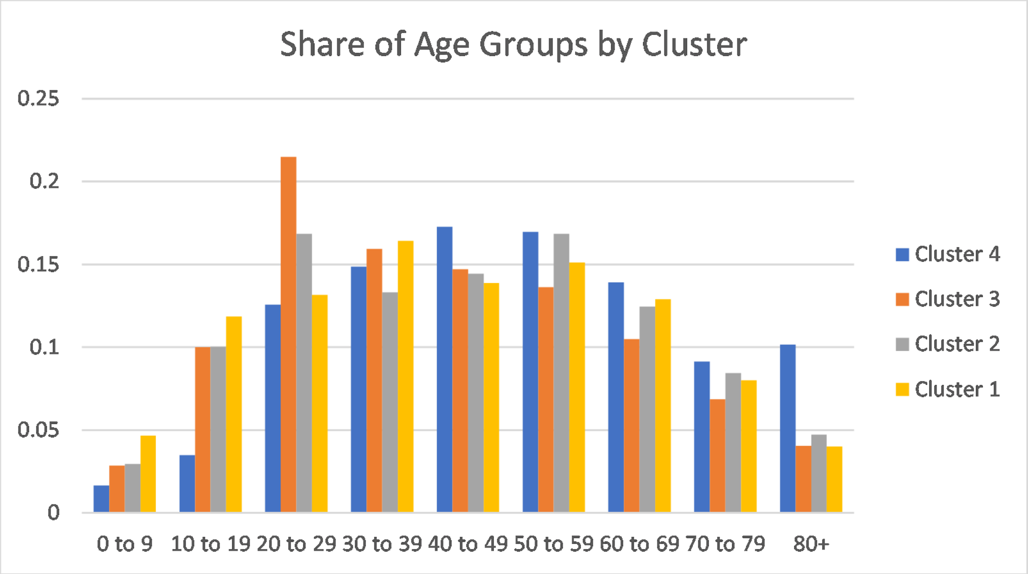

Figure 11. Proportion of individuals for each age group allocated to clusters. Given as a percentage. Cluster 4 was the high-mortality cluster.

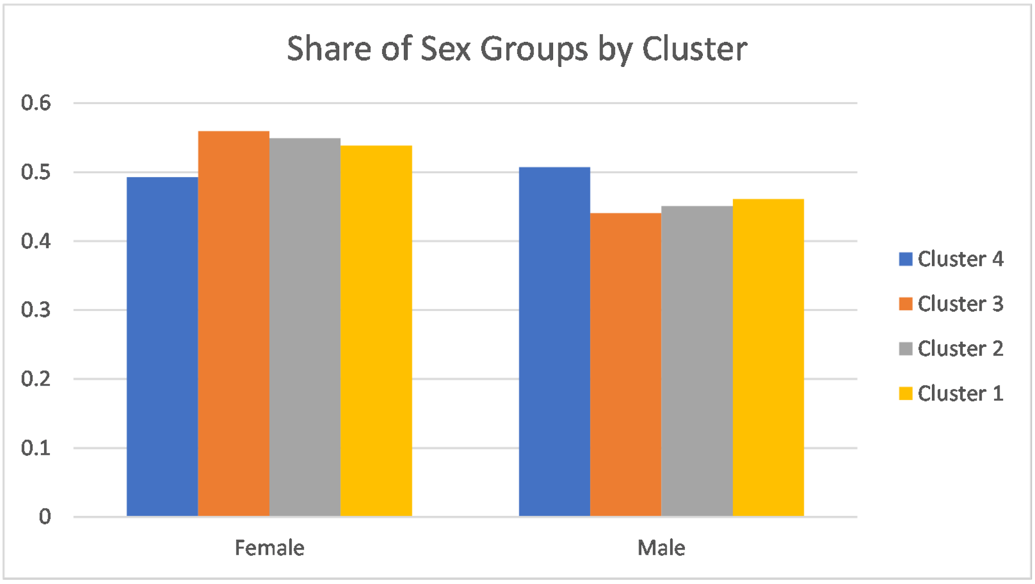

Figure 12. Proportion of individuals for each sex group allocated to clusters. Given as a percentage. Cluster 4 was the high-mortality cluster.

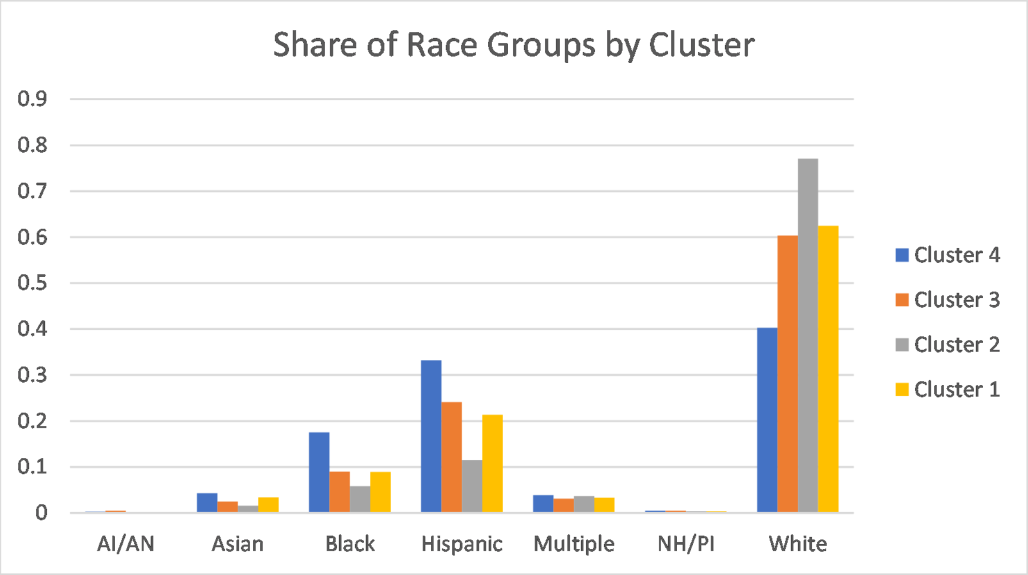

Figure 13. Proportion of individuals for each race group allocated to clusters. Given as a percentage. Cluster 4 was the high-mortality cluster.

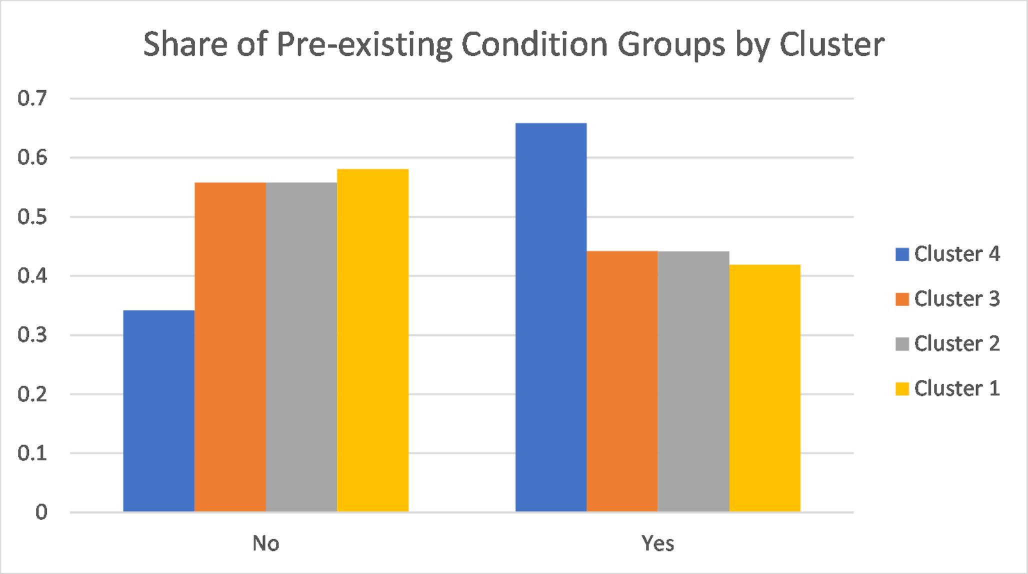

Figure 14. Proportion of individuals for each medical condition group allocated to clusters. Given as a percentage. Cluster 4 was the high-mortality cluster.

5. Conclusions

In this project, we set out to predict COVID-19 patient mortality using demographic data from patients in the United States. The Random Forest method was an attractive machine learning approach that had already been utilized for similar purposes. In fact, the AdaBoost boosted random forest had been established as the method that gave superior predictive performance when compared to several common techniques. While predictive performance is important, we were interested in the heterogeneity of COVID-19 risk factors. That is, we wanted to look at risk factors beyond the typically noted ones, such as having a pre-existing medical condition or being in an older age group.

Given the lack of structure that RF methods impose on the data, getting this information required the incorporation of additional tools. To this end, we incorporated a patient similarity measure that was derived from using a RF classifier in an unsupervised manner. This gave us a similarity measure between patients that allowed us to cluster our observations into four clusters. By building a RF classifier separately on each cluster, we hoped to make up the predictive performance gained by using the AdaBoost algorithm while simultaneously figuring out a way to get at how certain factors put some groups at higher risk of death from COVID-19.

The clustering approach did seem to pay off in prediction performance, as our approach yielded prediction performance metrics that were all within 0.02 of the Adaboost method. Additionally, the clusters did provide insight into factors affecting risk of mortality from COVID-19. We were able to look at which clusters were at higher or lower risk of death and note which demographic groups were over or under-represented in these clusters. The identified factors did in general mesh with conclusions put out by the CDC. The performance of our approach is hard to rank in the bigger picture of COVID-19 patient mortality prediction, given the different data that have been used for this purpose. As we use only the existence of a pre-existing condition as specific medical information, it should be understood that our approach will not give the same predictive performance as approaches that collect more extensive medical information. Rather, this approach should be utilized to identify at-risk demographics.

To better get at the intricacies of the added risk measure produced by the clustered RF approach, we built a neural network model to predict patient mortality. When this model was used to estimate risk for individuals from different racial, regional, and age groups, we did expose some hidden heterogeneity. The risk associated with a certain region, for example, did not affect all races or age groups equally. This neural network model certainly exposed the complexity in estimating COVID-19 mortality risk.

With the wide range of outcomes associated with COVID-19 infection, a way to assign some kind of order to the severity of outcome was needed. To this end, we utilized k-means clustering, along with a way to project the multi-dimensional observations onto the two-dimensional plane. This approach did shed some light on how discriminatory our features were in predicting COVID-19 mortality. While our density plots did show a general trend of death observations showing up more heavily in one area, our observations were far from being able to be neatly clustered into a survivor group and a death group. However, the subsequent clustering did produce a distinctly high-risk cluster whose membership seemed to be determined by the identified mortality risk factors.

In addition to providing some insight into the hidden heterogeneity of COVID-19 mortality risk, this project has seemingly succeeded in showing that RF methods when approached from a different angle can provide valuable insight into how certain features of data affect the prediction problem at hand. Specifically, our results could be used as a guide for which communities most need resources to combat COVID-19. We see clear trends of older, minority individuals in the northeast and south as being at an elevated risk of mortality from COVID-19, and this information should prompt a focus on these individuals.

While we were able to achieve good insights from our work, it should be noted that this approach is limited by the availability of uniformly good data. As the RF method cannot handle missing values, differing levels of reporting from different regions can weaken our results. This seemed to be evident in the unrealistic mortality rates predicted for the south region. Consequently, this method may be better served as part of a project that involves a gathering of data, rather than simply using what is available.

The ability to achieve good prediction performance and also get qualitative information about features of the data make a clustered RF approach worthy of exploration in similar problems. This approach could be used as a starting point to look at how certain features influence a dependent variable in cases where less is known at the outset. Further work in how the clustering is performed from the similarity measure developed from the RF method would be warranted. Only two methods to cluster the data were utilized in this project, and it is certainly possible that more creative ways to perform clustering might yield interesting results. Additionally, there is no reason that this approach could not be generalized to similar infectious diseases where frailty is a factor in patient survival.

Supplementary Materials: R code for this work may be accessed at https://github.com/edcorne/Cornelius-MS-Thesis-Summer-21-Illinois-State-University.git.

Author Contributions: O.A. and D.H. conceived the idea and formulated the direction of the project. E.C. had performed model development, modeling computations, analyzed results and wrote the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Access to the data used in this project may be requested from the CDC at https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Restricted-Access-Detai/mbd7-r32t (accessed on 30 March 2021).

Conflicts of Interest: The authors declare no conflict of interest.

Disclaimer: The CDC does not take responsibility for the scientific validity or accuracy of methodology, results, statistical analyses, or conclusions presented.

Abbreviations

The following abbreviations are used in this manuscript:

ML	Machine Learning
RF	Random Forest

References

1. Darcy, A.M.; Louie, A.K.; Roberts, L.W. Machine learning and the Profession of Medicine. *JAMA* **2016**, *315*, 551–552. [CrossRef]
2. Bates, D.W.; Saria, S.; Ohno-Machado, L.; Shah, A.; Escobar, G. Big Data In Health Care: Using Analytics To Identify And Manage High-Risk And High-Cost Patients. *Health Affairs* **2014**, *33*, 1123–1131. [CrossRef] [PubMed]
3. Jain, V.; Chatterjee, J.M. *Machine Learning with Health Care Perspective*; Springer: Berlin/Heidelberg, Germany, 2020
4. Chatterjee, J.M. Bioinformatics using Machine Learning. *Glob. J. Internet Interv. IT Fusion* **2018**, *1*, 28–35.
5. Khamparia, A.; Gupta, D.; de Albuquerque, V.H.C.; Sangaiah, A.K.; Jhaveri, R.H. Internet of health things-driven deep learning system for detection and classification of cervical cells using transfer learning. *J. Supercomput.* **2020**, *76*, 8590–8608. [CrossRef]
6. Waheed, A.; Goyal, M.; Gupta, D.; Khanna, A.; Al-Turjman, F.; Pinheiro, P.R. CovidGAN: Data Augmentation Using Auxiliary Classifier GAN for Improved Covid-19 Detection. *IEEE Access* **2020**, *8*, 91916–91923. [CrossRef] [PubMed]
7. Sakarkar, G.; Pillai, S.; Rao, C.V.; Peshkar, A.; Malewar, S. Comparative Study of Ambient Air Quality Prediction System Using Machine Learning to Predict Air Quality in Smart City. In Proceedings of the International Conference on IoT Inclusive Life (ICIIL 2019), NITTTR, Chandigarh, India, 19–20 December 2019; pp. 175–182.
8. CDC COVID Data Tracker. 2021. Available online: https://covid.cdc.gov/covid-data-tracker/#datatracker-home (accessed on 16 March 2021)
9. Aizenman, N. Protecting the Immuno-Compromised against COVID Could Be Key to Ending the Pandemic. 2021. Available online: https://www.npr.org/sections/goatsandsoda/2021/06/28/1011043650/the-key-to-ending-the-pandemic-may-be-protecting-immunocompromised-people#:~:text=All%20Things%20Considered-,Key%20To%20Ending%20Pandemic%20Could%20Be%20Protecting%20The%20Immuno%2DCompromised,slow%20the%20emergence%20of%20variants (accessed on 28 April 2021)
10. Wang, L.; Wong, A. COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-Ray Images. *arXiv* **2020**, arXiv:2003.09871.
11. Pal, R.; Sekh, A.; Kar, S.; Prasad, D. Neural Network Based Country Wise Risk Prediction of COVID-19. *Appl. Sci.* **2020**, *10*, 6448. [CrossRef]
12. Liu, D.; Clemente, L.; Poirier, C.; Ding, X.; Chinazzi, M.; Davis, J.T.; Vespignani, A.; Santillana, M. A machine learning methodology for real-time forecasting of the 2019–2020 COVID-19 outbreak using Internet searches, news alerts, and estimates from mechanistic models. *arXiv* **2020**, arXiv:2004.04019.
13. Beck, B.R.; Shin, B.; Choi, Y.; Park, S.; Kang, K. Predicting commercially available antiviral drugs that may act on the novel coronavirus (SARS-CoV-2) through a drug-target interaction deep learning model. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 784–790. [CrossRef] [PubMed]
14. Khalifa, N.E.M.; Taha, M.H.N.; Hassanien, A.E.; Elghamrawy, S. Detection of Coronavirus (COVID-19) Associated Pneumonia based on Generative Adversarial Networks and a Fine-Tuned Deep Transfer Learning Model using Chest X-ray Dataset. *arXiv* **2020**, arXiv:2004.01184.
15. Sujath, R.; Chatterjee, J.M.; Hassanien, A.E. A machine learning forecasting model for COVID-19 pandemic in India. *Stoch. Environ. Res. Risk Assess.* **2020**, *34*, 959–972. [CrossRef] [PubMed]
16. Pourhomayoun, M.; Shakibi, M. Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making. *Smart Health* **2021**, *20*, 100178 [CrossRef]
17. Karthikeyan, A.; Garg, A.; Vinod, P.; Priyakumar, U.D. Machine Learning Based Clinical Decision Support System for Early COVID-19 Mortality Prediction. *Front. Public Health* **2021**, *9*, 475. [CrossRef]
18. Kar, S.; Chawla, R.; Haranath, S.P.; Ramasubban, S.; Ramakrishnan, N.; Vaishya, R.; Sibal, A.; Reddy, S. Multivariable mortality risk prediction using machine learning for COVID-19 patients at admission (AICOVID). *Sci. Rep.* **2021**, *11*, 12801. [CrossRef]
19. Tang, Z.; Zhao, W.; Xie, X.; Zhong, Z.; Shi, F.; Liu, J.; Shen, D. Severity Assessment of Coronavirus Disease 2019 (COVID-19) Using Quantitative Features from Chest CT Images. *arXiv* **2020**, arXiv:2003.11988.
20. de Freitas Barbosa, V.A.; Gomes, J.C.; de Santana, M.A.; de Lima, C.L.; Calado, R.B.; Bertoldo, C.R., Jr.; de Almeida Albuquerque, J.E.; de Souza, R.G.; de Araujo, R.J.E.; de Souza, R.E.; et al. Covid-19 rapid test by combining a random forest based web system and blood tests. *medRxiv* **2020**. [CrossRef]
21. Gupta, V.K.; Kumar, D.; Sardana, A. Prediction of COVID-19 confirmed, death, and cured cases in India using random forest model. *Big Data Min. Anal.* **2021**, *4*, 116–123. [CrossRef]
22. Yesilkanat, C.M. Spatio-temporal estimation of the daily cases of COVID-19 in worldwide using random forest machine learning algorithm. *Chaos Solitons Fractals* **2020**, *140*, 110210. [CrossRef]
23. An, C.; Lim, H.; Kim, D.W.; Chang, J.H.; Choi, Y.J.; Kim, S.W. Machine learning prediction for mortality of patients diagnosed with COVID-19: A nationwide Korean cohort study. *Sci. Rep.* **2020**, *10*, 18716. [CrossRef]

24. Wang, J.; Yu, H.; Hua, Q.; Jing, S.; Liu, Z.; Peng, X.; Cao, C.; Luo, Y. A descriptive study of random forest algorithm for predicting COVID-19 patients outcome. *PeerJ* **2020**, *8*, e9945. [CrossRef]
25. Majhi, R.; Thangeda, R.; Sugasi, R.P.; Kumar, N. Analysis and prediction of COVID-19 trajectory: A machine learning approach. *J. Public Aff.* **2020**, e2537 [CrossRef] [PubMed]
26. Iwendi, C.; Bashir, A.K.; Peshkar, A.; Sujatha, R.; Chatterjee, J.M.; Pasupuleti, S.; Mishra, R.; Pillai, S.; Jo, O. COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm. *Front. Public Health* **2020**, *8*, 357. [CrossRef]
27. Risk for COVID-19 Infection, Hospitalization, and Death By Race/Ethnicity. 2021. Available online: https://www.cdc.gov/coronavirus/2019-ncov/covid-data/investigations-discovery/hospitalization-death-by-race-ethnicity.html (accessed on 9 April 2021)
28. Andreu-Perez, J.; Poon, C.; Merrifield, R.; Wong, S.; Yang, G. Big data for health. *IEEE J. Biomed. Health Inform.* **2015**, *19*, 1208. [CrossRef] [PubMed]
29. Joyner, M.J.; Paneth, N. Seven Questions for Personalized Medicine. *JAMA* **2015**, *314*, 999–1000. [CrossRef]
30. Celi, L.A.; Marshall, J.D.; Lai, Y.; Stone, D.J. Disrupting Electronic Health Records Systems: The Next Generation. *JMIR Med. Inform.* **2015**, *3*, e34. [CrossRef] [PubMed]
31. Xu, R.; Nettleton, D.; Nordman, D.J. Case-Specific Random Forests. *J. Comput. Graph. Stat.* **2016**, *25*, 49–65. [CrossRef]
32. Park, Y.J.; Kim, B.C.; Chun, S.H. New knowledge extraction technique using probability for case-based reasoning: Application to medical diagnosis. *Expert Syst.* **2006**, *23*, 2–20. [CrossRef]
33. Panahiazar, M.; Taslimitehrani, V.; Pereira, N.L.; Pathak, J. Using EHRs for heart failure therapy recommendation using multidimensional patient similarity analytics. *Stud. Health Technol. Inform.* **2015**, *210*, 369–373.
34. Brookhart, M.A.; Schneeweiss, S.; Rothman, K.J.; Glynn, R.J.; Avorn, J.; Sturmer, T. Variable selection for propensity score models. *Am. J. Epidemiol.* **2006**, *163*, 1149–1156. [CrossRef] [PubMed]
35. Lee, J. Patient-Specific Predictive Modeling Using Random Forests: An Observational Study for the Critically Ill. *JMIR Med. Inform.* **2017**, *5*, e3. [CrossRef]
36. Liaw, A.; Wiener, M. Classification and Regression by randomForest. *R News* **2002**, *2*, 18–22.
37. Alfaro, E.; Gámez, M.; García, N. adabag: An R Package for Classification with Boosting and Bagging. *J. Stat. Softw.* **2013**, *54*, 1–35. [CrossRef]
38. Maechler, M.; Rousseeuw, P.; Struyf, A.; Hubert, M.; Hornik, K. *Cluster: Cluster Analysis Basics and Extensions*; R package version 2.1.1—For new features, see the 'Changelog' file (in the package source); R Core Team: Vienna, Austria, 2021.
39. Fritsch, S.; Guenther, F.; Wright, M.N. *Neuralnet: Training of Neural Networks*; R package version 1.44.2; R Core Team: Vienna, Austria, 2019
40. Centers for Disease Control and Prevention, COVID-19 Response. *COVID-19 Case Surveillance Data Access, Summary, and Limitations (30 March 2021 Version)*; Centers for Disease Control and Prevention: Atlanta, GA, USA, 2021.
41. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
42. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 389–415.
43. R Core Team *R: A Language and Environment for Statistical Computing*; R package version 1.44.2; R Core Team: Vienna, Austria, 2019.
44. CDC People at Increased Risk. 2021. Available online: https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/index.html (accessed on 16 April 2021)