Medical & Biological Engineering & Computing (2022) 60:1881–1896
https://doi.org/10.1007/s11517-022-02570-8

ORIGINAL ARTICLE

Efficient analysis of COVID-19 clinical data using machine learning models

Sarwan Ali[1] · Yijing Zhou[1] · Murray Patterson[1]

Received: 17 October 2021 / Accepted: 27 March 2022 / Published online: 4 May 2022

Abstract

Because of the rapid spread of COVID-19 to almost every part of the globe, huge volumes of data and case studies have been made available, providing researchers with a unique opportunity to find trends and make discoveries like never before by leveraging such big data. This data is of many different varieties and can be of different levels of veracity, e.g., precise, imprecise, uncertain, and missing, making it challenging to extract meaningful information from such data. Yet, efficient analyses of this continuously growing and evolving COVID-19 data is crucial to inform — often in real-time — the relevant measures needed for controlling, mitigating, and ultimately avoiding viral spread. Applying machine learning-based algorithms to this big data is a natural approach to take to this aim since they can quickly scale to such data and extract the relevant information in the presence of variety and different levels of veracity. This is important for COVID-19 and potential future pandemics in general. In this paper, we design a straightforward encoding of clinical data (on categorical attributes) into a fixed-length feature vector representation and then propose a model that first performs efficient feature selection from such representation. We apply this approach to two clinical datasets of the COVID-19 patients and then apply different machine learning algorithms downstream for classification purposes. We show that with the efficient feature selection algorithm, we can achieve a prediction accuracy of more than 90% in most cases. We also computed the importance of different attributes in the dataset using information gain. This can help the policymakers focus on only certain attributes to study this disease rather than focusing on multiple random factors that may not be very informative to patient outcomes.

Keywords COVID-19 · Coronavirus · Clinical data · Classification · Feature selection

1 Introduction

Because of the rapid global spread of COVID-19 and the cooperation of medical institutions worldwide, a tremendous amount of public data — more data than ever before for a single virus — has been made available for researchers [1–3]. This "big data" opens up new opportunities to analyze the behavior of this virus [4, 5]. Despite these opportunities, the huge size of the data poses a challenge for its processing on smaller systems [1]. On the one hand, this creates scalability issues, and on the other hand, it creates the problem of high dimensionality (the curse of dimensionality) [6, 7]. Since such data was not previously available to the research community at this magnitude and ease of access, new and more sophisticated methods are required to extract useful information from this big data.

At the same time, the shortage of medical resources may occur when such a severe pandemic happens, and the surging number of patients exceeds the capacity of the clinical system. This situation happens in many countries and regions during continuous outbreaks of the COVID-19 pandemic and clinicians have to make the tough decision of which individual patients have a higher possibility to recover and should receive a limited amount of medical care. What is more difficult is the decision of which patients have little to no chance of survival, regardless of treatment level, and should hence be abandoned for the sake of optimizing the use of limited resources for others who still have a chance. In addition to this, patients with different levels of severity

✉ Sarwan Ali
sali85@student.gsu.edu

Yijing Zhou
yzhou43@student.gsu.edu

Murray Patterson
mpatterson30@gsu.edu

[1] Georgia State University, Atlanta, GA, USA

and urgency of symptoms require the medical system to create a complete plan to provide various treatments in proper order [8].

The clinical decision support system is of utmost importance to optimize the use of the limited medical resources and thus save more lives overall [8]. In order to develop such a clinical decision support system with high quality, it is necessary to build a model that can predict the possible complications of patients, assessing the likelihood that they will survive under certain levels of care. Machine learning (ML)-based algorithms are proven to perform well in terms of classification and clustering. Therefore, we work on building machine learning (ML) models that can scale to larger datasets and reduce the run time by selecting the proper attributes. Classification of clinical data as early as possible is an important goal as it could help relevant authorities (e.g., doctors) to make appropriate decisions on time. The earliest decision in many application domains could be more rewarding and support efficient decision-making. However, humans can take more time to process the information and come up with a conclusion. Since time is an essential factor in dealing with people's lives, a slight delay in decision-making could be very costly. Using ML models, we can speed up the information analysis part and make it more efficient in predictive performance than human-based analysis. In this paper, we first use ML models to study how we can improve the classifiers' predictive performance and improve the runtime so that it could help doctors make efficient decisions on time. Manual analysis of accuracy vs. time trade-off is not easy for humans, and hence that problem could be solved using ML. Secondly, we use the interpretability model to explain the reason behind the specific behaviors of the classifiers. We use a popular explainability model (SHAPE [9–11]) to understand the clinical data and impact of different features of the coronavirus patients. In this way, once doctors have a predictive decision (in less time) from an ML model and the reasons behind those decisions (computed using SHAPE), doctors can take decision early, which could save lives of people by focusing on high risk patients and also by focusing on only those clinical attributes of the patients that are highly correlated to their disease. Since ML models take a feature vector representation as input [12, 13], designing such vectors while preserving a maximum amount of information is a challenging task [14]. Moreover, when the size of the data becomes large, even scalability of these models becomes an issue.

In this paper, we propose a pipeline to efficiently predict with high accuracy (and low runtime) patient mortality and likelihood of testing positive/negative for COVID-19 as a function of many different factors. Our pipeline involves data cleaning, data preprocessing, feature selection, classification, and various statistical analyses on the results. Using the clinical findings, our model can help doctors to prescribe medications and design strategies in advance that can help to save the highest number of lives. We use two different datasets in this paper, which involve clinical findings from the Centers for Disease Control and Prevention (CDC), USA[1], on factors such as age group, sex, ethnicity, and residence, and another from the Israelita Albert Einstein Hospital in Sao Paulo, Brazil [15], on many factors which can be obtained from a blood test, such as leukocytes, platelets, and red blood cells counts. Our contributions can be summarized as follows:

1. We propose a pipeline to efficiently predict patient mortality as a function of a few dozen factors. We show that with basic information about a patient (gender, race, exposure, etc.), we can predict in advance the likelihood of mortality in the future. We also predict if a patient is COVID-19 positive or negative using attributes like red blood cells and hemoglobin.
2. We show that our model is scalable on larger datasets (achieves accuracies >90%).
3. From our results, it is evident that the proposed model (using efficient feature selection) outperforms the baselines (without using any feature selection) in terms of prediction accuracy and runtime.
4. We show the importance of each attribute by measuring the information gain of the attributes with the class labels. This study can help doctors and other relevant authorities to focus more on specific attributes rather than dealing with all information at once, which can be difficult for humans to fathom.
5. We also use other statistical analyses such as Pearson correlation and Spearman correlation to understand the behavior of data and find the correlations between different attributes and the class labels (patient's mortality and likelihood of being COVID-19 positive/negative).
6. We use a popular method, called SHAP analysis, to measure the impact of variables against other features. This study helps us understand which class of the label is highly impacted by a specific feature from the dataset.

The rest of the paper is organized as follows: Section 2 contains literature review for the problem. Our proposed model is given in Section 3. Dataset statistics and experimental details are given in Section 4. We show results and their comparisons in Section 5. We provide an array of statistical analysis, such as importance of attributes, correlation, and SHAP analysis in Section 6. Finally, in Section 7, we conclude our paper.

[1] https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w4/data

2 Related work

Machine learning-based models that take fixed-length feature vectors as input has been successfully applied (for data analytics tasks) in many domains such as graphs analytics [12, 16], smart grid [6, 7], electromyography (EMG) [17], and text classification [18–20]. It is important to perform an objective evaluation of the underlying model rather than just doing subjective evaluation [21]. Many data science methodologies have been applied to objectively analyze the data of COVID-19 and provide support to the medical system. The synergy between data scientists and the biomedical communities is helpful to improve the detection of diseases and illnesses, as well as the prediction of possible complications. Authors in [22] developed a framework to predict cancer trends accurately. This type of framework could be used for the analysis of other clinical data. Authors in [23] used spike sequences to classify the variants of the COVID-19 infected humans. An effective approach to cluster the spike sequences based on the virus's variants is conducted in [24].

Several types of methods have been used to study and understand the behavior of the COVID-19 pandemic. One way is to use the genomic data to classify/cluster the coronavirus sequences that could be present in humans [1, 23, 25] or different hosts [26]. Some effort is made to understand the locality of the virus by evaluating the spike sequences of the coronavirus [2]. Another type of study involves using the computed tomography (CT) scan images of the human chest to identify the coronavirus [27, 28]. Authors in [27] used deep learning techniques to differentiate the CT scan images of COVID-19 and non-COVID 19 patients. A Convolutional Neural Network-based method to classify these CT scan images of COVID-19 patients is presented in [29]. A fast COVID-19 cases detection method using X-ray and CT scan images of the chest is proposed in [30], which uses a deep transfer learning algorithm to detect the COVID-19 positive cases in ≤ 2 seconds.

Several studies discussed different data mining techniques to study the behavior of the COVID-19 [5, 31, 32]. Authors in [33] used neural networks, which take advantage of few-shot learning and autoencoder to perform predictive analysis on COVID-19 data. Some studies also focus on finding the conditional dependencies between features, which can be used to analyze the behavior of different features towards the prediction of label [34]. A study for predicting the clinical outcomes of patients and indicating whether patients are more likely to recover from coronavirus or in danger of death is performed in [4]. They presented a tool called online analytical processing (OLAP), which can help the researchers learn more about the confirmed cases and mortality of COVID-19 by conducting machine learning methods on the big dataset of COVID-19.

3 Proposed approach

Most of the machine learning (ML) models take fixed-length feature vectors as an input to perform different tasks such as classification and clustering. We design a fixed-length feature vector representation, which includes the values of different attributes of the clinical data. One important point to mention here is that not all the features in the vectors are important in terms of predicting the class labels. Therefore, it is required to apply feature selection to not only improve the predictive performance of the underlying classifiers (by removing unnecessary features), but also improve the training runtime. The feature selection methods that we used in this paper are discussed below.

3.1 Feature selection methods

We use different supervised and unsupervised feature selection methods to improve the underlying classifiers' runtime and improve the predictive performance. For supervised models, we use Boruta (shadow features) [35], and Ridge Regression (RR) [36]. For unsupervised methods, we use the approximate kernel approach called Random Fourier Features (RFF) [37].

3.1.1 Boruta (shadow features)

The main idea of Boruta is that features do not compete among themselves, but rather they compete with a randomized version of them. Boruta captures the non-linear relationships and interactions using the random forest algorithm. It then extracts the importance of each feature (corresponding to the class label) and only keeps the features that are above a specific threshold of importance. To compute the importance of the features, it performs the following task: From the original features set in the data, it creates dummy features (shadow features) by randomly shuffling each feature. Now the shadow features are combined with the original features set to obtain a new dataset, which has twice the number of features of the original data. Using random forest, it computes the importance of the original and shadow features separately. Now the importance of the original features is compared with the threshold. The threshold is defined as the highest feature importance recorded among the shadow features. The feature from the original feature set is selected if its importance (computed using random forest) is greater than the threshold (highest importance value among shadow features). In Boruta, a feature is useful only if it is capable of doing better than the best randomized feature. Note that we are using two datasets in this paper, namely Clinical Data1, and Clinical Data2 (see Section 4.1 for detail regarding datasets). For Clinical Data1, Boruta selected 11 features out of 19 and removed Year, Gender, Race, Case Positive Specimen Interval, Case Onset Interval, Exposure, Current Status, and Symptom Status. For the Clinical Data2, Boruta selected 7 features from 18 features in total. The selected features are Red Blood Cells, Platelets, Hematocrit, Monocytes, Leukocytes, Eosinophils, and Proteina C reativa mg/dL.

3.1.2 Ridge regression

Ridge Regression (RR) is a supervised algorithm for parameter estimation that is used to address the collinearity problem that arises in multiple linear regression frequently [38, 39]. Its main idea is to increase the bias (it first introduces a Bias term for the data) to improve the variance, which shows the generalization capability of RR as compared to simple linear regression. RR ignores the data points that are far away from others, and it tries to make the regression line more horizontal. RR is useful for Feature selection because it gives insights on which independent variables are not very important (can reduce the slope close to zero). The un-important independent variables are then removed to reduce the dimensions of the overall dataset. The objective function of ridge regression is the following

$$min(\text{Sum of square residuals} + \alpha \times \text{slope}^2) \tag{1}$$

where $\alpha \times slope^2$ is called penalty terms.

3.1.3 Random Fourier features (RFF)

A popular approach for classification is using kernel-based algorithms, which computes a similarity matrix that can be used as input for traditional classification algorithms such as support vector machines. However, pair-wise computation for the kernel matrix is an expensive task. To make this task efficient, a method called the kernel trick is used.

Definition 1 It works by taking the dot product between the pairs of input points. Kernel trick avoids the need to map the input data (explicitly) to a high-dimensional feature space.

The main idea of the Kernel Trick is the following: *Any positive definite function f(x,y), where $x, y \in \mathcal{R}^d$, defines an inner product and a lifting ϕ for the purpose of computing the inner product quickly between the lifted data points* [37]. More formally:

$$\langle \phi(x), \phi(y) \rangle = f(x, y) \tag{2}$$

The main problem of the kernel method is that when we have large-sized data, they suffer from high initial computational and storage costs. To solve these problems, we use an approximate kernel method called Random Fourier Features (RFF) [37]. The RFF maps the given data to a low dimensional randomized feature space (euclidean inner product space). More formally:

$$z : \mathcal{R}^d \rightarrow \mathcal{R}^D \tag{3}$$

RFF basically approximate the inner product between a pair of transformed points. More formally:

$$f(x, y) = \langle \phi(x), \phi(y) \rangle \approx z(x)'z(y) \tag{4}$$

In Eq. (4), z is low dimensional (unlike the lifting ϕ). In this way, we can transform the original input data with z. Now, z is the approximate low dimensional embedding for the original data. We can then use z as the input for different classification algorithms.

3.2 Classification algorithms

For classification, we use Support Vector Machine (SVM), Naive Bayes (NB), Multiple Linear Regression (MLP), K-Nearest Neighbors (KNN), Random Forest (RF), Logistic Regression (LR), and Decision Tree (DT). All algorithms are used with default parameters. The value for K in the case of KNN is taken as 5 (using a standard validation set approach [40]).

We are also using a model called Keras Classifier for classification purposes. For this model, we use a sequential constructor. We create a fully connected network with one hidden layer that contains p neurons, where p is equal to the length of the feature vector. We use "rectifier" as an activation function and "softmax" activation function in the output layer. We also use an efficient Adam gradient descent optimization algorithm with "sparse categorical crossentropy" loss function (used for multi-class classification problem), which computes the crossentropy loss between the labels and predictions. For training the model, the batch size and number of epochs are taken as 100 and 10, respectively. Since the keras classification model does not require feature selection, we use the original data without using any feature selection method to input Keras classifiers.

Remark 1 We use "sparse categorical crossentropy" instead of simple "categorical crossentropy" because we are using integer labels rather than the one-hot representation of labels.

4 Experimental setup

In this section, we describe our dataset in detail. All experiments are performed on a Core i5 system running the Windows 10 OS, 32GB memory, and a 2.4 GHz processor. Implementation of the algorithms is done in Python. Our code and the prepossessed dataset are available online[2].

4.1 Dataset statistics

In this paper, we are using clinical data from two different sources. The description of both datasets is given below.

4.2 Clinical Data1

We use COVID-19 Case Surveillance dataset (we call it Clinical Data1 for reference), which is publicly available on the Centers for Disease Control and Prevention CDC, USA's website[3]. After preprocessing (removing missing values), we got 95984 patients data record. The attributes in the dataset are following:

1. **Year:** The earlier of year the Clinical Date. date related to the illness or specimen collection or the Date Received by CDC.
2. **Month:** The earlier of month the Clinical Date. date related to the illness or specimen collection or the Date Received by CDC (see Fig. 1 for month and year distribution).
3. **State of residence:** This attribute shows the name of the state (of the USA) in which the patient is living (see Fig. 2 for states distribution).
4. **State FIPS code:** Federal Information Processing Standards (FIPS) code for different states.
5. **County of residence:** Name of the County.
6. **County fips code:** Federal Information Processing Standards (FIPS) code for different Counties.
7. **Age group:** Age groups of patients that include 0–17 years, 18–49 years, 50–64 years, and 65 + years.
8. **Gender:** Female, Male, Other, Unknown.
9. **Race:** American Indian/Alaska Native, Asian, Black, Multiple/Other, Native Hawaiian/Other Pacific Islander, White, Unknown (see Table 1 for the distribution of values for race attribute).
10. **Ethnicity:** Hispanic, Non-Hispanic, Unknown.
11. **Case positive specimen interval:** Weeks between earliest date and date of first positive specimen collection.
12. **Case onset interval:** Weeks between earliest date and date of symptom onset.
13. **Process:** Under what process was the case first identified, e.g., Clinical evaluation, Routine surveillance, Contact tracing of case patient, Multiple, Other, Unknown. (see Table 2).
14. **Exposure:** In the 14 days prior to illness onset, did the patient have any of the following known exposures, e.g., domestic travel, international travel, cruise ship or vessel travel as a passenger or crew member, workplace, airport/airplane, adult congregate living facility (nursing, assisted living, or long-term care facility), school/university/childcare center, correctional facility, community event/mass gathering, animal with confirmed or suspected COVID-19, other exposure, contact with a known COVID-19 case? Possible values for this attribute are Yes and Unknown.
15. **Current status:** What is the current status of this person? Possible values are Laboratory-confirmed case, Probable case.
16. **Symptom status:** What is the symptom status of this person? Possible values are Asymptomatic, Symptomatic, Unknown, Missing.
17. **Hospital:** Was the patient hospitalized? Possible values are "Yes", "No", and "Unknown".
18. **ICU:** Was the patient admitted to an intensive care unit (ICU)? Possible values are "Yes", "No", and "Unknown".
19. **Death/Deceased:** This attribute highlights whether the patient died as a result of this illness. The possible values are "Yes", "No", and "Unknown".
20. **Underlying Conditions:** This attribute highlights if the patient has single or multiple medical conditions and risk behaviors. These conditions include diabetes mellitus, hypertension, severe obesity (occurs when BMI is greater than 40), cardiovascular disease, chronic renal disease, chronic liver disease, chronic lung disease, other chronic diseases, immunosuppressive condition, autoimmune condition, current smoker, former smoker, substance abuse or misuse, disability, psychological/psychiatric, pregnancy, other. The possible values for this attribute are "Yes" and "No".

The Distribution of values for different attributes is shown in Fig. 3. To check if there is any natural clustering in Clinical Data1, we use the t-distributed stochastic neighbor embedding (t-SNE) approach [41]. We map input data to 2d real vectors representation using t-SNE and Deceased attribute (for Clinical Data1) as a class label (see Fig. 4). We can observe in the figure that there is no visible clustering corresponding to different values of the deceased attribute. All values (No, Yes, and Unknown) are scattered around in

[2] https://github.com/sarwanpasha/COVID_19_Clinical_Data_Analytics

[3] https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w4/data

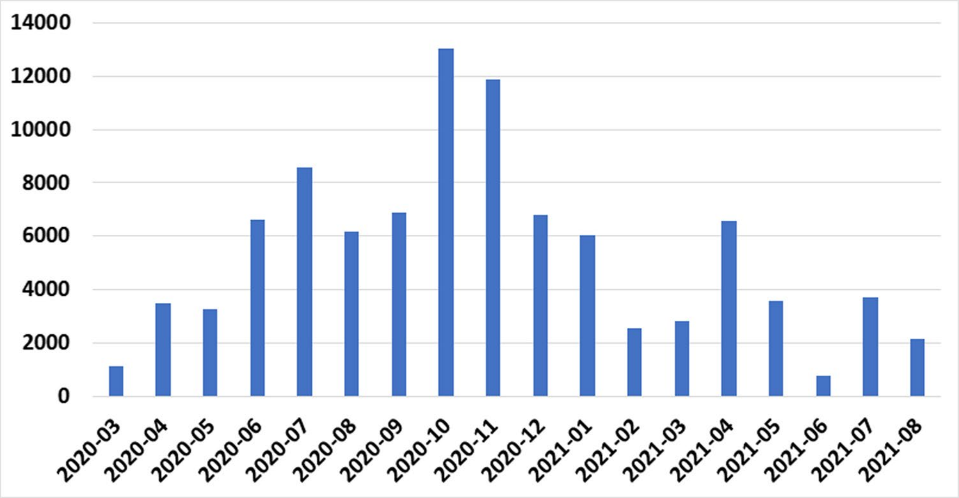


Fig. 1 Month and Year attribute distribution

the plot. This behavior shows that performing any ML task on such data will not directly give us efficient results (since the data is not properly grouped together).

4.3 Clinical Data2

We obtained the Clinical Data2 data from [15]. This study used a laboratory dataset of patients with COVID-19 in the Israelita Albert Einstein Hospital in Sao Paulo, Brazil. The patient samples were collected to identify who were infected by COVID-19 at the beginning of the year 2020. The laboratory dataset contains information on 608 patients with 18 laboratory findings. In this dataset, 524 had no findings, and 84 were patients with COVID-19. The attributes are Red blood Cells, Hemoglobin, Platelets, Hematocrit, Aspartate transaminase, Lymphocytes, Monocytes, Sodium, Urea, Basophils, Creatinine, Serum Glucose, Alanine transaminase, Leukocytes, Potassium, Eosinophils, Proteina C reativa mg/dL, Neutrophils, SARS-Cov-2 exam result (positive or negative). All the attributes (other than "SARS-Cov-2 exam result") contain integer values.

4.4 Evaluation metrics

To measure the performance of underlying machine learning classifiers, we use different evaluation metrics such as Average Accuracy, Precision, Recall, weighted and Macro F1, and Receiver Operator Curve (ROC) Area Under the Curve (AUC). We also computed the training runtime of all ML models to see which model is the best in terms of runtime. We use 5-fold cross validation to test the performance of classifiers and compare the results of different models.

5 Results and discussion

The average and standard deviation results for Clinical Data1 are given in Table 3 and Table 4, respectively. For classifying the Deceased attribute, we can see that all methods are

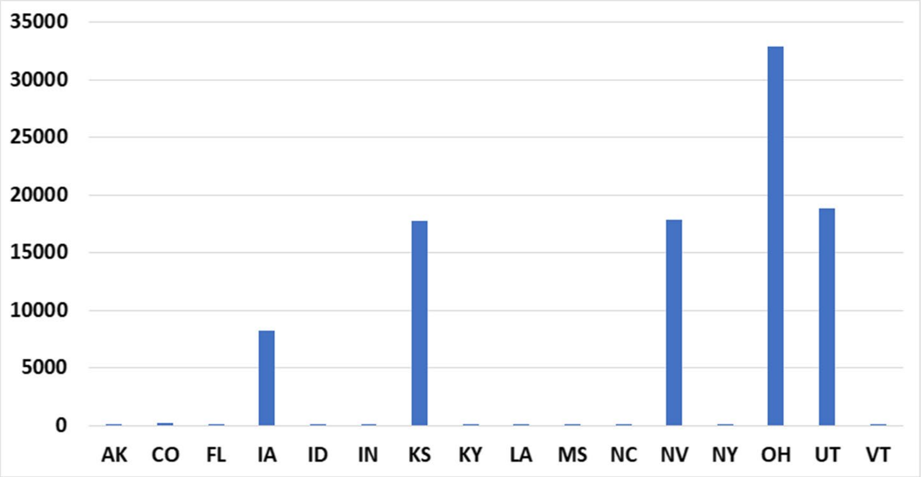


Fig. 2 State of residence distribution

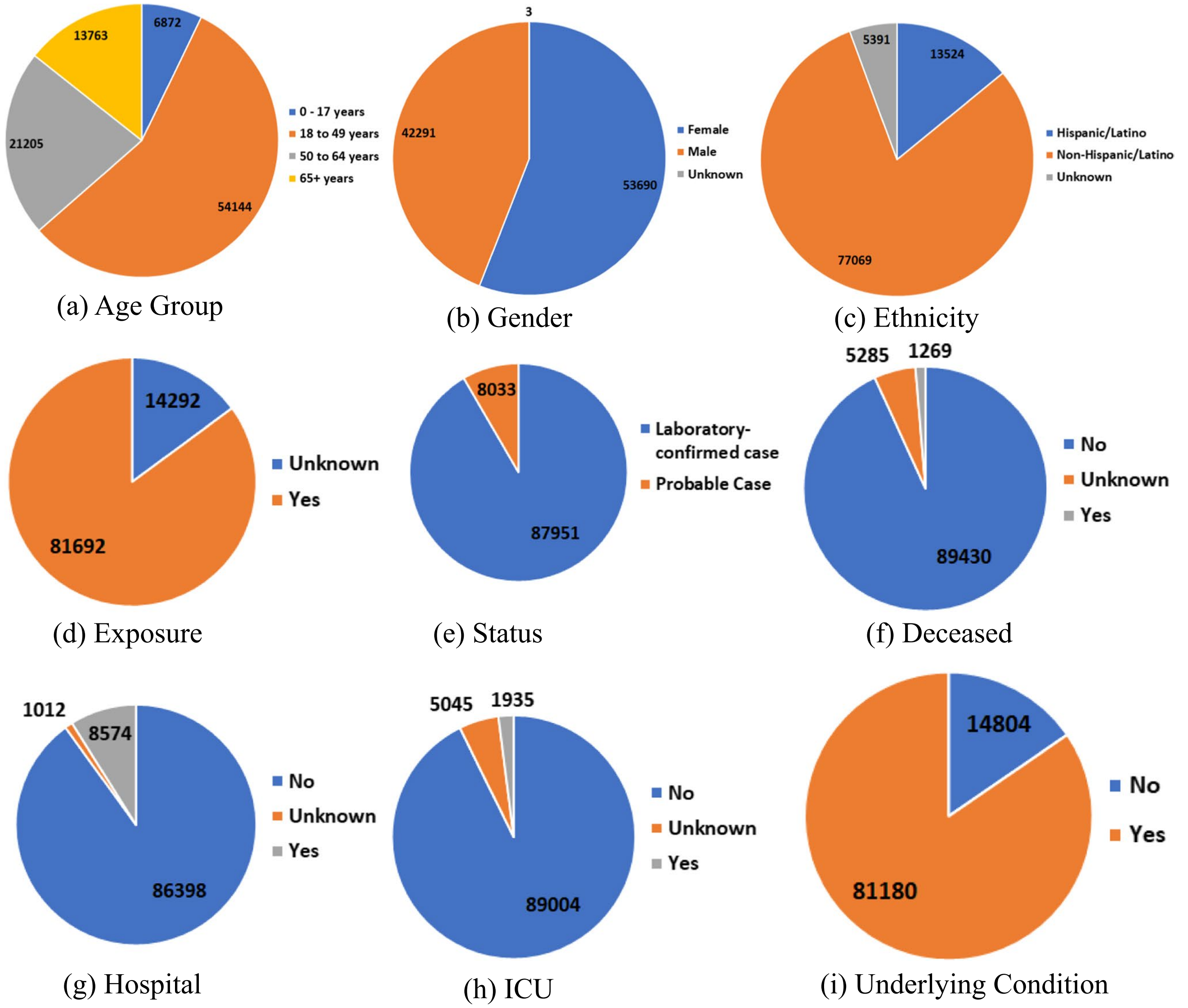
Table 1 Race attribute distribution

Race	Count
American Indian/ Alaska Native	94
Asian	3067
Black	8806
Multiple/Other	1833
Native Hawaiian/Other Pacific Islander	859
Unknown	3081
White	78244

Table 2 Process attribute distribution

Process	Count
Clinical evaluation	43768
Contact tracing of case patient	6813
Laboratory reported	11848
Multiple	22595
Other	556
Other detection method (specify)	164
Provider reported	212
Routine physical examination	22
Routine surveillance	8641
Unknown	1365

able to classify the label (Deceased attribute) with very high accuracy (accuracy > 90 in most of the cases). Note that feature selection-based models are better in terms of prediction accuracy and outperform the setting in which we are not using any feature selection approach (No Feat. Selec.). Also, the Boruta feature selection model outperforms all



(a) Age Group

(b) Gender

(c) Ethnicity

(d) Exposure

(e) Status

(f) Deceased

(g) Hospital

(h) ICU

(i) Underlying Condition

Fig. 3 Pie charts for the distribution of different attributes values

Fig. 4 t-SNE plot for deceased attribute of Clinical Data1

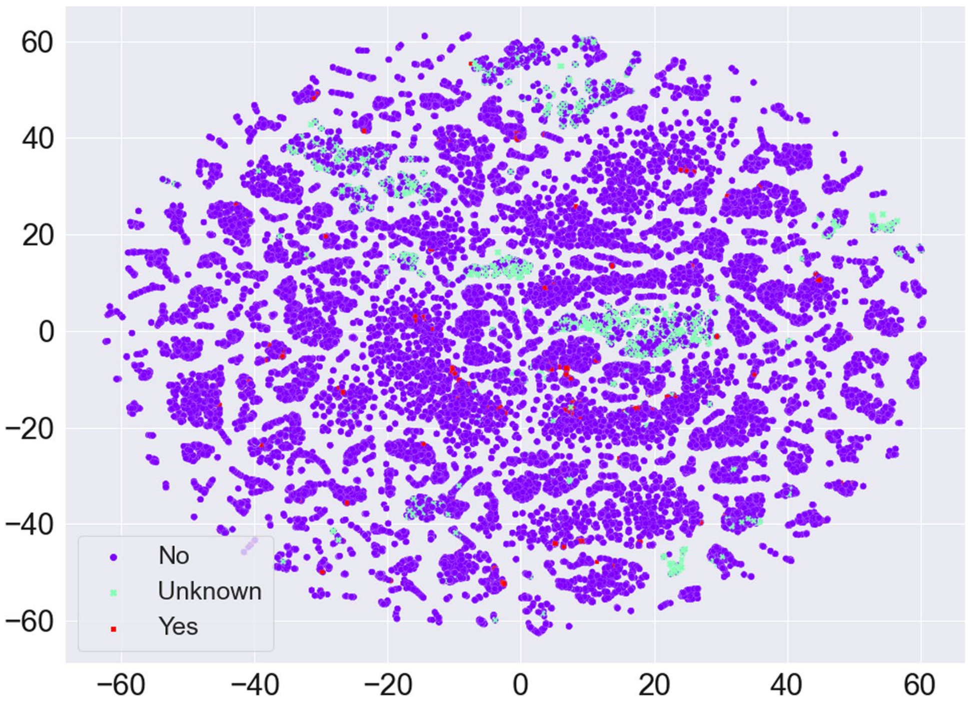


Table 3 Average classification results for Clinical Data1 (95984 patients). Best values are shown in bold

		Acc.	Prec.	Recall	F1 (Weighted)	F1 (Macro)	ROC AUC	Train Time (Sec.)
No Feat. Selec.	NB	0.78	0.93	0.78	0.83	0.49	0.80	0.19
	MLP	0.94	0.93	0.94	0.93	0.59	0.66	35.28
	KNN	0.94	0.93	0.94	0.93	0.60	0.69	4.71
	RF	0.94	0.94	0.94	0.94	0.64	0.71	4.88
	LR	0.93	0.87	0.93	0.90	0.32	0.50	1.38
	DT	0.93	0.93	0.93	0.93	0.62	0.73	0.37
Boruta	NB	0.83	0.94	0.83	0.87	0.54	**0.81**	0.149
	MLP	0.94	0.93	0.94	0.93	0.58	0.66	22.76
	KNN	0.94	0.94	0.94	0.94	0.62	0.70	1.814
	RF	**0.95**	**0.94**	**0.95**	**0.94**	**0.64**	0.72	3.346
	LR	0.93	0.89	0.93	0.90	0.33	0.50	0.968
	DT	0.94	0.94	0.94	0.94	0.64	0.73	0.227
RR	NB	0.84	0.93	0.84	0.87	0.45	0.72	**0.129**
	MLP	0.93	0.87	0.93	0.90	0.32	0.50	5.658
	KNN	0.93	0.92	0.93	0.92	0.48	0.60	1.660
	RF	0.94	0.93	0.94	0.93	0.51	0.64	2.214
	LR	0.93	0.87	0.93	0.90	0.32	0.50	0.338
	DT	0.94	0.93	0.94	0.93	0.51	0.64	0.154
RFF	NB	0.93	0.87	0.93	0.90	0.32	0.50	0.144
	MLP	0.93	0.89	0.93	0.90	0.32	0.50	24.22
	KNN	0.93	0.91	0.93	0.92	0.45	0.58	3.280
	RF	0.94	0.93	0.94	0.93	0.56	0.64	27.87
	LR	0.93	0.87	0.93	0.90	0.32	0.50	0.261
	DT	0.91	0.92	0.91	0.91	0.51	0.65	1.461
Keras Class.	-	0.93	0.87	0.93	0.90	0.32	0.50	11.582

Table 4 Standard deviation classification results for Clinical Data1 (95984 patients)

		Acc.	Prec.	Recall	F1 (Weig.)	F1 (Macro)	ROC AUC
No Feat. Selec.	NB	0.008176	0.003993	0.008176	0.006843	0.004180	0.002588
	MLP	0.002156	0.002679	0.002156	0.004162	0.005809	0.005512
	KNN	0.001271	0.001829	0.001271	0.001646	0.000916	0.000800
	RF	0.001647	0.001966	0.001647	0.001792	0.002193	0.001533
	LR	0.002894	0.004888	0.002894	0.004087	0.000851	0.000000
	DT	0.002333	0.002524	0.002333	0.002377	0.002928	0.001914
Boruta	NB	0.007808	0.001084	0.007808	0.00551	0.001686	0.0051
	MLP	0.002467	0.001547	0.002467	0.003225	0.016451	0.019329
	KNN	0.001507	0.002445	0.001507	0.002081	0.003452	0.004234
	RF	0.000776	0.000812	0.000776	0.000664	0.002328	0.001993
	LR	0.001721	0.002914	0.001721	0.002433	0.000505	0.00010
	DT	0.001746	0.001532	0.001746	0.001504	0.004998	0.00426
RR	NB	0.007741	0.002295	0.007741	0.005476	0.003985	0.001675
	MLP	0.002266	0.00272	0.002266	0.005542	0.016571	0.015832
	KNN	0.002009	0.001808	0.002009	0.00176	0.004993	0.004776
	RF	0.001906	0.001691	0.001906	0.001642	0.002921	0.003788
	LR	0.002098	0.003551	0.002098	0.002965	0.000616	0.002100
	DT	0.002505	0.002503	0.002505	0.002378	0.002449	0.003133
RFF	NB	0.007876	0.001905	0.007876	0.005162	0.00449	0.006048
	MLP	0.001429	0.001099	0.001429	0.004531	0.018336	0.018966
	KNN	0.00149	0.001927	0.00149	0.001736	0.003214	0.003358
	RF	0.001339	0.001612	0.001339	0.001479	0.00307	0.00353
	LR	0.001556	0.002629	0.001556	0.002198	0.000457	0.00340
	DT	0.00122	0.001468	0.00122	0.001202	0.002676	0.003387
Keras Class.	-	0.00213	0.004201	0.00201	0.000992	0.001967	0.002376

other feature selection approaches. In terms of training runtime, RFF with Logistic Regression classifier is performing better than the other classifiers. We note that, while Boruta feature selection is close to (No. Feat. Selec.) in the case of predictive performance, that it has a significantly shorter runtime. Since both predictive performance and runtime are important in an overwhelmed clinical setting where quick decisions are needed, a method which strikes a balance between the two is of most value. The overall performance gain for Boruta in case of RF classifier is 1% in terms of accuracy, 0.5% in terms of precision, 1% in terms of recall, 0.6% in terms of F1 weighted, 0.8% in terms of F1 macro, and 1% in terms of ROC-AUC as compared no “No Feature Selection” model.

The average and standard deviation results for Clinical Data2 are given in Table 5 and Table 6, respectively. To classify whether a patient is COVID-19 positive or negative, we can see that the random forest classifier with the Boruta feature selection approach outperforms all other feature selection methods. The overall performance gain for Boruta in case of RF classifier is 6% in terms of accuracy, 11% in terms of precision, 6% in terms of recall, 8% in terms of F1 weighted, 28% in terms of F1 macro, and 24% in terms of ROC-AUC as compared no “No Feature Selection” model. In terms of runtime, the logistic regression classifier with RFF outperforms other approaches.

6 Statistical analysis

To evaluate importance positions in spike sequences, we find the importance of each attribute with respect to class labels (for Clinical Data1). For this purpose, we computed the Information Gain (IG) between each attribute and the true class label. The IG is defined as follows:

$$IG(Class, position) = H(Class) - H(Class|position) \tag{5}$$

$$H = \sum_{i \in Class} -p_i \log p_i \tag{6}$$

where H is the entropy, and p_i is the probability of the class i. The IG values for different attributes (for Clinical Data1) are given in Fig. 5.

What is particularly interesting is that the State and County code are two major predictors of patient outcome (Clinical Data1). This is likely due to the current

Table 5 Average classification results for Clinical Data2 (608 patients). Best values are shown in bold

		Acc.	Prec.	Recall	F1 (Weighted)	F1 (Macro)	ROC AUC	Train Time (Sec.)
No Feat. Selec.	NB	0.89	0.88	0.89	0.88	0.71	0.70	0.025
	MLP	0.86	0.85	0.86	0.85	0.65	0.64	1.327
	KNN	0.88	0.87	0.88	0.87	0.68	0.66	0.013
	RF	0.85	0.79	0.85	0.82	0.49	0.50	0.178
	LR	0.87	0.86	0.87	0.86	0.65	0.63	0.013
	DT	0.81	0.81	0.81	0.81	0.56	0.56	0.01
Boruta	NB	0.83	0.89	0.83	0.85	0.71	0.79	0.01
	MLP	0.87	0.89	0.87	0.88	0.75	0.78	1.621
	KNN	0.86	0.85	0.86	0.86	0.66	0.66	0.015
	RF	**0.91**	**0.90**	**0.91**	**0.90**	**0.77**	**0.74**	0.125
	LR	0.87	0.88	0.87	0.88	0.73	0.74	0.01
	DT	0.85	0.86	0.85	0.86	0.68	0.69	0.007
RR	NB	0.83	0.80	0.83	0.81	0.57	0.56	0.016
	MLP	0.85	0.84	0.85	0.85	0.67	0.66	1.024
	KNN	0.85	0.84	0.85	0.84	0.66	0.64	0.01
	RF	0.85	0.83	0.85	0.84	0.65	0.64	0.137
	LR	0.87	0.84	0.87	0.84	0.61	0.59	0.009
	DT	0.82	0.82	0.82	0.82	0.62	0.62	0.009
RFF	NB	0.89	0.78	0.89	0.83	0.47	0.50	0.022
	MLP	0.77	0.79	0.77	0.78	0.48	0.48	1.565
	KNN	0.86	0.80	0.86	0.83	0.50	0.51	0.019
	RF	0.88	0.78	0.88	0.83	0.47	0.50	0.163
	LR	0.89	0.78	0.89	0.83	0.47	0.50	**0.008**
	DT	0.73	0.80	0.73	0.76	0.50	0.52	0.009
Keras Class.	-	0.83	0.76	0.83	0.79	0.48	0.50	10.928

vaccination situation in the USA, which varies quite widely from county to county [42]. The IG values for Clinical Data2 are given in Table 6. We can observe that four attributes, namely "Platelets", "Monocytes", "Leukocytes", and "Eosinophils" are most important towards prediction of "SARS-Cov-2 Exam Result" attribute in the Clinical Data2 (Fig. 6).

Since information gain does not give us the negative (or opposite) contribution of a feature corresponding to the class label, we use other statistical measures such as Pearson correlation [43] and Spearman correlation [44] to further evaluate the contribution of features in the dataset towards the prediction of labels. The Pearson Correlation is computed using the following expression:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2(y_i - \bar{y})^2}} \tag{7}$$

where r is the correlation coefficient, x_i is the values of the x-variable in a sample, $\bar{x}$ is the mean of the values of the x-variable, y_i is the values of the y-variable in a sample, and $\bar{y}$ is the mean of the values of the y-variable. The Spearman Correlation is computed using the following expression:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)} \tag{8}$$

where ρ is the Spearman's rank correlation coefficient, d_i is the difference between the two ranks of each observation, and n is the total number of observations.

The Pearson correlation values are given in Fig. 7 (for Clinical Data1). Similarly, the Spearman correlation values are given in Fig. 8 (for Clinical Data1). The Pearson and Spearman correlation values for Clinical Data2 are given in Fig. 9 and Fig. 10, respectively. We can observe that most of the attributes/features are contributing towards the prediction of labels, and only a few attributes have correlation values close to zero (in the case of both Pearson and Spearman correlations.

6.1 SHAP analysis

We also use SHAP analysis [9–11] to understand how significant each factor in determining the final label prediction of the model outputs. For this purpose,

Table 6 Standard deviation classification results for Clinical Data2 (608 patients)

		Acc.	Prec.	Recall	F1 (Weig.)	F1 (Macro)	ROC AUC
No Feat. Selec.	NB	0.007186	0.002933	0.007166	0.005813	0.003170	0.003598
	MLP	0.002526	0.002459	0.001856	0.003863	0.004708	0.005815
	KNN	0.001165	0.001727	0.001574	0.001247	0.001019	0.000974
	RF	0.001245	0.001764	0.001449	0.001995	0.002498	0.001431
	LR	0.003854	0.004189	0.002496	0.003057	0.000757	0.000109
	DT	0.003231	0.003421	0.002234	0.002576	0.003026	0.001619
Boruta	NB	0.006806	0.002084	0.006804	0.00355	0.002666	0.004101
	MLP	0.002145	0.001643	0.001964	0.003124	0.017452	0.016315
	KNN	0.001305	0.002542	0.001701	0.002151	0.002451	0.005231
	RF	0.000873	0.000711	0.000674	0.000761	0.003358	0.002194
	LR	0.001951	0.003215	0.001629	0.002234	0.000403	0.001109
	DT	0.002143	0.001235	0.001348	0.001601	0.005979	0.00514
RR	NB	0.006721	0.003491	0.006731	0.001475	0.002945	0.002625
	MLP	0.002156	0.00191	0.003246	0.006541	0.015572	0.016831
	KNN	0.003016	0.001914	0.001405	0.00293	0.003923	0.005786
	RF	0.002501	0.002661	0.001401	0.001741	0.002962	0.002948
	LR	0.002361	0.002599	0.003065	0.002371	0.000557	0.003016
	DT	0.003163	0.002317	0.002619	0.003545	0.003145	0.002931
RFF	NB	0.006134	0.002357	0.004826	0.008152	0.00344	0.0040483
	MLP	0.003459	0.003049	0.002479	0.003511	0.019316	0.014716
	KNN	0.00237	0.002421	0.00264	0.001831	0.002918	0.002951
	RF	0.001731	0.001811	0.001438	0.001578	0.00206	0.002511
	LR	0.002556	0.003627	0.002576	0.001197	0.000551	0.00249
	DT	0.00225	0.001963	0.00224	0.002247	0.001973	0.002356
Keras Class.	-	0.00194	0.003251	0.00361	0.001062	0.001562	0.002471

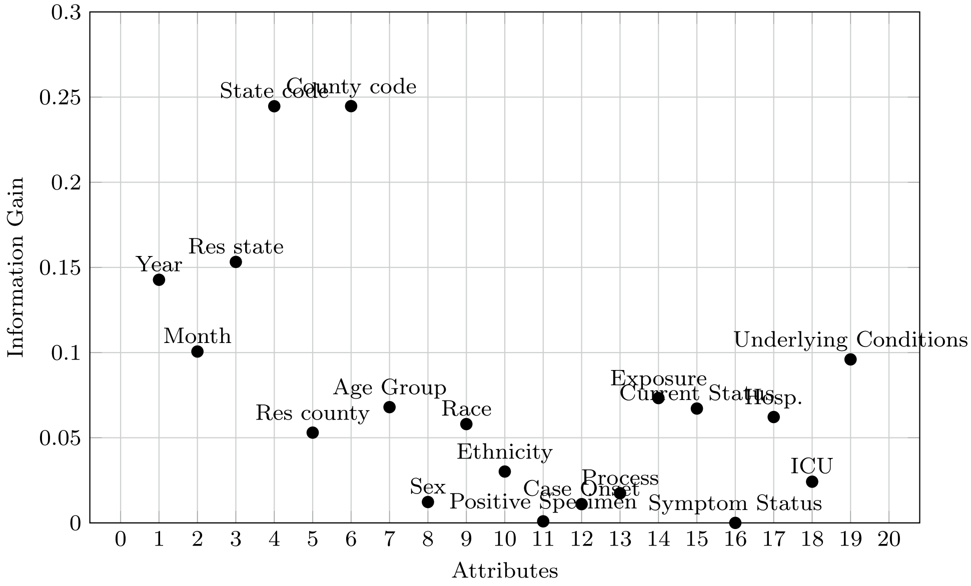


Fig. 5 Information Gain of different attributes with respect to Class label (deceased attribute) for Clinical Data1

SHAP analysis runs a large number of predictions and compares a variable's impact against the other features. The SHAP analysis for COVID Data1 (for the prediction of deceased label) is given in Fig. 11.

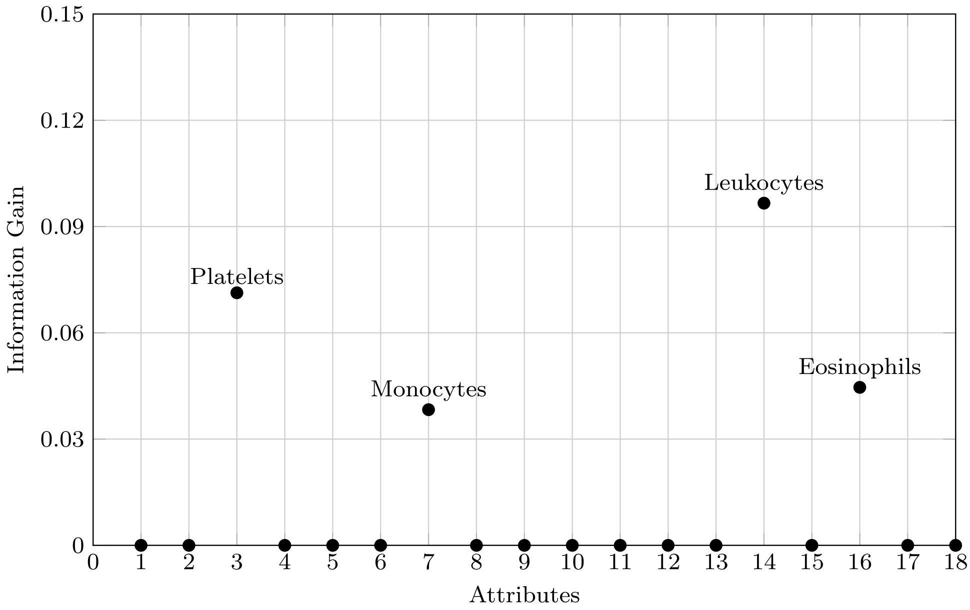


Fig. 6 Information Gain of different attributes with respect to Class label (SARS-Cov-2 Exam Result) for Clinical Data2

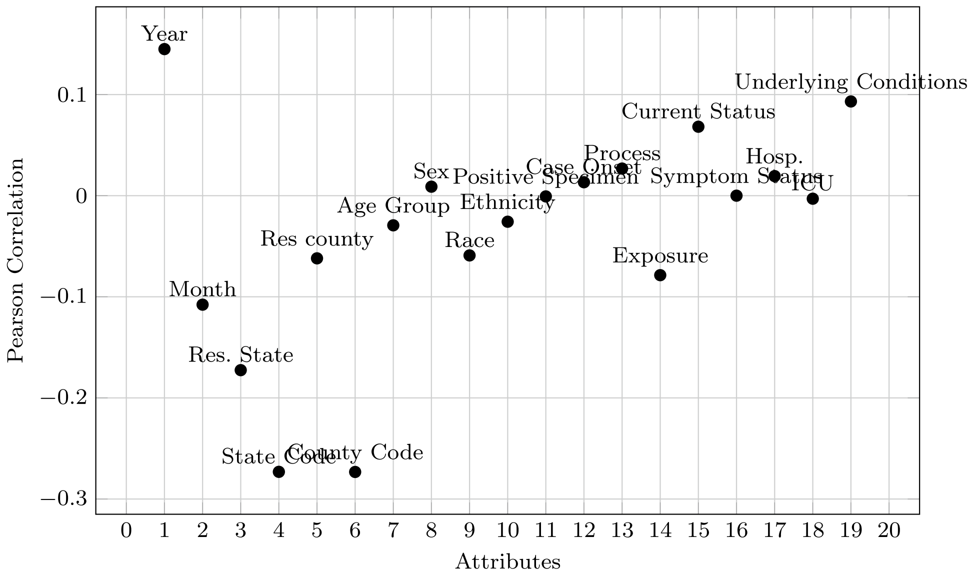


Fig. 7 Pearson Correlation for Clinical Data1

We can observe that the attributes “Sex”, “Age Group”, and “County Fips Code” play a major role in predicting different classes of the “Deceased” label. The SHAP analysis for COVID Data2 (for the prediction of covid positive/negative status)) is given in Fig. 12. We can observe that attributes “Leukocytes” and “Eosinophils” play a major role in predicting different classes in the “SARS-Cov-2 Exam Result” attribute. The code for SHAP analysis is also available online [4].

7 Conclusion

We propose an efficient model for the classification of COVID-19 patients using efficient feature selection methods and machine learning classification algorithms. With Boruta for feature selection, we show that simple classification algorithms like the random forest can also outperform the keras classification model when the dataset size is not too big. We also show the importance of each attribute in Clinical Data1 by computing the information gain values for each attribute corresponding to the class label. In the future, we will extract more data and apply sophisticated deep learning

[4] https://github.com/slundberg/shap

Fig. 8 Spearman Correlation for Clinical Data1

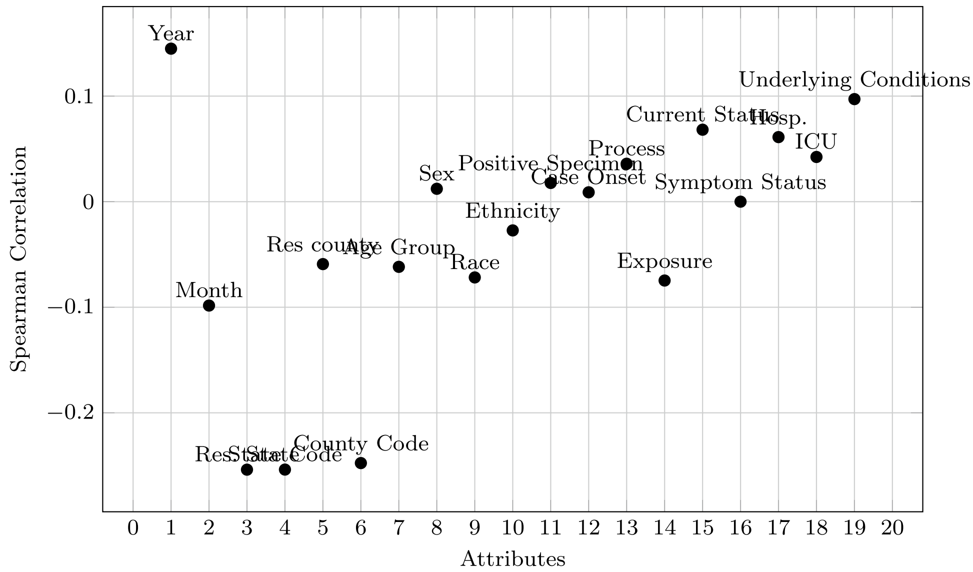
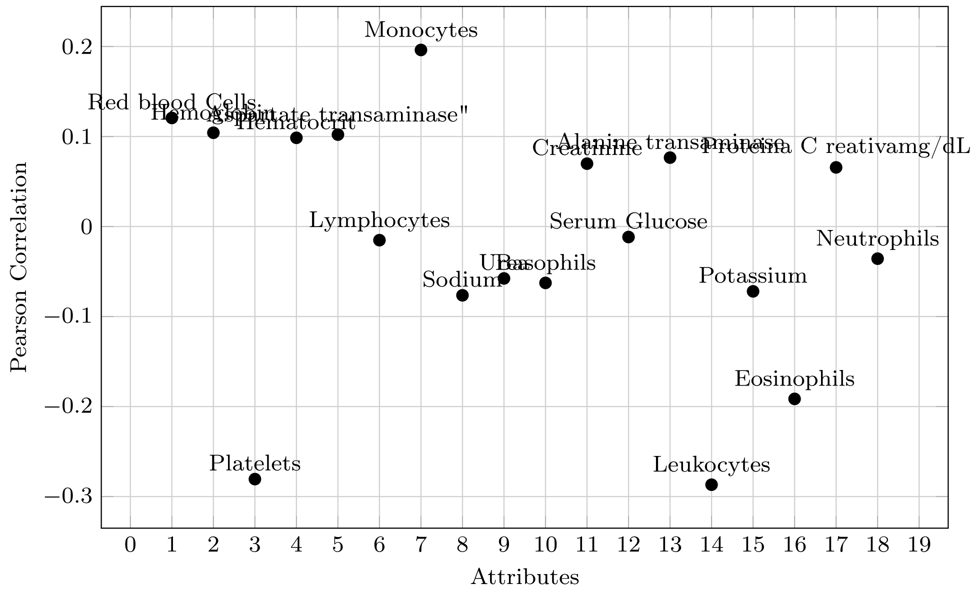


Fig. 9 Pearson Correlation for Clinical Data2



models such as LSTM and GRU to improve the predictive performance. We will also use other factors such as weather along with the clinical data to further improve the classification results.

These results have many practical meanings. The most direct real-world application of the machine learning model is to provide support to medical doctors during the COVID-19 pandemic. By predicting the risk level of individual patients, our model enables clinicians to assign wisely, in real-time, limited medical resources, especially during periods of medical shortage, and provide immediate treatment to the most vulnerable groups. With the help of the risk prediction system, clinicians learn which individual patients may be in danger of death and can thus conduct personalized prevention treatment in due time. Moreover, our research can be used to build a general clinical decision support system that serves not only COVID-19 but also other potential future pandemics. The patterns found in this data may also help biologists to design effective vaccines or vaccination strategies. Finally, these methodologies can be applied for future studies on big data and machine learning in the broader sense.

Fig. 10 Spearman Correlation for Clinical Data2

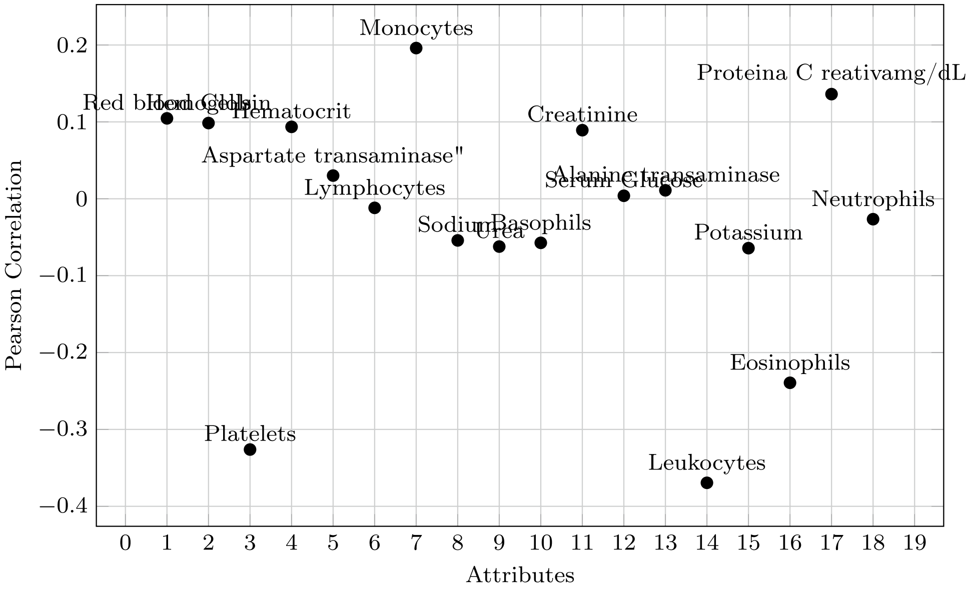


Fig. 11 Mean absolute value for the SHAP values for Clinical Data1 (for "Deceased" label)

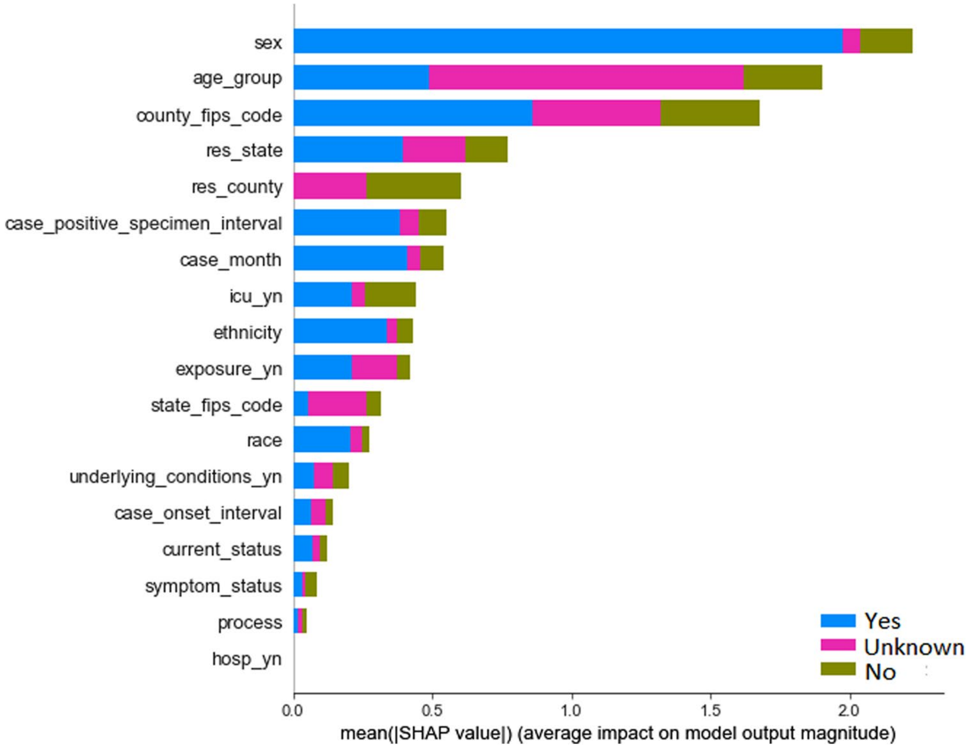
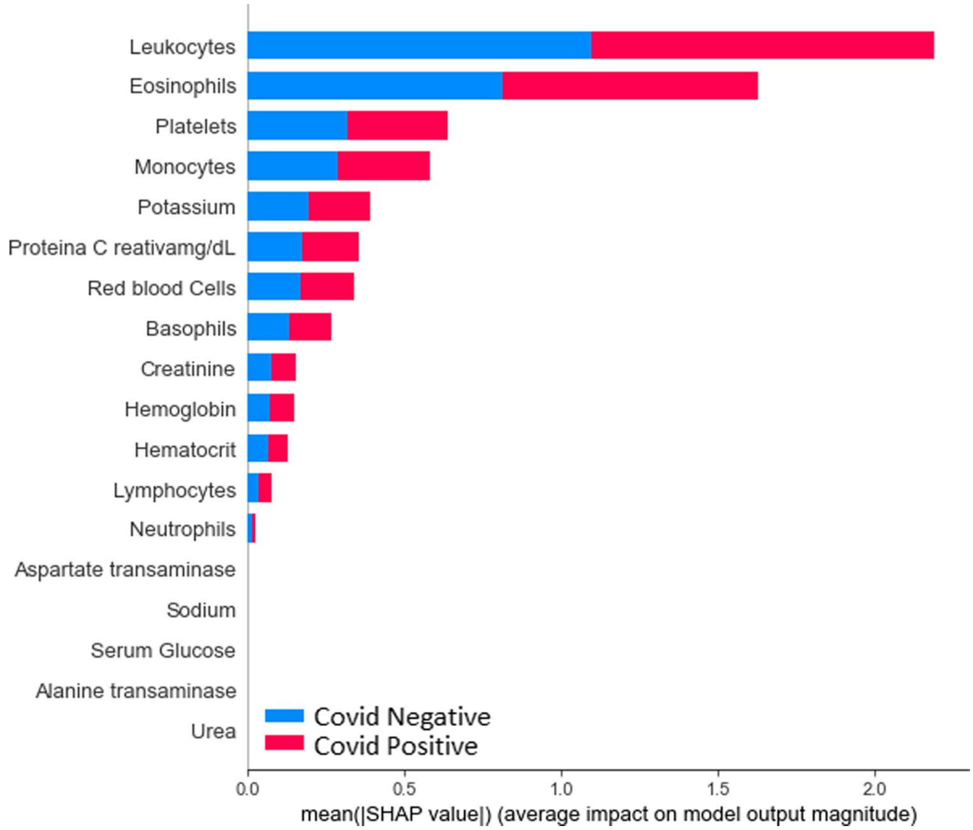


Fig. 12 Mean absolute value for the SHAP values for Clinical Data2 (for "SARS-Cov-2 Exam Result" label)



References

1. Ali S, Patterson M. Spike2vec: An efficient and scalable embedding approach for covid-19 spike sequences. In 2021 IEEE International Conference on Big Data (Big Data) 2021 Dec 15 (pp. 1533-1540).
2. Ali S, Bello B, Patterson M (2021a) Classifying covid-19 spike sequences from geographic location using deep learning. arXiv preprint arXiv:211000809
3. GISAID Website (Accessed: 10-12-2021) . https://www.gisaidorg/
4. Leung CK, Chen Y, Hoi CS, Shang S, Cuzzocrea A (2020a) Machine learning and olap on big covid-19 data. In: 2020 IEEE International Conference on Big Data (Big Data), pp 5118–5127
5. Leung CK, Chen Y, Shang S, Deng D (2020b) Big data science on covid-19 data. In: 2020 IEEE 14th International Conference on Big Data Science and Engineering (BigDataSE), pp 14–21
6. Ali S, Mansoor H, Arshad N, Khan I (2019a) Short term load forecasting using smart meter data. In: International Conference on Future Energy Systems, pp 419–421
7. Ali S, Mansoor H, Khan I, Arshad N, Khan MA, Faizullah S (2019b) Short-term load forecasting using ami data. arXiv preprint arXiv:191212479
8. Abdulkareem KH, Mohammed MA, Salim A, Arif M, Geman O, Gupta D, Khanna A (2021) Realizing an effective covid-19 diagnosis system based on machine learning and iot in smart hospital environment. IEEE Internet of Things Journal
9. Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in Neural Information Processing Systems 30, pp 4765–4774
10. Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, Liston DE, Low DKW, Newman SF, Kim J et al (2018) Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat Biomed Eng 2(10):749
11. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee SI (2020) From local explanations to global understanding with explainable ai for trees. Nat Mach Intell 2(1):2522–5839
12. Ali S, Shakeel MH, Khan I, Faizullah S, Khan MA (2021) Predicting attributes of nodes using network structure. ACM Trans Intell Syst Technol 12(2):1–23
13. Grover A, Leskovec J (2016) node2vec: Scalable feature learning for networks. In: International Conference on Knowledge Discovery & Data Mining (KDD), pp 855–864
14. Yang L, Guo Y, Cao X (2018) Multi-facet network embedding: Beyond the general solution of detection and representation. In: AAAI Conference on Artificial Intelligence (AAAI), pp 499–506
15. Alakus TB, Turkoglu I (2020) Comparison of deep learning approaches to predict covid-19 infection. Chaos, Solitons & Fractals 140:110120
16. Ahmad M, Ali S, Tariq J, Khan I, Shabbir M, Zaman A (2020) Combinatorial trace method for network immunization. Inf Sci 519:215–228
17. Ullah A, Ali S, Khan I, Khan MA, Faizullah S (2020) Effect of analysis window and feature selection on classification of hand movements using EMG signal. In: SAI Intelligent Systems Conference (IntelliSys), pp 400–415
18. Shakeel MH, Karim A, Khan I (2019) A multi-cascaded deep model for bilingual sms classification. In: International Conference on Neural Information Processing, pp 287–298
19. Shakeel MH, Faizullah S, Alghamidi T, Khan I (2020a) Language independent sentiment analysis. In: 2019 International Conference

on Advances in the Emerging Computing Technologies (AECT), pp 1–5

20. Shakeel MH, Karim A, Khan I (2020b) A multi-cascaded model with data augmentation for enhanced paraphrase detection in short texts. Information Processing & Management 57(3):102204
21. Hassan IU, Haseeb A, Ali S (2021) Locally weighted mean phase angle (lwmpa) based tone mapping quality index (tmqi-3). Accepted at: International Conference on Intelligent Vision and Computing (ICIVC)
22. Leung CK, Fung DL, Mushtaq SB, Leduchowski OT, Bouchard RL, Jin H, Cuzzocrea A, Zhang CY (2020c) Data science for healthcare predictive analytics. In: Proceedings of the 24th Symposium on International Database Engineering & Applications, pp 1–10
23. Ali S, Sahoo B, Ullah N, Zelikovskiy A, Patterson M, Khan I (2021d) A k-mer based approach for sars-cov-2 variant identification. In: International Symposium on Bioinformatics Research and Applications, pp 153–164
24. Ali S, Ali TE, Khan MA, Khan I, Patterson M. Effective and scalable clustering of SARS-CoV-2 sequences. In 2021 the 5th International Conference on Big Data Research (ICBDR) 2021 Sep 25 (pp. 42-49).
25. Tayebi Z, Ali S, Patterson M (2021) Robust representation and efficient feature selection allows for effective clustering of sars-cov-2 variants. Algorithms 14(12):348
26. Kuzmin K, Adeniyi AE, DaSouza Jr AK, Lim D, Nguyen H, Molina NR, Xiong L, Weber IT, Harrison RW (2020) Machine learning methods accurately predict host specificity of coronaviruses based on spike sequences alone. Biochem Biophys Res Commun 533(3), 553–558
27. Shah V, Keniya R, Shridharani A, Punjabi M, Shah J, Mehendale N (2021) Diagnosis of covid-19 using ct scan images and deep learning techniques. Emerg Radiol 28(3):497–505
28. Zaffino P, Marzullo A, Moccia S, Calimeri F, De Momi E, Bertucci B, Arcuri PP, Spadea MF (2021) An open-source covid-19 ct dataset with automatic lung tissue classification for radiomics. Bioengineering 8(2):26
29. Teli MN (2021) Telinet: Classifying ct scan images for covid-19 diagnosis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 496–502
30. Panwar H, Gupta P, Siddiqui MK, Morales-Menendez R, Bhardwaj P, Singh V (2020) A deep learning and grad-cam based color visualization approach for fast detection of covid-19 cases using chest x-ray and ct-scan images. Chaos, Solitons & Fractals 140:110190
31. Albahri AS, Hamid RA, Alwan JK, Al-Qays Z, Zaidan A, Zaidan B, Albahri A, AlAmoodi A, Khlaf JM, Almahdi E, et al. (2020) Role of biological data mining and machine learning techniques in detecting and diagnosing the novel coronavirus (covid-19): a systematic review. J Med Syst 44:1–11
32. Li WT, Ma J, Shende N, Castaneda G, Chakladar J, Tsai JC, Apostol L, Honda CO, Xu J, Wong LM, et al. (2020) Using machine learning of clinical data to diagnose covid-19: a systematic review and meta-analysis. BMC medical informatics and decision making 20(1):1–13
33. Fung DL, Hoi CS, Leung CK, Zhang CY (2021) Predictive analytics of covid-19 with neural networks. In: 2021 International Joint Conference on Neural Networks (IJCNN), pp 1–8
34. Ali S (2021) Cache replacement algorithm. arXiv preprint arXiv:210714646
35. Kursa MB, Rudnicki WR, et al. (2010) Feature selection with the boruta package. J Stat Softw 36(11), 1–13
36. Hoerl AE, Kannard RW, Baldwin KF (1975) Ridge regression: some simulations. Communications in Statistics-Theory and Methods 4(2), 105–123
37. Rahimi A, Recht B, et al. (2007) Random features for large-scale kernel machines. In: NIPS, vol 3, p 5
38. Ali S, Ciccolella S, Lucarella L, Vedova GD, Patterson M (2021b) Simpler and faster development of tumor phylogeny pipelines. J Comput Biol 28(11), 1142–1155
39. McDonald GC (2009) Ridge regression. Wiley Interdisciplinary Reviews: Comput Stat 1(1), 93–100
40. Devijver P, Kittler J (1982) Pattern recognition: A statistical approach. In: London, GB: Prentice-Hall, pp 1–448
41. Van der M L, Hinton G (2008) Visualizing data using t-SNE. J Mach Learn Res (JMLR) 9(11)
42. NewYork Times (NYT) (2021) https://www.nytimes.com/interactive/2020/us/covid-19-vaccine-doses.html, [Online; Accessed: 15-12-2021]
43. Benesty J, Chen J, Huang Y, Cohen I (2009) Pearson correlation coefficient. In: Noise reduction in speech processing, pp 1–4
44. Myers L, Sirois MJ. Spearman correlation coefficients, differences between. Encyclopedia of statistical sciences. 2004 Jul 15;12.

Sarwan Ali is a Ph.D. student at Department of Computer Science, Georgia State University, working in the field of Bioinformatics, Data mining, Big data, and Machine Learning.

Yijing Zhou is an Undergraduate student at Georgia State University, studying Computer Science and Biology. She is working in the fields of Bioinformatics, Computational Biology, and Combinatorics.

Murray Patterson is an Assistant Professor at Georgia State University working in the fields of Bioinformatics, Computational Biology, Algorithms, and Combinatorics.