Comprehensive Machine Learning Analysis on the Phenotypes of COVID-19 Patients Using Transcriptome Data

Pratheeba Jeyananthan

Department of Computer Engineering, University of Jaffna, Sri Lanka.

E-mail: pratheeba@eng.jfn.ac.lk

Received: 08/11/2021
Revised: 08/02/2022
Accepted: 28/02/2022

Abstract

Purpose: Evolving technologies allow us to measure human molecular data in a wide reach. Those data are extensively used by researchers in many studies and help in advancements of medical field. Transcriptome, proteome, metabolome, and epigenome are few such molecular data. This study utilizes the transcriptome data of COVID-19 patients to uncover the dysregulated genes in the SARS-COV-2.

Method: Selected genes are used in machine learning models to predict various phenotypes of those patients. Ten different phenotypes are studied here such as time since onset, COVID-19 status, connection between age and COVID-19, hospitalization status and ICU status, using classification models. Further, this study compares molecular characterization of COVID-19 patients with other respiratory diseases.

Results: Gene ontology analysis on the selected features shows that they are highly related to viral infection. Features are selected using two methods and selected features are individually used in the classification of patients using six different machine learning algorithms. For each of the selected phenotype, results are compared to find the best prediction model.

Conclusion: Even though, there are not any significant differences between the feature selection methods, random forest and SVM performs very well throughout all the phenotype studies.

Keywords: COVID-19, Transcriptome data, Phenotype analysis, Machine learning models, Respiratory diseases, Dysregulated genes.

Introduction

Due to its high mortality rate and high spreading rate, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) poses a critical challenge to public health (Li, Liu, Yu, Tang, & Tang, 2020). Even though studies show that the vaccines reduce the severity of the disease, the world continues to suffer in controlling the spread of this disease (Christie, et al., 2021). As of 27th August 2021, there are 214, 468, 601 total confirmed cases and 4, 470, 969 total deaths reported to World Health Organization. Further, they reported that 4, 953, 887, 422 vaccine doses have been administered till 24 August 2021. This pandemic has a great impact not only on health, but also on economics, politics, social and other important aspects of many countries (Tisdell, 2020; Padhan & Prabheesh, 2021).

This pandemic has a significant effect on the clinical trials and clinical research (Sathian, et al., 2020; Fu, et al., 2020). Contemporary studies progress in various directions to find a solution to this global health issue. It opens out in many directions such as vaccination related studies, finding drugs, study on post complications of COVID-19, studies on molecular landscapes of the patients, biomarker identification and dysregulated genes of the patients. Molecular data of the patients are used in almost all these areas.

The field of biology becomes data-driven mainly due to the vast amount of molecular data currently used in clinical research. These molecular data have already shown clinical significance, pathological significance and physiological significances in many diseases including cancers. In this sequence, now molecular data are being vastly used in the COVID-19 related studies. For example, analyzing RNA data of 27 different tissues showed that ACE2 gene is a receptor of SARS-COV-2 virus (Islam & Khan, 2020). Further, they showed that this gene plays a potential anti-tumor role in cancer. Besides, transcriptome data is used to propose novel therapies to COVID-19 and in identification of new consequences of COVID-19 (Moni, Lin, Quinn, & Eapen, 2021). Studying the transcriptome data also showed a persistent neuroinflammation in acute COVID-19 patients (Fullard, et al., 2021), and the comorbidity of COVID-19 patients with psychiatric disorders (Shen, et al., 2020).

Apart from transcriptomic studies, research on other molecular data such as proteome and epigenome are also very active and lead to many significant identifications. Some characteristic proteins and metabolite changes are identified with the potential to be used as biomarkers of severity prediction (Völlmy, et al., 2021). Serum protein also used in the prediction of the mortality of severe COVID-19 patients (Shirvaliloo, 2021). Likewise, DNA methylation is used in this field to identify new methylations of the infection and in the predictions of other complications (Castro de Moura, et al., 2021; Nagpal & Singh, 2018). Research on multi-omic data also provide very prominent insights in this field.

This study uses two different COVID-19 transcriptome datasets to predict the clinical outcomes of the patient. Those two sets of data were measured on completely different sets of patients and provided with distinct sets of phenotype data. After the preprocessing of data, for each clinical outcome, 100 differently expressed genes are selected using two different feature selection techniques, mutual information, and feature importance. Those selected genes are individually studied using gene ontology (GO) analysis to check their enrichment functions. After that, selected features are used with six classification models to check their ability in the corresponding prediction. Performance of various machine learning algorithms are compared in each prediction to choose the best model.

Materials and Methods

1. Material

Both sets of data are acquired from Gene Expression Omnibus (GEO) with the accession numbers GSE157103 and GSE161731. They are high throughput RNA-seq transcriptome data of COVID-19 patients along with control data. The former one used plasma and leukocyte samples from hospitalized patients, while the latter one is measured on peripheral blood sample.

GSE157103

This set contains 126 samples from 100 COVID-19 positive patients and 26 negative patients. Among them, 66 patients were admitted in ICU and 60 of them were not admitted. Transcriptome was measured on over 17 000 genes. This data was presented with four different phenotype data: COVID status (have/not), age, gender, and ICU status (admitted/not).

GSE161731

This data consists of 195 samples of COVID-19 patients, patients with other respiratory diseases and healthy individuals. Transcriptome of 77 COVID-19 patients were measured along with 23 bacterial pneumonia, 17 influenza, 59 seasonal corona virus and 19 healthy controls. Among those patients 12 were hospitalized and 65 were not hospitalized. Others were under the category of hospitalization not seen necessary. Moreover, 19 COVID-19 patients were in early stage (<= 10 days), 36 patients in middle stage (11 – 21 days) and 22 longer patients (> 21 days). Altogether 93 male patients and 82 female patients are used in this study. More than 15 000 genes were measured with comparably more phenotype data than the previous one.

2. Data Preprocessing

Few preprocessing steps are applied before feeding the data into the machine learning model.

i. Feature Scaling

GSE161731 is already normalized dataset, which is ready to use (Figure S1). However, GSE157103 was not normalized, and the data spread over a wide range (Figure S2). Hence it is normalized using min-max scalar.

Min-Max Scalar:

In this normalization, data is scaled to the range of 0 and 1. Equation 1 is used here to reduce the difference between the expression levels of features.

$$\boldsymbol{x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}}$$ *Equation 1*

ii. Feature Selection Methods

As the transcriptome is measured on over 15,000 genes, mutual information and feature importance are used to select top 100 features and they are individually used in the machine learning models.

- **Mutual Information**

Mutual information calculates the dependency between two variables. This is widely used for feature selection in many fields such as computational biology, image processing and speech recognition. It has been used in bioinformatics for gene selection (Jansi Rani & Devaraj, 2019; Ng, et al., 2021). Mutual Information can be measured using Equation 2, where is the entropy.

$$\begin{aligned} \boldsymbol{I(X;Y)} &= \boldsymbol{H(X) - H(X|Y)} \\ &= \boldsymbol{H(Y) - H(Y|X)} \\ &= \boldsymbol{H(X) + H(Y) - H(X,Y)} \end{aligned}$$ *Equation 2*

Entropy measures the expected uncertainty in a random variable. X

$$\boldsymbol{H(X) = \sum_{i=1}^{n} P(x_i)\ \log P(x_i)}$$

Mutual information is used in the feature selection of each task and the selected features are presented from Table S-2.1 to Table S-2.10. Mutual information values are plotted against selected features and illustrated from Figure S2-1 to Figure S2-20.

- **Feature Importance**

Feature importance is another feature selection method with the capacity of selecting the most relevant features of a particular target. The feature importance is used in this study along with random forest classifier during feature selection and the selected features are used in various models. Selected features are given from Table S-1.1 to Table S-1.10. Illustration of feature importance values are given from Figure S1-3 to Figure S1-40.

3. Machine Learning Algorithms

Six different classification algorithms, one with two different kernels are used in this classification-oriented study.

i. Support Vector Machine (SVM)

SVM is one of the classic supervised machine learning algorithms for binary classification tasks. This is applied to linearly separable data. SVM finds a hyper plane to separate the given set of data into two classes. Our objective here is to find the weight vector which is normal to the hyperplane. Finding the best value for is an optimization problem described by Equation 3.

$$\boldsymbol{min} \quad \frac{\mathbf{1}}{\mathbf{2}} \left|\left|\boldsymbol{w}\right|\right|^2$$

$$\boldsymbol{s.t.} \qquad \boldsymbol{y_i(w.X_i + b) - 1 = 0}$$

Equation 3

As this is designed to stratify linearly separable data, SVM kernels are available to handle data with different properties. This study uses polynomial kernel and linear SVM.

ii. Naïve Bayes Classifier

This classifier is defined by probabilistic classification models based on Bayes theorem. They are simple classifiers and can attain higher accuracies using kernel density estimation. Basically, it uses Bayes theorem under the assumption that all feature values are independent given the target.

iii. Decision Tree

This is a tree-based classifier, predicts by questioning (yes/no question) on each feature. As it is in the tree structure, first step would be finding out the root node attribute on the top. To select the best attribute, algorithm scans all attributes and their values. The one leads to best split of the data would be selected as the root node. After this selection, data will be split into two classes and algorithm will find the next feature in the same process. In general, this algorithm continues either until each sample is classified or until it encounters a specified stopping criteria.

iv. Random Forest Classifier

This is a simple classification model under ensemble learning, yields comparably higher performance. In ensemble method, we combine more than one machine learning algorithms or execute the same algorithm multiple times to get a more powerful model than the original. In random forest, multiple decision trees are used, obtain accuracies from all the trees and the final output would be the average of all the prediction values.

v. K-Nearest Neighbor Classifier (KNN)

This is a non-parametric classification algorithm under supervised learning. This classifier makes prediction based on the number of closest objects of the new instance. New object will be assigned to the class with maximum number of neighbors. To find the closest object, some distance measures such as Euclidean distance is used. Number of nearest objects should be defined by the programmer, wherein this study has two.

vi. Perceptron

This is a supervised algorithm for binary classifiers with single layer and multilayer perceptron, where the application of former is restricted to linearly separable data. Generally, their prediction depends on a threshold. Perceptron function is given by Equation 4.

$$\boldsymbol{f(x)} = \begin{cases} \mathbf{1}, & \boldsymbol{if\ w.x + b > 0} \\ \mathbf{0}, & \boldsymbol{otherwise} \end{cases} \qquad \textit{Equation 4}$$

4. Accuracy Measures

Accuracy is used as the performance measure in these classification models.

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{True\ positive + True\ negative}}{\boldsymbol{True\ positive + True\ negative + False\ positive + False\ negative}}$$

5. Cross validation (CV)

Cross validation is used in machine learning to validate the model. It helps to interpret the result with a confidence level. Even though 10-fold cross validation is used in this study, leave one out cross validation is also used to confirm the result, as we have comparably low number of samples.

Results

This section describes the GO terms related to selected features and their performance on different models. This study uses ten different phenotypes such as time since onset, disease cohort, health status, comparison between COVID-19 and healthy people, stratification of different respiratory diseases, gender specificity of COVID-19 patients, hospitalization requirement of the COVID-19 patients, age classification between COVID-19 patients, COVID-19 status, and ICU status. Here, first eight targets are received from the dataset GSE161731 and latter two are from GSE157103. More details of the dataset can be retrieved from Gene Expression Omnibus (GEO).

1. Time since onset:

This phenotype contains 3 different classes of patients, early, middle, and longer. Subjects identified with COVID-19 within 10 days are considered as early patients, who lies between 11 and 21 days are considered as middle patients, and greater than 21 days are considered as late patients.

First step of this study is selecting appropriate set of features for this prediction. Figure S1-5 and Figure S2-6 show the distribution of mutual information and feature importance of the selected features accordingly. To confirm the propriety of the selected features, they are subjected to a GO analysis.

GO analysis on the features selected using feature importance shows that they are closely related to viral and bacterial related activities such as toll-like receptor TLR6:TLR2 signaling pathways (Gardinassi, Souza, Sales-Campos, & Fonseca, 2020; Overmyer, et al., 2021), response to diacyl bacterial lipopeptide, cellular response to diacyl bacterial lipopeptide and hematopoietic stem cell homeostasis (Table S-1.11). This list elongates with further GO terms related to immune response (Gardinassi, Souza, Sales-Campos, & Fonseca, 2020; Overmyer, et al., 2021; Liu, Jia, Fang, & Zhao, 2020; Sardar, Sharma, & Gupta,, 2021), viral related activities (Gardinassi, Souza, Sales-Campos, & Fonseca, 2020; Loganathan, Ramachandran, Shankaran, Nagarajan, & Mohan, 2020; Jain, et al., 2021), T cell proliferation (Gardinassi, Souza, Sales-Campos, & Fonseca, 2020; Overmyer, et al., 2021; Loganathan, Ramachandran, Shankaran, Nagarajan, & Mohan, 2020)and blood coagulation (Liu, Jia, Fang, & Zhao, 2020; Sardar, Sharma, & Gupta,, 2021).

Function enrichment of features selected by mutual information gives even more viral-related functions (Table S-2.11) with farther appropriate p-value (<0.1) such as viral process, viral genome replication, viral life cycle, many immune related processes, and cellular processes. As the selected features show remarkable outcomes in this study, it was hypothesized that it could be able to construct a classifier that accurately discriminate the stage of the COVID-19 patients.

Hence selected features are fed to six different machine learning models. Although the selected genes show prominent enrichment functions, maximum accuracy of this stratification is 0.79 ± 0.13 by support vector machine using polynomial kernel (Figure 1). As the number of samples are very low, this result is confirmed with leave one out cross validation (LOOCV) with the accuracy of 0.77 (Table 2). However, random forest (0.75 ±0.09) and linear SVM (0.78 ±0.14) closely perform to the maximum accuracy. Both performances are on mutual information selected features.

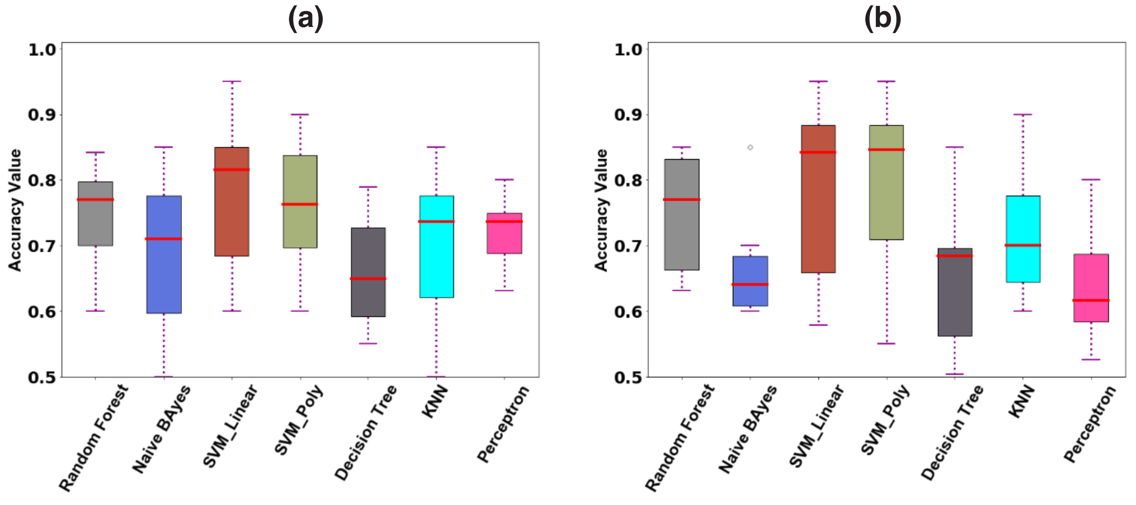


Figure 1. Selected 100 features are used in the classification of patients into their disease stage. Two different feature selection techniques are used. (a) Performance of Feature Importance selected features (b) Mutual information selected features on this classification with various classification algorithms.

2. Disease cohort

This dataset has patients with different respiratory diseases such as bacterial, influenza, COVID-19 and other seasonal COVID along with healthy individuals. Data preprocessing and the steps followed are same throughout the study.

Features selected with feature importance study (Table S-1.2) show GO enrichment functions mostly related to golgi, glycosylation and demannosylation, which are not related to any viral activities and not identified in the literature. However, features selected using mutual information (Table S-2.2) are more viral-related, like viral gene expression, viral process, and viral life cycle. This list also contains some new functions such as beta-catenin—CF complex, regulation of heparin sulfate proteoglycan biosynthesis process and catabolic processes.

Feeding these features to the machine learning model provides highest accuracy of 0.85 with very narrow standard deviation of 0.04. Simultaneously, linear SVM gives this performance with LOOCV accuracy of 0.84. Figure 2 shows that random forest performs equally well like SVM, with the accuracy of 0.82 ± 0.11 (Table 2).

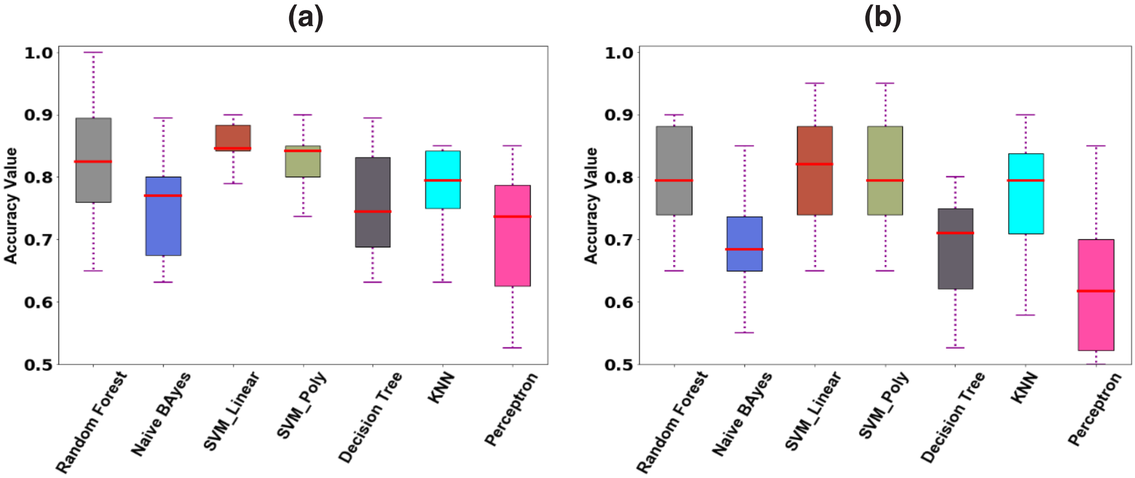


Figure 2. Patients with different respiratory diseases are classified into their corresponding cohort using their transcriptome data. Two sets of features are selected using feature importance and mutual information. Performance of the features selected using, (a) Feature importance and (b) Mutual information on different classifiers.

3. Health status

This study is between the healthy individuals and all other respiratory system diseased patients including SARS-COV-2. Patients with any respiratory disease are grouped together and considered as one group. They are studied against the healthy people to observe the dysregulatioins of genes and classification characteristics.

GO analysis on the 100 features selected by feature importance (Table S-1.31) are related to viral transcription, viral gene expression, viral life cycle and catabolic and metabolic processes (Gardinassi, Souza, Sales-Campos, & Fonseca, 2020; Overmyer, et al., 2021; Loganathan, Ramachandran, Shankaran, Nagarajan, & Mohan, 2020). All functions are significant with very low p-value (8.77x10-5). Enrichment analysis on mutual information selected features on the other hand (Table S-2.21), gives functions such as viral gene expression, viral transcription and other functions such as SRP-dependent cotranslational protein targeting to membrane, catabolic processes and metabolic processes. These enrichment terms and their level of significance show that the selected features are prominent with high information.

Machine learning models on the selected features reveal that all the models perform well with accuracy more than 0.9. This performance is on both the feature sets of this study (Figure 3). Like previous subgroup studies, Linear SVM (0.93 ± 0.06) and random forest (0.95 ± 0.04) perform better here as well compared to other models. Corroboration of this performance using LOOCV gives the accuracy of 0.94 and 0.95 respectively. However, other models also perform equally well (Table 2) on this classification.

4. Study between COVID-19 and healthy people

Next study is on distinction between healthy people and COVID-19 patients. All the other samples are removed in this section. Enrichment terms such as viral transcription, viral gene expression and viral life cycle are identified in both feature sets. Genes associated with various kind of catabolic and life

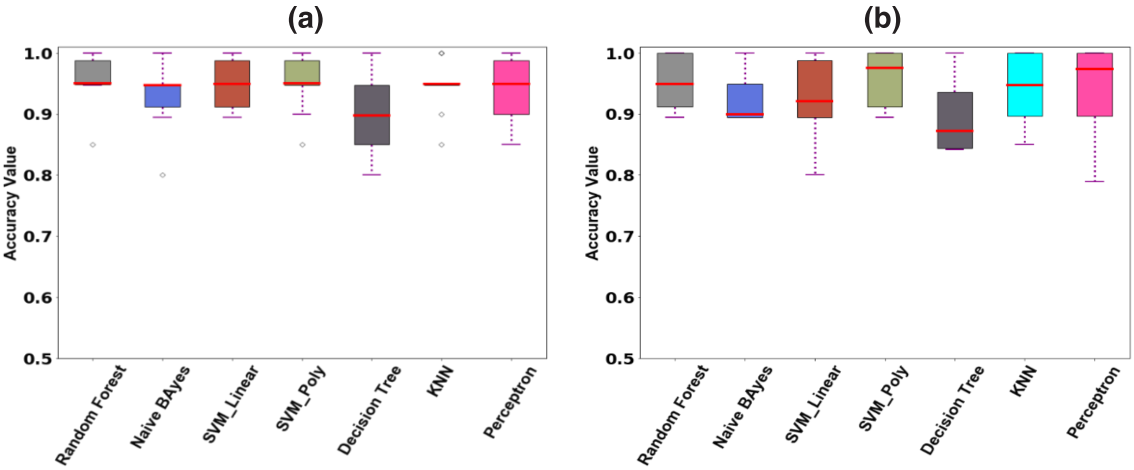


Figure 3. Healthy individuals are classified against the patients with any respiratory diseases. Performance of the 100 features selected using (a) Feature importance and (b) mutual information on different machine learning algorithms

cycle is identified in both feature sets. Genes associated with various kind of catabolic and metabolic processes are also identified (Table S-1.41 and Table S-2.41).

As expected, this classification also yields high performance (> 0.9) with all classifiers except perceptron, which gives 0.88 ± 0.14 on mutual information selected features. Linear SVM and random forest perform comparably well in this classification as well compared to others (Figure 4). Table 2 shows that their performances are high while using LOOCV as well. COVID versus not COVID patients were already studied with 0.91 of AUROC and 85.2% of accuracy (Liu, Fruit, Ward, & Correll,, 1999).

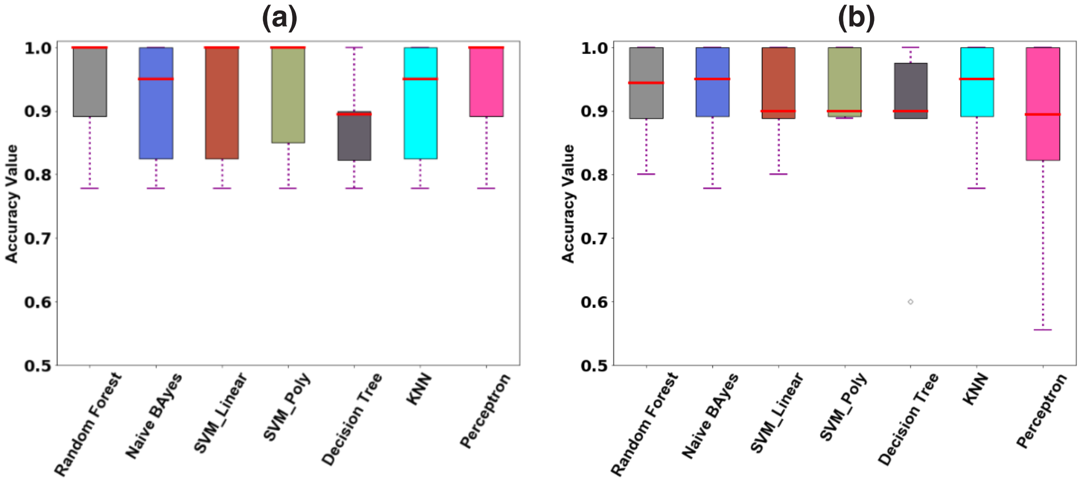


Figure 4. COVID-19 patients are classified against healthy individuals. Other samples are removed from the study. Performance of (a) Feature Importance selected and (b) Mutual Information selected features using different classification methods

5. Stratification between COVID19 patients and other respiratory diseases

COVID-19 is one of the respiratory diseases with comparably high mortality than others. To study the difference between COVID-19 and other respiratory diseases, this study compares patients from both cohorts. Healthy individuals are omitted in this section and COVID-19 patients are grouped against all the patients with any other respiratory diseases. Feature selection on this study gives prominent features with important functions highly related to immune related activities. Feature importance gives features related to immune response, defense response (Jain, et al., 2021) and metabolic processes. Along with immune response genes, genes associated with negative regulation of interleukin-10 production (Gardinassi, Souza, Sales-Campos, & Fonseca, 2020; Sardar, Sharma, & Gupta,, 2021; Patterson, et al., 2021) interferon-gamma biosynthetic process and interferon-gamma production (Gardinassi, Souza, Sales-Campos, & Fonseca, 2020; Overmyer, et al., 2021; Jain, et al., 2021)are selected from mutual information. Interleukin-10 is an important anti-inflammatory cytokine in the body which determines outcomes of many inflammatory diseases (Arimoto, Miyauchi, Stoner, Fan, & Zhang, 2018). Interferon-gamma is also closely related to antimicrobial activities and killing of intracellular pathogens (Liu, Fruit, Ward, & Correll,, 1999; Arslan, 2021).

This study gives highest accuracy of 0.92 ± 0.05 (Figure 5) with polynomial SVM on feature importance selected features. Here LOOCV performance is 0.93 (Table 2). Using CpG island features on this classification resulted with the accuracy of 0.93 using random forest classifier (Bwire, 2020). All the classifiers in this study perform greater than 0.8 (Figure 5) on transcriptome data.

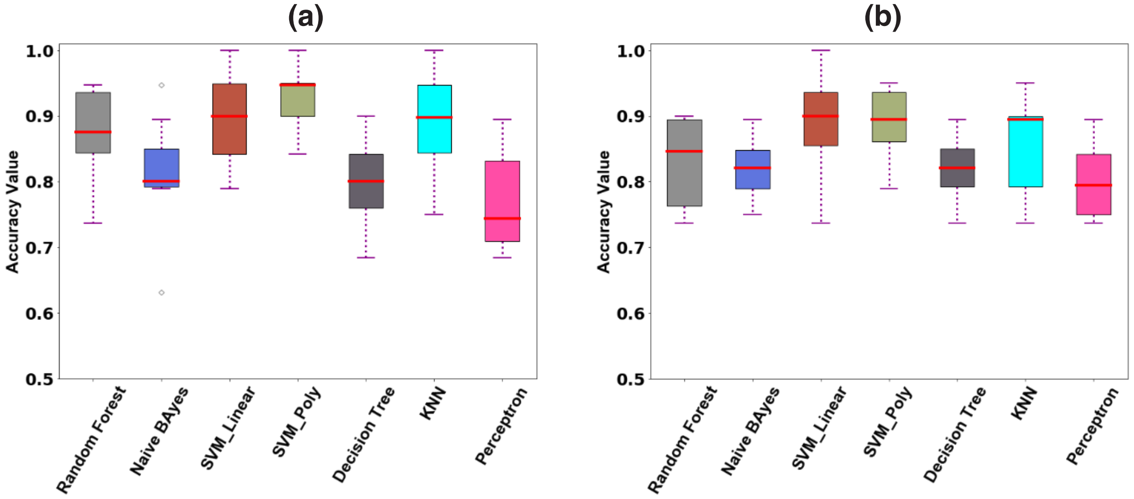


Figure 5. Stratification between COVId-19 and other respiratory diseases. Healthy individuals are not considered in this study. Features are selected using (a) Feature Importance and (b) Mutual Information and their performances are compared on different machine learning classification algorithms.

6. Gender specificity of COVID-19 patients

Studies show that men are more vulnerable to COVID-19 than women (Jin, et al., 2020) and there is a gender difference in COVID-19 patients (Bajaj, et al., 2021). To check the gender specificity of COVID-19 patients, this study considers all the COVID-19 patients to classify them into male or female. Test for functional enrichment of the selected genes surprisingly not shown any gender-related GO terms (Table S-1.61, Table S-2.61). Rather, they are related to gene expression regulation, coagulation and hemostasis (Gardinassi, Souza, Sales-Campos, & Fonseca, 2020; Liu, Jia, Fang, & Zhao, 2020; Sardar, Sharma, & Gupta,, 2021), metabolic processes, platelet activation and dealkylation and demethylation (Liu, Jia, Fang, & Zhao, 2020). Individual checking on the selected features shows that there are no X or Y chromosome related genes except XIST. Using the selected set of features individually for classifying the patients into male or female gives very high accuracy. All the models perform with an accuracy higher than 0.95 (Figure 6), with the highest of 0.97 ± 0.06 (Table 2).

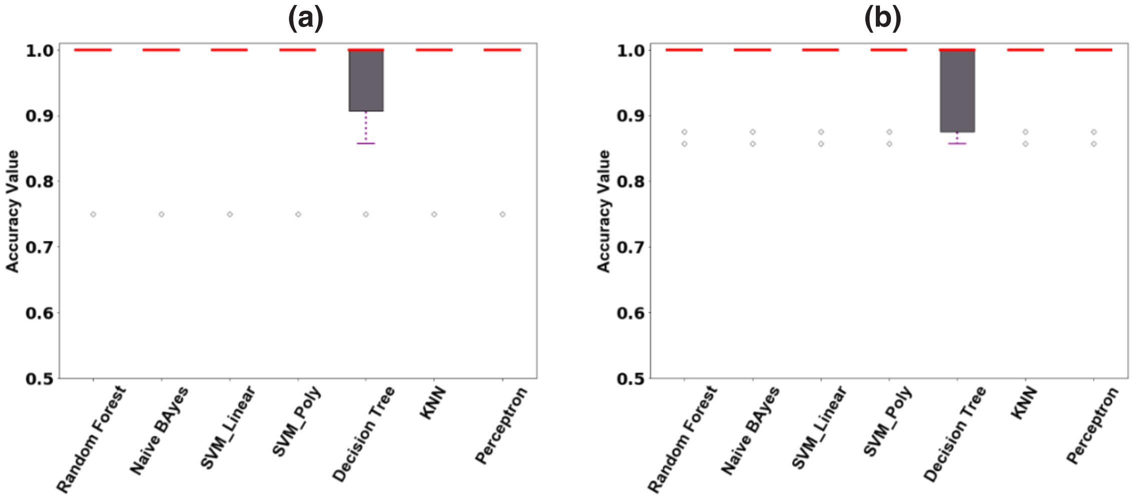


Figure 6. Gender specificity of COVID 19 patients. Only COVID-19 patients are studied here to classify them into male and female. Top 100 features are selected using (a) Feature Importance and (b) Mutual Information, and fed into different classifiers to compare their performances

7. Hospitalization requirement of COVID-19 patients

Several factors determine the hospitalization of a COVID-19 patient including their physical condition. As other complications of the patients are not provided with the dataset, transcriptome data of the patients in this study is used to predict the hospital status of a COVID-19 patient. Functional enrichment analysis on the selected features shows functions closely related to sugar (fructose and glucose) (Overmyer, et al., 2021), catabolic and metabolic activities and terms related to regulation of few activities such as cell cycle. As glucose related genes are highly enriched in the selected features, the hospitalized patients might have other complications such as diabetes. Performance of these features in this classification provides very high accuracy up to 0.99 ± 0.04, (Table 2). Naïve Baye’s algorithms gives highest performance on mutual information selected features, while other classifiers also give comparable performance (Figure 7). SVM and random forest perform comparably well than other classifiers in this study as well.

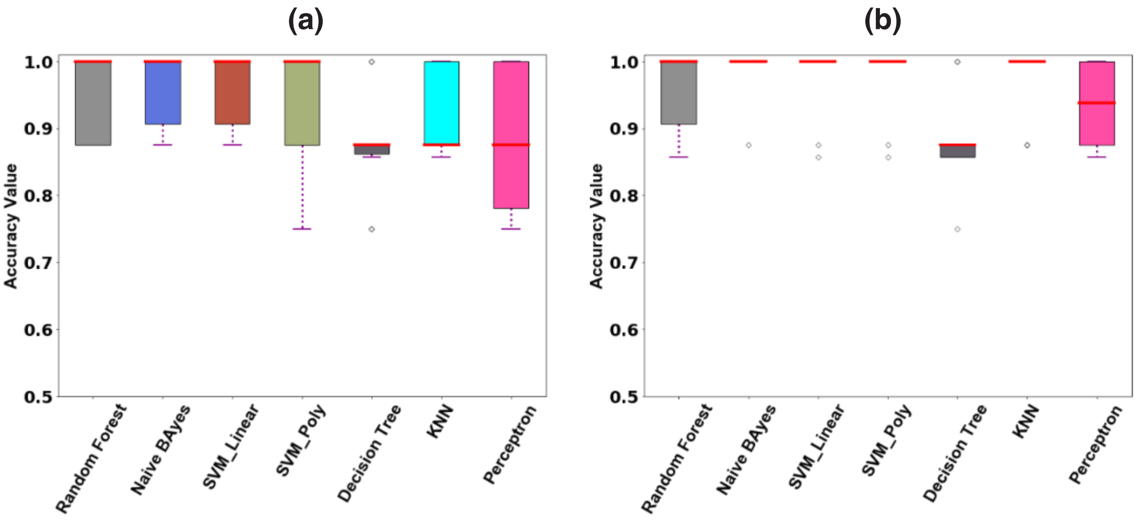


Figure 7. Classification of the COVID-19 patients into their hospital status. Selected transcriptome features are used in this classification using different machine learning algorithms. Performance of the features selected using (a) Feature Importance and (b) Mutual Information

8. Age clusters between COVID-19 patients

Age plays a crucial role in the history of SARS-CoV-2. Here patients with age <=50 are grouped against the patients with age >50. Important enrichment functions of feature importance selected features are mostly related to regulation activities (of T cell differentiation, thymocyte aggregation, leukocyte, and lymphocyte differentiation (Gardinassi, Souza, Sales-Campos, & Fonseca, 2020)). GO terms related to features selected by mutual information show unique functions such as cranial skeletal system development, regulation of cellular response to stress, metabolic processes (Table S-1.81, Table S-2.81).

Using these features in different machine learning models gives maximum accuracy of 0.89 ± 0.08 on feature importance selected features using polynomial SVM (Table 2). However, perceptron seems to be a good classifier to this problem with good mean value and narrow variance (Figure 8).

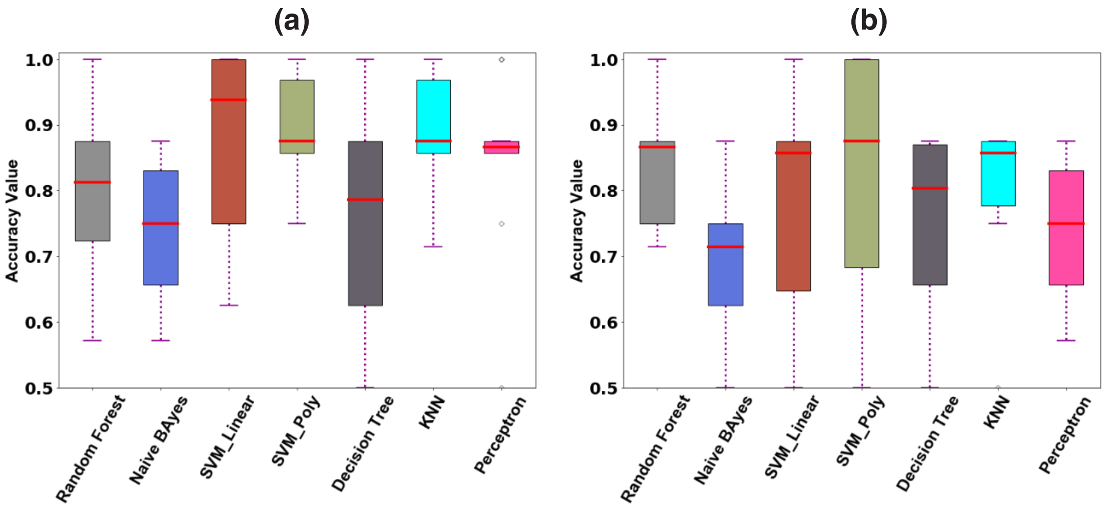


Figure 8. Age classification of COVID-19 patients. Then selected genes are used in the classification using different classifiers. Performance of the classifiers are compared between (a) Feature Importance and (b) Mutual Information selected features

9. COVID-19 status

Final two studies use a different dataset GSE157103, which measured transcriptome data on hospitalized patients who had and did not have COVID-19. Combined analysis are not performed between these two datasets, as the experimental setup, nature of the patients and the samples used are different among them.

This section tries to distinguish COVID-19 patients from non-COVID-19 using their transcriptome data. First step is feature selection. As mentioned before, 100 features are individually selected in the same way and tested for their functional enrichments. This analysis shows that selected genes are highly related to mitotic cell related activities, cell cycle related activities, DNA replication and nuclear division (Gardinassi, Souza, Sales-Campos, & Fonseca, 2020; Overmyer, et al., 2021).

In this classification, random forest performs best among these classifiers (0.95 ± 0.06) (Table 2). Even though all the others perform well in this classification (Figure 9), SVM, which performs best in all the other classifiers performs comparably low here. In the mutual information selected features, decision tree classifier gives the accuracy of 0.92 with very narrow standard deviation (0.05). All these classifiers give almost equivalent performance using LOOCV as well (Table 2).

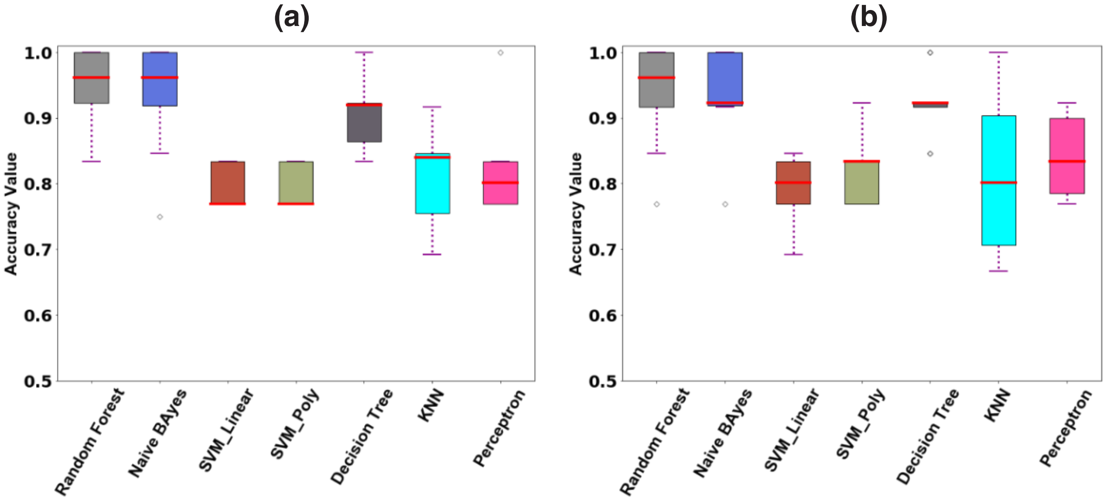


Figure 9. COVID-19 patients were classified against non-COVID-19 patients using the genes selected by, (a) Feature Importance and (b) Mutual Information. Six different classifiers are used in this classification and their performances are compared.

10. ICU status of COVID-19 patients

The same dataset was given with the ICU status of the hospitalized patients. Lastly, this data is used to predict the ICU status of the patients. Selected features are mostly related to immune related activities, virus defend activities, T cell receptor, metabolic process, cell activities and DNA, RNA related activities. These features give an accuracy of 0.86 ± 0.11 on random forest classifier (Table 2). Figure 10 illustrates performances by different classifiers.

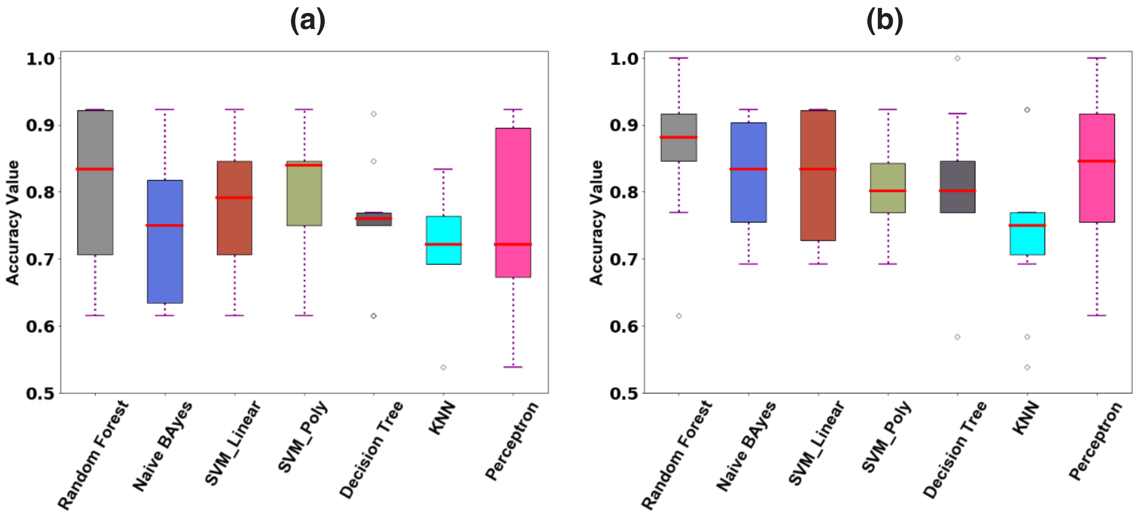


Figure 10. Last classification of the study. ICU status of the hospitalized patients are predicted using (a) Feature Importance and (b) Mutual Information selected features. Those features are applied on six different classifiers and their performance are compared

Discussion

Two different transcriptome datasets are used in this study. Both are provided with different sets of clinical data. Altogether ten different clinical data related to COVID-19 are studied here. Each dataset has around 15 000 features measured on more than hundred patients including healthy individuals. As the number of features are too high compared to number of samples, first step of the study was to select suitable features of the study. Two different feature selection techniques, mutual information and feature importance are used. These methods are widely used in bioinformatics (Jansi Rani & Devaraj, 2019; Ng, et al., 2021) for feature selection. Table 2 shows that both these methods selected features with high importance (Figure S1-1 - Figure S2-10) and provide very high classification performance. Those selected features are individually tested for their functional enrichment. GO analysis on the features shows that they are mostly related to viral transcription, viral gene expression and viral life cycle (Gardinassi, Souza, Sales-Campos, & Fonseca, 2020; Loganathan, Ramachandran, Shankaran, Nagarajan, & Mohan, 2020; Jain, et al., 2021). Immune system related genes (Overmyer, et al., 2021; Liu, Jia, Fang, & Zhao, 2020) have also been preferably chosen in this study to be used in the classification processes. Apart from these functions, cellular processes related genes (Overmyer, et al., 2021; Loganathan, Ramachandran, Shankaran, Nagarajan, & Mohan, 2020; Sardar, Sharma, & Gupta,, 2021), genes related to catabolic and metabolic process (Gardinassi, Souza, Sales-Campos, & Fonseca, 2020; Overmyer, et al., 2021; Loganathan, Ramachandran, Shankaran, Nagarajan, & Mohan, 2020) and coagulation processes (Gardinassi, Souza, Sales-Campos, & Fonseca, 2020; Jain, et al., 2021; Sardar, Sharma, & Gupta,, 2021) are widely identified in almost all the subgroups. Various unique functional enrichment terms are identified in this study and listed from Table S-1.11 to Table S-2.101.

In section 3.h, patients are classified as two groups based on their age. For having enough patients in both groups, 50 is considered as the boundary age between two cohorts. Actual value should be greater than this limit, 50 (Mahase, 2020). So, keep this boundary will not affect the reported performance of this study. If we actually increase this limit to proper value, the classifier would perform even well, as the each group will carry more accurate information.

Finally, the selected features are used to predict the corresponding phenotype of the patients. Table 2 shows that all these phenotypes can be predicted with high accuracy using the transcriptome of COVID-19 patients. Either SVM (linear or polynomial) or random forest can yield the accurate prediction in any of this classification.

Because of the low number of samples in each study, every result is confirmed using LOOCV technique. LOOCV confirms the results obtained using 10-fold cross validation with almost equal accuracy value. Performance of the model shows that there are significant dysregulations in the genes of COVID-19 patients.

Table 2. Summary of the classifications done, and their performances using six different classifiers and two feature selection methods, Mutual Information (MI), Feature Importance (FI). Models are validated using (a) 10-fold cross validation (10-F CV) and (b) LOOCV. Highest performance in each classification problem is highlighted.

(a) Accuracies using 10-fold cross validation

Classification Problem	Feature selection method	Random Forest	Naïve Bayes	SVM - Linear	SVM – Polynomial	Decision tree	KNN	Perceptron
Time since onset	MI	0.75±0.09	0.65±0.09	0.78±0.14	0.79±0.13	0.66±0.12	0.72±0.09	0.64±0.08
	FI	0.75±0.07	0.68±0.12	0.78±0.12	0.76±0.11	0.66±0.08	0.7±0.11	0.72±0.6
Disease cohort	MI	0.79±0.09	0.69±0.08	0.81±0.1	0.81±0.1	0.69±0.09	0.77±0.1	0.58±0.17
	FI	0.82±0.11	0.74±0.12	0.85±0.04	0.83±0.05	0.75±0.09	0.78±0.07	0.71±0.12
Health status	MI	0.95±0.04	0.92±0.04	0.93±0.06	0.96±0.05	0.9±0.06	0.94±0.06	0.94±0.08
	FI	0.95±0.04	0.93±0.05	0.95±0.04	0.95±0.05	0.9±0.07	0.94±0.04	0.94±0.05
COVID-19 VS healthy	MI	0.92±0.08	0.93±0.08	0.92±0.08	0.94±0.05	0.9±0.12	0.92±0.09	088±0.14
	FI	094±0.09	0.92±0.09	0.93±0.1	0.94±0.1	0.89±0.08	0.92±0.09	0.94±0.09
COVID-19 VS other respiratory	MI	0.83±0.07	0.82±0.05	0.88±0.09	0.89±0.06	0.81±0.06	0.86±0.07	0.8±0.05
	FI	0.87±0.08	0.82±0.08	0.89±0.07	0.92±0.05	0.81±0.07	0.89±0.08	0.77±0.07
Gender specificity	MI	0.97±0.06	0.97±0.06	0.97±0.06	0.97±0.06	0.95±0.07	0.97±0.06	0.97±0.08
	FI	0.97±0.06	0.97±0.08	0.97±0.08	0.97±0.08	0.95±0.09	0.97±0.08	0.97±0.08
Hospitalization	MI	0.96±0.06	0.99±0.04	0.97±0.06	0.97±0.06	0.88±0.07	0.97±0.05	0.93±0.07
	FI	0.95±0.06	0.96±0.06	0.96±0.06	0.94±0.09	0.87±0.08	0.92±0.07	0.89±0.11
Age clusters	MI	0.84±0.1	0.81±0.12	0.77±0.17	0.83±0.19	0.75±0.14	0.81±0.12	0.72±0.15
	FI	0.8±0.15	0.74±0.12	0.88±0.15	0.89±0.08	0.76±0.18	0.88±0.1	0.84±0.14
COVID-19 status	MI	0.94±0.08	0.94±0.07	0.79±0.05	0.82±0.05	0.92±0.05	0.82±0.12	0.84±0.06
	FI	0.95±0.06	0.94±0.08	0.79±0.03	0.79±0.03	0.91±0.05	0.81±0.07	0.82±0.07
ICU status	MI	0.86±0.11	0.82±0.08	0.83±0.1	0.81±0.07	0.81±0.11	0.74±0.12	0.83±0.12
	FI	0.81±0.12	0.75±0.12	0.79±0.1	0.8±0.1	0.76±0.09	0.7±0.12	0.75±0.12

Conclusion

This study uses transcriptome data of COVID-19 patients to classify them into their corresponding clinical data. Ten different phenotypes are predicted in this study, using seven different machine learning methods. From tens of thousands of measured transcriptomes, hundred are selected using feature importance and mutual information. Each set of selected features is used separately in the prediction of corresponding clinical data. Results show that both feature selection methods perform equally well in these classifications. Even though random forest and SVM (either linear or polynomial) give the best performance throughout all the phenotype classifications, while other classifiers also provide very high accuracy.

(b) Accuracies using Leave One Out Cross Validation

Classification Problem	Feature selection method	Random Forest	Naïve Bayes	SVM - Linear	SVM – Polynomial	Decision tree	KNN	Perceptron
Time since onset	MI	0.74	0.65	0.76	0.77	0.64	0.73	0.65
	FI	0.73	0.68	0.77	0.77	0.64	0.7	0.68
Disease cohort	MI	0.81	0.69	0.84	0.82	0.71	0.76	0.59
	FI	0.81	0.73	0.84	0.87	0.75	0.77	0.72
Health status	MI	0.95	0.92	0.94	0.96	0.9	0.95	0.88
	FI	0.95	0.93	0.94	0.95	0.91	0.94	0.93
COVID-19 VS healthy	MI	0.93	0.94	0.91	0.93	0.86	0.93	0.9
	FI	0.94	0.92	0.93	0.93	0.89	0.91	0.93
COVID-19 VS other respiratory	MI	0.84	0.82	0.91	0.88	0.77	0.86	0.7
	FI	0.86	0.82	0.92	0.93	0.81	0.88	0.82
Gender specificity	MI	0.97	0.97	0.97	0.97	0.92	0.97	0.97
	FI	0.97	0.97	0.97	0.97	0.95	0.97	0.95
Hospitalization of COVID-19 patients	MI	0.96	0.97	0.97	0.97	0.78	0.97	0.94
	FI	0.96	0.96	0.95	0.95	0.78	0.92	0.91
Age clusters	MI	0.77	0.82	0.75	0.84	0.79	0.82	0.68
	FI	0.82	0.75	0.9	0.86	0.87	0.88	0.82
COVID-19 status	MI	0.94	0.94	0.8	0.8	0.9	0.82	0.79
	FI	0.95	0.94	0.79	0.79	0.91	0.81	0.76
ICU status	MI	0.84	0.83	0.83	0.81	0.76	0.75	0.74
	FI	0.83	0.74	0.79	0.8	0.7	0.68	0.74

Data accessibility

Datasets used in this study are publicly available in GEO data repository under the accession numbers of GSE157103 and GSE161731. All the models use pre-built functions in python libraries, none of them are implemented here from the scratch.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Arimoto, K., Miyauchi, S., Stoner, S., Fan, J., & Zhang, D. (2018). Negative regulation of type I IFN signaling. Journal of Leukocyte Biology, 10(6), 1099-1116. doi:10.1002/JLB.2MIR0817-342R

Arslan, H. (2021). Machine Learning Methods for COVID-19 Prediction Using Human Genomic Data. Proceedings, 74(1). doi:10.3390/proceedings2021074020

Bajaj, V., Gadi, N., Spihlman, A., Wu, S., Choi, C., & Moulton, V. (2021). Aging, Immunity, and COVID-19: How Age Influences the Host Immune Response to Coronavirus Infections?

Frontiers in Physiology, 11. doi:10.3389/fphys.2020.571416

Bwire, G. (2020). Coronavirus: Why Men are More Vulnerable to Covid-19 Than Women? SN Comprehensive Clinical Medicine, 2(7), 874-876. doi:10.1007/s42399-020-00341-w

Castro de Moura, M., Davalos, V., Planas-Serra, L., Alvarez-Errico, D., Arribas, C., & Ruiz, M. (2021). Epigenome-wide association study of COVID-19 severity with respiratory failure. EBioMedicine, 66. doi:10.1016/j.ebiom.2021.103339

Christie, A., Henley, S., Mattocks, L., Fernando, R., Lansky, A., & Ahmad, F. (2021). Decreases in COVID-19 Cases, Emergency Department Visits, Hospital Admissions, and Deaths Among Older Adults Following the Introduction of COVID-19 Vaccine — United States, September 6, 2020-May 1, 2021. MMWR Morbidity and Mortality Weekly Report, 70(23), 858-864. doi:DOI: 10.15585/mmwr.mm7023e2

Fu, J., Zhou, B., Zhang, L., Balaji, K., Wei, C., & Liu, X. (2020). Expressions and significances of the angiotensin-converting enzyme 2 gene, the receptor of SARS-CoV-2 for COVID-19. Molecular Biology Reports, 47(6), 4383-4392. doi:10.1007/s11033-020-05478-4

Fullard, J., Lee, H., Voloudakis, G., Suo, S., Javidfar, B., & Shao, Z. (2021). Single-nucleus transcriptome analysis of human brain immune response in patients with severe COVID-19. Genome Medicine, 13(1). doi:10.1186/s13073-021-00933-8

Gardinassi, L., Souza, C., S.-C. H., & Fonseca, S. (2020). Immune and Metabolic Signatures of COVID-19 Revealed by Transcriptomics Data Reuse. Frontiers in Immunology, 11. doi:10.3389/fimmu.2020.01636

Islam, A., & Khan, M. (2020). Lung transcriptome of a COVID-19 patient and systems biology predictions suggest impaired surfactant production which may be druggable by surfactant therapy. Scientific Reports, 10(1). doi:10.1038/s41598-020-76404-8

Jain, R., Ramaswamy, S., Harilal, D., Uddin, M., Loney, T., & Nowotny, N. (2021). Host transcriptomic profiling of COVID-19 patients with mild, moderate, and severe clinical outcomes. Computational and Structural Biotechnology Journal, 19, 153-160. doi:10.1016/j.csbj.2020.12.016

Jansi Rani, M., & Devaraj, D. (2019). Two-Stage Hybrid Gene Selection Using Mutual Information and Genetic Algorithm for Cancer Data Classification. Journal of Medical Systems, 43(8). doi:10.1007/s10916-019-1372-8

Jin, J., Bai, P., He, W., Wu, F., Liu, X., & Han, D. (2020). Gender Differences in Patients With COVID-19: Focus on Severity and Mortality. Frontiers in Public Health, 8. doi:10.3389/fpubh.2020.00152

Li, H., Liu, S., Yu, X., Tang, S., & Tang, C. (2020). Coronavirus disease 2019 (COVID-19): current status and future perspectives. International Journal of Antimicrobial Agents., 55(5).

Liu, Q., Fruit, K., Ward, J., & C. P. (1999). Negative regulation of macrophage activation in response to IFN-gamma and lipopolysaccharide by the STK/RON receptor tyrosine kinase. J Immunol, 16(12), 6606-6613.

Liu, T., Jia, P., Fang, B., & Zhao, Z. (2020). Differential Expression of Viral Transcripts from Single-Cell RNA Sequencing of Moderate and Severe COVID-19 Patients and Its Implications for Case Severity. Frontiers in Microbiology, 11. doi:10.3389/fmicb.2020.603509

Loganathan, T., Ramachandran, S., Shankaran, P., Nagarajan, D., & Mohan, S. S. (2020). Host transcriptome-guided drug repurposing for COVID-19 treatment: a meta-analysis-based approach. PeerJ, 8. doi:10.7717/peerj.9357

Mahase, E. (2020). Covid-19: Why are age and obesity risk factors for serious disease? BMJ. doi:10.1136/bmj.m4130

Moni, M., Lin, P., Quinn, J., & Eapen, V. (2021). COVID-19 patient transcriptomic and genomic profiling reveals comorbidity interactions with psychiatric disorders. Translational Psychiatry, 11(1). doi:org/10.1038/s41398-020-01151-3

Nagpal, A., & Singh, V. (2018). A Feature Selection Algorithm Based on Qualitative Mutual Information for Cancer Microarray Data. Procedia Computer Science, 132, 244-252. doi:10.1016/j.procs.2018.05.195

Ng, D., G. A., Santos, Y., Servellita, V., Goldgof, G., & Meydan, C. (2021). A diagnostic host response biosignature for COVID-19 from RNA profiling of nasal swabs and blood. Science Advances, 7(6). doi:10.1126/sciadv. abe5984

Overmyer, K., Shishkova, E., Miller, I., Balnis, J., Bernstein, M., & Peters-Clarke, T. (2021). Large-Scale Multi-omic Analysis of COVID-19 Severity. Cell Systems, 12(1), 23-40. doi:10.1016/j.cels.2020.10.003

Padhan, R., & Prabheesh, K. (2021). The economics of COVID-19 pandemic: A survey. Economic Analysis and Policy, 70, 220-237. doi:10.1016/j.eap.2021.02.012

Patterson, B., Guevara-Coto, J., Yogendra, R., Francisco, E., Long, E., & Pise, A. (2021). Immune-Based Prediction of COVID-19 Severity and Chronicity Decoded Using Machine Learning. Frontiers in Immunology, 12. doi:10.3389/fimmu.2021.700782

Sardar, R., Sharma, A., & G. D. (2021). Machine Learning Assisted Prediction of Prognostic Biomarkers Associated With COVID-19, Using Clinical and Proteomics Data. Frontiers in Genetics, 12. doi:10.3389/fgene.2021.636441

Sathian, B., Asim, M., B. I., Pizarro, A., Roy, B., & Van Teijlingen, E. (2020). Impact of COVID-19 on clinical trials and clinical research: A systematic review. Nepal Journal of Epidemiology, 10(3), 878-887. doi:10.3126/nje.v10i3.31622

Shen, B., Yi, X., Sun, Y., Bi, X., Du, J., & Zhang, C. (2020). Proteomic and Metabolomic Characterization of COVID-19 Patient Sera. SSRN Electronic Journal. doi:10.1016/j.cell.2020.05.032

Shirvaliloo, M. (2021). Epigenomics in COVID-19; the link between DNA methylation, histone modifications and SARS-CoV-2 infection. Epigenomics, 13(10), 745-750. doi:10.2217/epi-2021-0057

Tisdell, C. (2020). Economic, social and political issues raised by the COVID-19 pandemic. Economic Analysis and Policy, 68, 17-28. doi:DOI: 10.1016/j.eap.2020.08.002

Völlmy, F., van den Toorn, H., Zenezini Chiozzi, R., Zucchetti, O., Papi, A., & Volta, C.

(2021). A serum proteome signature to predict mortality in severe COVID-19 patients. Life Science Alliance, 4(9). doi:10.26508/lsa.202101099

Annexes

1. Data visualization:

Two different transcriptome datasets are used in this study. Visualizing the data shows that GSE 157103 ranges over a wide range, hence not normalized and GSE161731 is normalized, and it can be fed to the machine learning model without normalization.

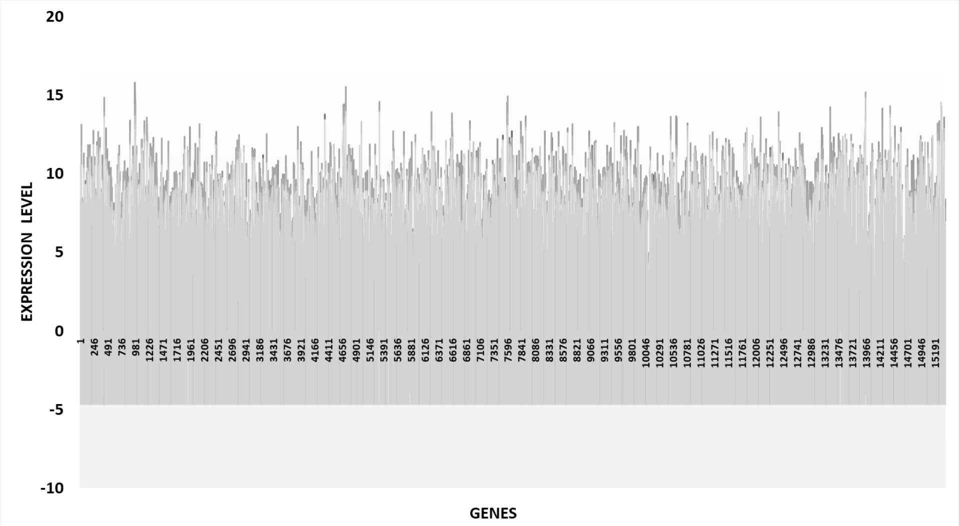


Figure S1. Distribution of expression levels of transcriptomes. Transcriptome data is presented in GSE161731 for COVID-19 patients along with control samples. It is a normalized data between -5 and 15.

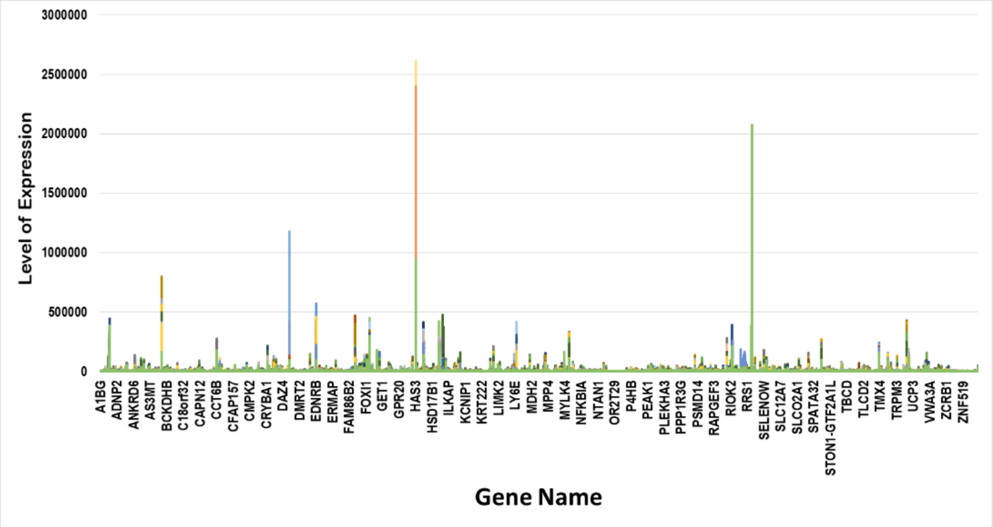


Figure S2. Distribution of gene expression levels reported in GSE157103. GSE157103 is another dataset consists of transcriptome of COVID-19 and controls. This is not a normalized data.

2. Feature Selection:

As the transcriptomes are measured on around 15,000 genes, mutual information and feature importance are used to select the features with high importance. For each phenotype study, top 100 features are selected using these methods. They are individually studies to see their biological functions using GO analysis and then used in the classification models.

I. Features selected using Feature Importance:

i. Phenotype: Time since onset (Early, middle, late)

Patients are provided with three different time scales from their COVID infection: early, middle and late. Selecting features on this classification give the features presented in Table S-1.1.

Table S-1.1. Feature importance selected features for the phenotype analysis of time onset. COVID-19 patients are classified into their level of infection. For this classification, 100 features are selected using feature importance.

TCF7L2	TDRD7	MZB1	LYPD3	HORMAD1
SCARB2	KIAA1109	RP11-597D13.9	ZNF263	FCGR1B
DISC1	HIVEP1	CNTRL	MEA1	IREB2
RP11-383F6.1	ANAPC13	TET2	SORT1	SON
FNDC3B	IGHV2-70	DDX60L	MBOAT1	SRM
C3orf18	ADAR	KIDINS220	MXI1	MANBA
IGLV3-25	PLSCR1	MXD4	LARP4B	ASXL2 WNK1
C18orf25	GPATCH4	NEAT1	POU2AF1	AURKB
NCOA2	GNA12	PDLIM5	GUK1	SOD1
INO80D	MR1	GALC	FOXP3	FBXW4
EIF2AK2	SERPING1	NFKBIZ	KBTBD2	TMEM9
KDM1B	DIP2B	SCAF11	ETV7	SLA2
REL	VPS13B	PAICS	FAM46A	RP11-169D4.2
EIF4G3	SP110	RP11-680H20.1	LPCAT2	LRRK2
TMEM144	CYBB	FAM8A1	AC079630.4	TMTC2
PDCD6IP	KMT2C	TJP2	TRIM65	RP11-68I3.11
IGHG1	TNRC6B	ENSG00000282939	LBH	SAMD9L
RNF213	CD36	IGHV1-24	PPIH CD274	UBE4A
CLPTM1L	MLKL	ABT1	SKAP1	
CACNA1I	RCOR1	TLR2		

Feature importance value of selected features are plotted against the selected features. Bar plot is illustrated in Figure S1-1.

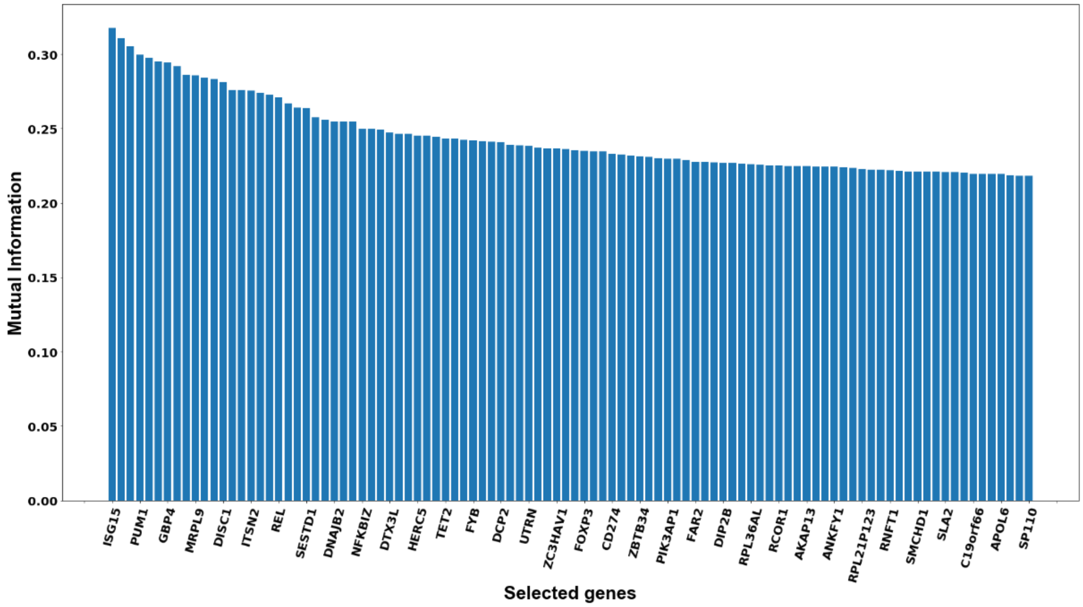


Figure S1-1. For the classification of patients into their 'time since onset' stage, features are selected using feature importance. Selected features are plotted against the corresponding feature importance.

Those selected features are tested for their functional enrichments and the result is summarized in:

Table S-1.11. Functional enrichment of the genes selected using feature importance in the classification of patients into their disease stage.

ID	Name	P Value
GO:0038124	toll-like receptor TLR6:TLR2 signaling pathway	0.4833621434638550
GO:0071724	response to diacyl bacterial lipopeptide	0.4833621434638550
GO:0071726	cellular response to diacyl bacterial lipopeptide	0.4833621434638550
GO:0061484	hematopoietic stem cell homeostasis	0.8029174680793630
GO:0070339	response to bacterial lipopeptide	1.0
GO:0071220	cellular response to bacterial lipoprotein	1.0
GO:0071221	cellular response to bacterial lipopeptide	1.0
GO:0002367	cytokine production involved in immune response	1.0
GO:0032493	response to bacterial lipoprotein	1.0
GO:0045069	regulation of viral genome replication	1.0
GO:0010468	regulation of gene expression	1.0
GO:0038180	nerve growth factor signaling pathway	1.0
GO:0070391	response to lipoteichoic acid	1.0
GO:0071223	cellular response to lipoteichoic acid	1.0
GO:0046007	negative regulation of activated T cell proliferation	1.0
GO:1900225	regulation of NLRP3 inflammasome complex assembly	1.0
GO:0030193	regulation of blood coagulation	1.0
GO:1900046	regulation of hemostasis	1.0
GO:0044546	NLRP3 inflammasome complex assembly	1.0
GO:0002440	production of molecular mediator of immune response	1.0
GO:0050818	regulation of coagulation	1.0

GO:0002474	antigen processing and presentation of peptide antigen via MHC class I	1.0
GO:0050665	hydrogen peroxide biosynthetic process	1.0
GO:1901068	guanosine-containing compound metabolic process	1.0
GO:0032613	interleukin-10 production	1.0
GO:0004385	guanylate kinase activity	1.0
GO:0019079	viral genome replication	1.0
GO:0007596	blood coagulation	1.0
GO:1903555	regulation of tumor necrosis factor superfamily cytokine production	1.0

Steps described above are followed in all the other nine phenotype classifications. They are presented in the following sections.

ii. Phenotype: Cohort (COVID-19, Bacterial, Influenza, Seasonal Covid and healthy)

Table S-1.2. List of selected features in the classification of patients into their disease cohort. Patients are classified into their disease cohort using transcriptome data. Feature importance selected features in this classification

DIP2B	SP110	EEF1B2	SASH1	CD302
KIAA1109	ENSG00000282939	SLK	ENTPD7	ITPRIPL2
DDX60L	IFI6	CLK4	DAPP1	MCEMP1
SPG11	SLA2	IFI27	GYG1	MLKL
CYBB	TMOD3	WBSCR22	ZNF445	GCA
FBXW2	THAP7	STRN	COA1	RPL36
DTX3L	VCAN	E2F3	SOS2	ACOT9
SCARB2	PABPC1	ICAM2	CPEB4	NXPE3
CTD-2047H16.3	CASC4	DMXL2	CCL5	ATP13A1
ADRBK2	RAB21	HERC5	SAP30L	CTA-384D8.35
KIDINS220	AKAP10	PTPN7	IFIT1	STXBP5
SORT1	RPS27A	KLHL2	EPHB2	BANP
ROCK1	UBL3 MPI	UBASH3A	AGO4	RPS17
CCDC58	KMT2C	GSR	CD55	AGTPBP1
IFT27	TET2	SMCHD1	KDM1B	SULT1B1
ST6GALNAC6	ST3GAL4-AS1	RPL6P27	DPY19L3	KCNE1
MPEG1	WDFY3	ZKSCAN1	CDK4	BNIP2
TRBC2	CD274	SLC25A40	VPS51	SLC1A3
IRF9	GUCD1	SAMSN1	NF1	EDEM3
KLF7		PLXNC1	AMBRA1	MAN1A1

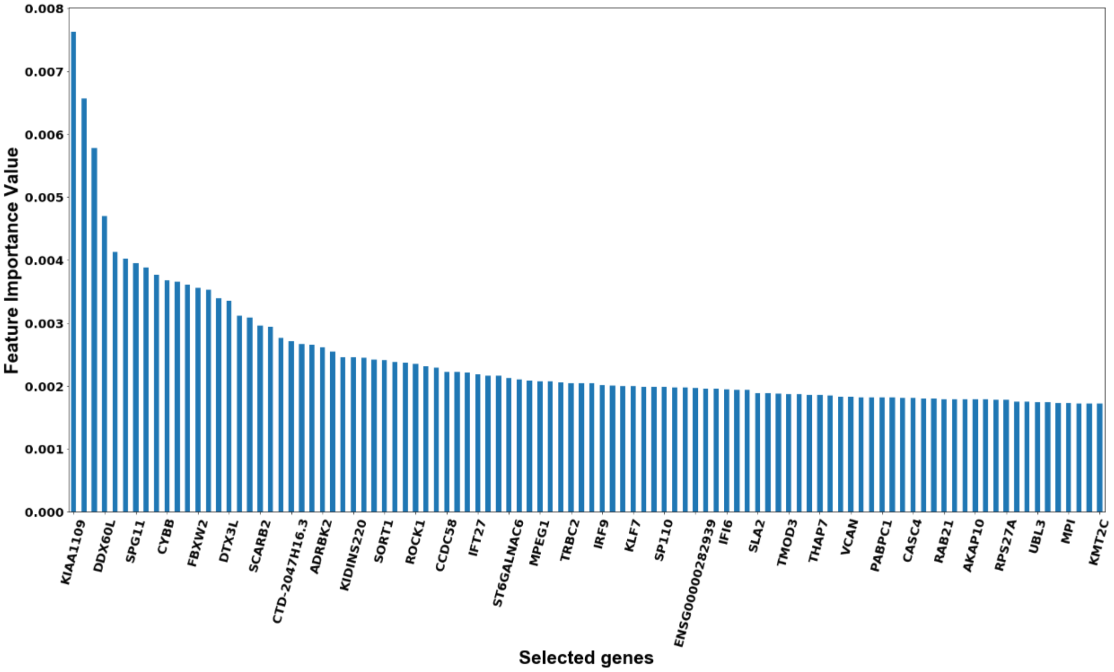


Figure S1-2. Feature importance is plotted against the features selected using feature importance in the classification of the patients into their disease cohort.

Table S-1.21. GO terms related to the selected features in the classification of patients into their disease cohort.

ID	Name	P Value
GO:1904381	Golgi apparatus mannose trimming	1.0
GO:0004571	mannosyl-oligosaccharide 1,2-alpha-mannosidase activity	1.0
GO:0015924	mannosyl-oligosaccharide mannosidase activity	1.0
GO:0036507	protein demannosylation	1.0
GO:0036508	protein alpha-1,2-demannosylation	1.0
GO:0004559	alpha-mannosidase activity	1.0
GO:0015923	mannosidase activity	1.0
GO:0006491	N-glycan processing	1.0
GO:0006517	protein deglycosylation	1.0
GO:0009311	oligosaccharide metabolic process	1.0
GO:0006487	protein N-linked glycosylation	1.0
GO:0004553	hydrolase activity, hydrolyzing O-glycosyl compounds	1.0
GO:0005793	endoplasmic reticulum-Golgi intermediate compartment	1.0
GO:0016798	hydrolase activity, acting on glycosyl bonds	1.0
GO:0006486	protein glycosylation	1.0
GO:0043413	macromolecule glycosylation	1.0
GO:0070085	glycosylation	1.0
GO:0009101	glycoprotein biosynthetic process	1.0
GO:0009100	glycoprotein metabolic process	1.0
GO:0044723	single-organism carbohydrate metabolic process	1.0
GO:0000139	Golgi membrane	1.0
GO:0005509	calcium ion binding	1.0

GO:1901137	carbohydrate derivative biosynthetic process	1.0
GO:0005975	carbohydrate metabolic process	1.0
GO:0044431	Golgi apparatus part	1.0
GO:1901135	carbohydrate derivative metabolic process	1.0
GO:0005794	Golgi apparatus	1.0
GO:0005783	endoplasmic reticulum	1.0
GO:0098588	bounding membrane of organelle	1.0

iii. Phenotype: Healthy individuals VS all other patients

Table S-1.3. Hundred features selected in the stratification of patients into their healthy status. Here healthy people are classified against patients having any respiratory disease. Features are selected using feature importance.

PFDN5	DNASE1L3	ENY2	LARP7	CRTAP
RP11-291B21.2	DCAF13	TMA7	EFHD2	TBCA
RP11-466H18.1	RP11-244J10.1	RP4-800G7.1	SCARNA21	HAT1
SNHG8	IFT20	HSF2	RP11-51O6.1	PLEKHO2
PMPCB	CETN3	GRN	HSPB11	C1orf162
RPL34	EEF1B2	UQCRC1	KIAA0930	CLNS1A
SNHG6	MRFAP1L1	SH3BP5L	RRN3P1	PAIP1
POLR2K	MRPL22	TPT1	RPL21P28	RPS8
FAM204A	RPS3AP47	CCDC53	PPIA	GALNT10
RPL35A	RPS27	RPL7	KLRB1	ARHGAP30
TOMM20	RWDD1	C8orf59	RPS4X	SNX19
MECP2	DPH5	MAP4K3	DBNL	SHFM1
UQCRB	ZNF83	KLRC1	LAMTOR1	CTIF
TAF1D	GIMAP7	RP11-175B9.3	UBE2K	LTK
NKIRAS2	TRIB3	MRPL21	LY75	E2F2
RAB1B	RPL30	RPS27A	MTERF2	KCTD21
COX7C	RPL11	RPL9	ATP5F1	THOC1
CHMP4B	EEF1A1	GATAD2A	APOBEC3B	EMR1
RPL5	COX7B	RP11-92K2.2	RPS7P1	NIT2
TRGV9	RPS3AP6	NR1D2	ENTPD5	TRAM1

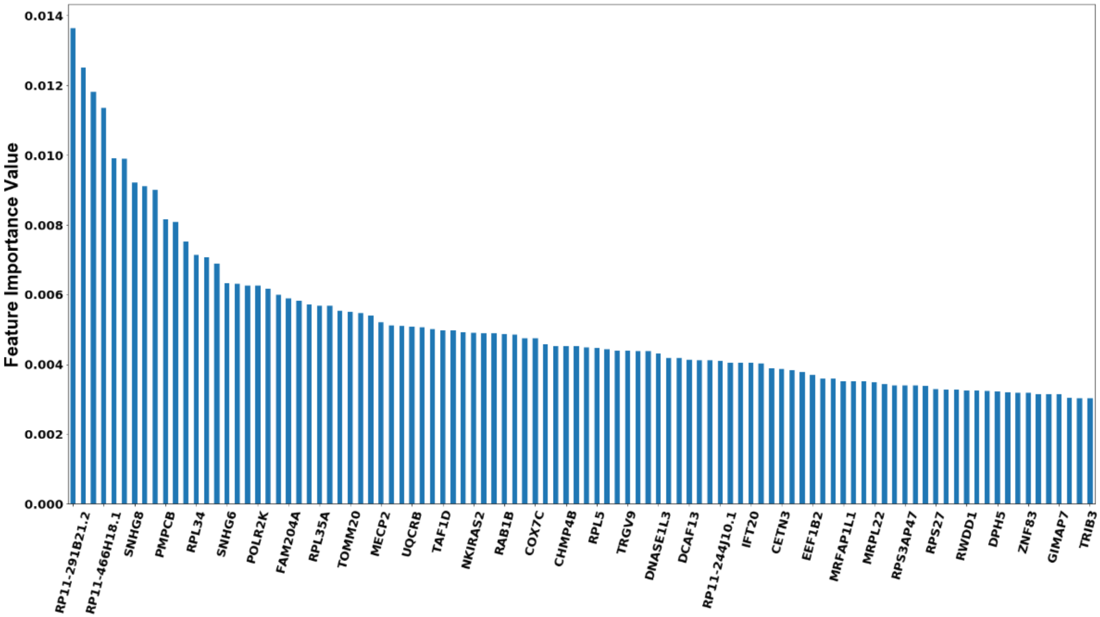


Figure S1-3. Feature importance selected features along with their feature importance value. These features are selected in the classification of the patients into their healthy status (are they healthy or having any respiratory disease)

Table S-1.31. Functional enrichment terms related to the selected features in the classification of the patients into their healthy status.

ID	Name	P Value
GO:0045047	protein targeting to ER	3.247722888800848E-10
GO:0006613	cotranslational protein targeting to membrane	4.141639530870768E-10
GO:0072599	establishment of protein localization to endoplasmic reticulum	5.254335296600144E-10
GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	3.309177474367533E-9
GO:0070972	protein localization to endoplasmic reticulum	4.445564238228691E-9
GO:0006614	SRP-dependent cotranslational protein targeting to membrane	4.690264427913616E-9
GO:0044391	ribosomal subunit	5.693976146308744E-9
GO:0022626	cytosolic ribosome	3.301177863935551E-8
GO:0006612	protein targeting to membrane	3.539886388981291E-8
GO:0006413	translational initiation	4.939551671891652E-8
GO:0000956	nuclear-transcribed mRNA catabolic process	7.27694205110292E-8
GO:0006402	mRNA catabolic process	1.712547050729687E-7
GO:0003735	structural constituent of ribosome	2.71685412866491E-7
GO:0019083	viral transcription	2.811613727107954E-7
GO:0019080	viral gene expression	5.68260633541329E-7
GO:0005840	ribosome	5.836119067477666E-7
GO:0006401	RNA catabolic process	7.975315636380336E-7
GO:0044033	multi-organism metabolic process	1.548222721589113E-6
GO:0019058	viral life cycle	2.000166306005748E-6
GO:0015934	large ribosomal subunit	3.53282526751206E-6
GO:0034655	nucleobase-containing compound catabolic process	1.186775690116567E-5

GO:0006364	rRNA processing	2.02465402854651E-5
GO:0016072	rRNA metabolic process	2.621843540223931E-5
GO:0046700	heterocycle catabolic process	3.165373365122187E-5
GO:0044270	cellular nitrogen compound catabolic process	3.837234414165562E-5
GO:0019439	aromatic compound catabolic process	4.636167944450116E-5
GO:0044445	cytosolic part	5.175334098833948E-5
GO:1901361	organic cyclic compound catabolic process	8.766545644947198E-5

iv. Phenotype: Healthy individuals VS COVID-19 patients

Table S-1.4. Top hundred features are selected while classifying the COVID-19 patients against healthy individuals. Patients with other diseases are not considered in this section

RPL34	TMA7	CDC37L1	CD69	CNTLN
NFYB	DAXX	BLOC1S2	CD48	RN7SK
TUBB4B	KIF1C	GIMAP7	RPS18	KBTBD3
RPS24	RPS3A	AP000936.1	DCTPP1	UBE2Q2P6
DCAF11	CCDC88A	REV1	SLC7A5	OPRL1
GPATCH11	C10orf88	UBL4A	RPL30	ARFRP1
EEF1A1	LONRF1	HSDL1	RP11-543P15.1	AB019441.29
OTUD6B-AS1	ZFC3H1	MCM6	ADRM1	RPL7P9
RPS27	RPS27A	BAZ2B	RNF185	RP11-705C15.2
DDX5	RPL7	CCNK	RPL5	IREB2
C9orf69	LINC01506	RBM39	LRRC28	ZNF613
DRAM2	BTF3L4	DEF6	RNF5	TMC8
RPL39	RPL21P28	TMEM184B	GPSM3	RP11-613F7.1
TRAPPC1	PRPF4B	PRPF40A	NAP1L1	KIAA1109
RPS15A	SRM	POLDIP3	RPL4P5	NELFE
RPS3AP47	ASNSD1	RCN2	VCPKMT	RPS2
ACTR1A	HMGN2	ATG101	UQCRB	PLEKHM1
RPL26	DBI	BAP1	PAIP1	TUBA1C
CRTAM	SCAF11	AP1S2	LPAR6	COX7C
PIGT	UBE2V2	SH3BGRL3	EIF2A	RUFY3

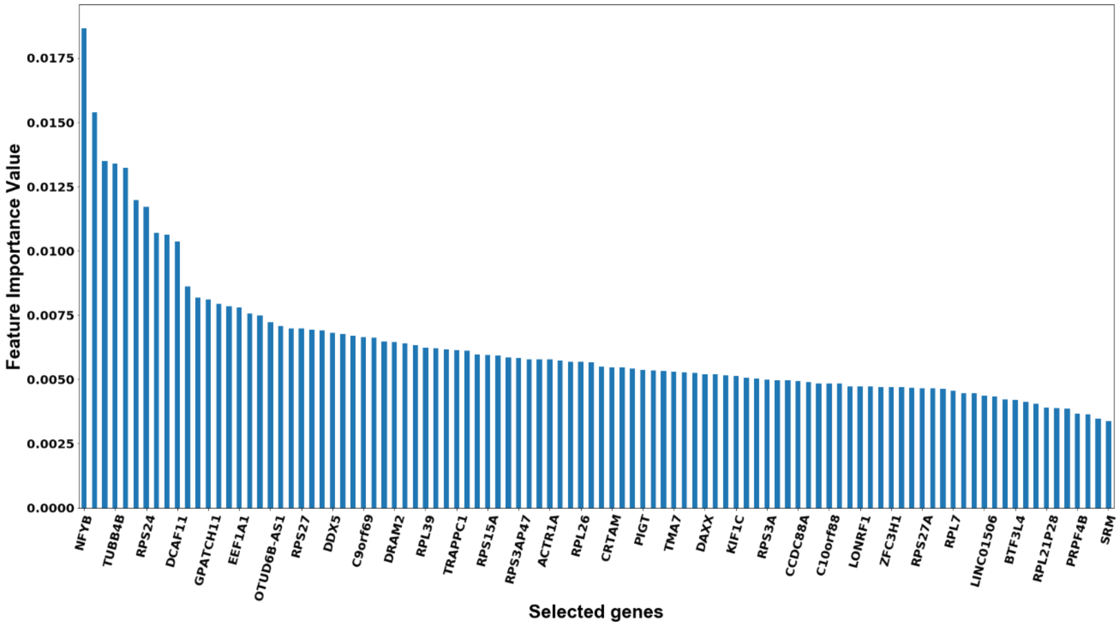


Figure S1-4. Bar plot between selected features and their corresponding feature importance value. This is on the classification of patients into, whether they are COVID-19 or healthy

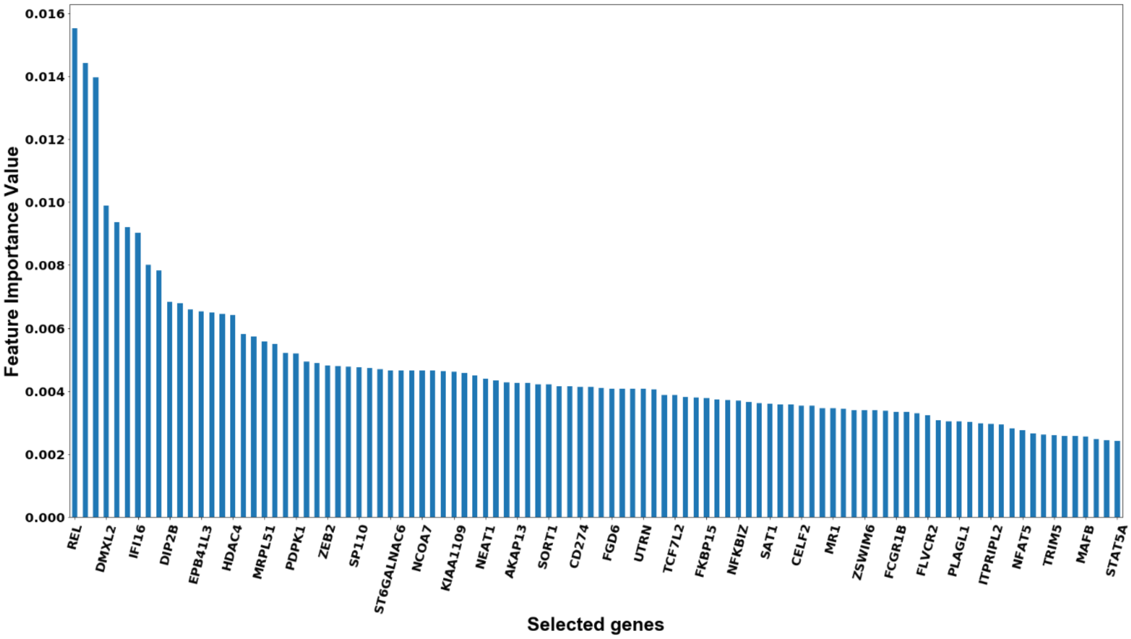
Table S-1.41. Selected genes are tested for gene enrichment analysis. List of GO terms on the features selected while classifying them into healthy or COVID-19

ID	Name	P Value
GO:0006614	SRP-dependent cotranslational protein targeting to membrane	4.064907915386337E-12
GO:0045047	protein targeting to ER	9.354948144612602E-12
GO:0006613	cotranslational protein targeting to membrane	1.220150587726358E-11
GO:0072599	establishment of protein localization to endoplasmic reticulum	1.582424825314545E-11
GO:0022626	cytosolic ribosome	4.248636400627166E-11
GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	1.181198906422864E-10
GO:0006413	translational initiation	1.431742455207831E-10
GO:0070972	protein localization to endoplasmic reticulum	1.630606696102211E-10
GO:0019083	viral transcription	8.055995468651928E-10
GO:0044445	cytosolic part	1.285681350947508E-9
GO:0019080	viral gene expression	1.85559162312773E-9
GO:0000956	nuclear-transcribed mRNA catabolic process	4.056787154843965E-9
GO:0044391	ribosomal subunit	5.245469197708598E-9
GO:0044033	multi-organism metabolic process	6.093194142133936E-9
GO:0006402	mRNA catabolic process	1.027320336211091E-8
GO:0016071	mRNA metabolic process	2.869930681835403E-8
GO:0006612	protein targeting to membrane	3.261054233398337E-8
GO:0006401	RNA catabolic process	5.461067905155586E-8
GO:0003735	structural constituent of ribosome	2.502851132563645E-7
GO:0005840	ribosome	5.376415709514282E-7
GO:0090150	establishment of protein localization to membrane	1.022327030813531E-6
GO:0034655	nucleobase-containing compound catabolic process	1.147391261808475E-6
GO:0006364	rRNA processing	1.622175611705192E-6
GO:0019058	viral life cycle	1.842615859086327E-6
GO:0016072	rRNA metabolic process	2.152548386190508E-6
GO:0042254	ribosome biogenesis	2.410339630698724E-6
GO:0046700	heterocycle catabolic process	3.319329023154692E-6
GO:0044270	cellular nitrogen compound catabolic process	4.088880996591498E-6

v. Phenotype: COVID-19 patients VS all other respiratory diseases

Table S-1.5. Features selected in the classification between COVID-19 and all other respiratory diseases. Healthy individuals are omitted in this section. COVID-19 is studies against all other respiratory diseases. Feature importance is used in the feature selection.

NOTCH2	TCF7L2	TMEM165	C3orf58	EMR1
REL	FKBP15	CCL5	NLRC5	TRADD
DMXL2	NFKBIZ	SPAG7	MSL3	RPL22
IFI16	SAT1	RPL32	SPG11	GOT1
DIP2B	CELF2	TET3	EMD	TSPAN18
EPB41L3	MR1	TLR8	PARP9	C1QC
HDAC4	ZSWIM6	RP11-476D10.1	HABP4	RP11-68I3.11
MRPL51	FCGR1B	ELF1	IGHG1	FNIP1
PDPK1	FLVCR2	DOCK8	TRUB2	FBXO6
ZEB2	PLAGL1	TIMM44	CD163	FAHD2B
SP110	ITPRIPL2	DTX3L	RPS23	C1orf21
ST6GALNAC6	NFAT5	CD96	U2AF1L4	RLIM
NCOA7	TRIM5	STX12	ABT1 PSMC3	IGHV1-69-2
KIAA1109	MAFB	PCGF3	MYOF	LMO2
NEAT1	STAT5A	CACNA1I	MX2	KLHL6
AKAP13	CARD8	GALC	FAM83G	WDFY3
SORT1	PCNXL2	ASXL2	DICER1	PDLIM5
CD274	CYBB	SSBP1	C1QB	WARS
FGD6	KDM1B	DENND5A	STK38	TNFSF13B
UTRN	PDCD6IP	DCTPP1		SLC15A2



FigureS1-5. Feature importance value is plotted for each of the selected feature. This is for the classification of patients into whether they are COVID-19 patients or any other respiratory disease patient.

Table S-1.51. GO terms of the selected features while predicting whether patients are COVID-19 or having any other respiratory disease

ID	Name	P Value
GO:0045087	innate immune response	0.0112011004327660
GO:0006955	immune response	0.02197450286045
GO:0002376	immune system process	0.2011742693452200
GO:0006357	regulation of transcription from RNA polymerase II promoter	0.3313541551559750
GO:0050776	regulation of immune response	0.428988222323575
GO:0005829	cytosol	0.4773897847715090
GO:0006366	transcription from RNA polymerase II promoter	0.5524944864421730
GO:0045944	positive regulation of transcription from RNA polymerase II promoter	0.6258469263521210
GO:0002682	regulation of immune system process	0.676156002876787
GO:0090304	nucleic acid metabolic process	0.6924506434499850
GO:0034097	response to cytokine	0.736874362128793
GO:0006952	defense response	0.9244818180842800
GO:0032774	RNA biosynthetic process	0.969496135969837
GO:0045321	leukocyte activation	1.0
GO:0046649	lymphocyte activation	1.0
GO:0007249	I-kappaB kinase/NF-kappaB signaling	1.0
GO:0005515	protein binding	1.0
GO:0016070	RNA metabolic process	1.0
GO:0045088	regulation of innate immune response	1.0
GO:0050778	positive regulation of immune response	1.0
GO:0005634	nucleus	1.0
GO:0045089	positive regulation of innate immune response	1.0
GO:0002684	positive regulation of immune system process	1.0
GO:2001141	regulation of RNA biosynthetic process	1.0
GO:0009059	macromolecule biosynthetic process	1.0
GO:0034645	cellular macromolecule biosynthetic process	1.0
GO:0044445	cytosolic part	1.0
GO:0040029	regulation of gene expression, epigenetic	1.0
GO:0051239	regulation of multicellular organismal process	1.0

vi. Phenotype: Gender classification among COVID-19 patients

Table S-1.6. Selected features in the study of gender specification of COVID-19 patients. Only COVID-19 patients are considered in this study to predict their gender. Feature importance selected features in this study

KDM5D	ASGR2	PPP1R3F	EIF2S3L	ZNF559
XIST	METAP1D	ZFY	ZNF347	PYGB
DDX3Y	PDLIM1	MAL	PRKCQ-AS1	ANO10
RPS4Y1	U1	HELB	SLC39A7	FHIT
TXLNG	FANCF	PTPLAD1	ZNF506	C1orf132
BCORP1	CHRM3-AS2	TRAV8-2	ZNF83	SLC9A1
CATSPERB	ALDH3B1	PCED1B	ITM2A	GNAZ
MGLL	MTERF4	OBSCN	RPL5P1	CTD-2555O16.4
PRKX	LBH	LEF1	EPHX2	ZNF836
TXLNGY	RP3-477M7.5	RPL13P12	CD3G	TRIM68
GOLGA8B	DAPK3	WDR43	PSMD2	GNA15
PARVB	PSMC3	DDHD1	GP6	DDX17
GPA33	PSAP	EIF1AX	EIF2S3	TPTEP1
ITPKC	TSPAN9	CD99	RP11-498C9.3	FAM159A
USP9Y	ARHGAP12	GPR171	KYNU	WDR1
ZNF266	STAC3	EIF5	EIF1AY	SCPEP1
TRDV2	DACT1	ZNF862	AK3	SIL1
ZBTB5	CPVL	IGLV3-9	5-Sep	PABPC4
C6orf25	RP11-747H7.3	PADI6	TRA2A	GAL3ST4
THBS1	RP11-705C15.3	UBE2Q2	CEBPA	IGHV4-4

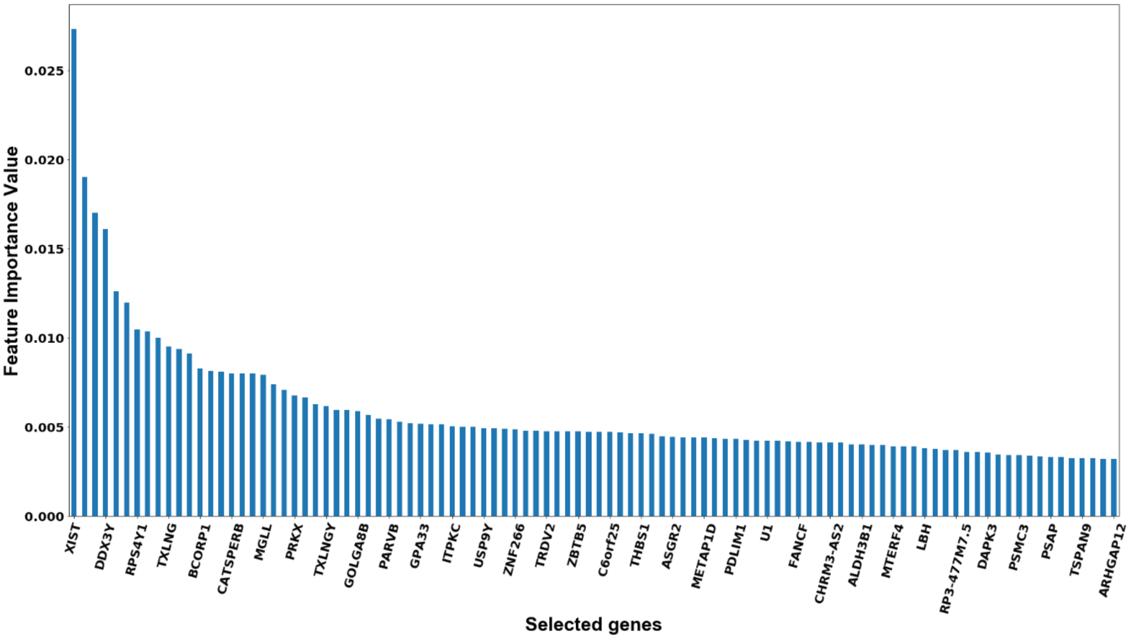


Figure S1-6. Feature importance value VS features. This is in the classification of COVID-19 patients into either male or female

Table S-1.61. GO analysis on the selected features while predicting the gender of COVID-19 patients. GO terms of this study.

ID	Name	P Value
GO:0003743	translation initiation factor activity	0.01981759159354100
GO:0006413	translational initiation	0.1752233683245370
GO:0008135	translation factor activity, RNA binding	0.3180717389104840
GO:0097197	tetraspanin-enriched microdomain	1.0
GO:0043604	amide biosynthetic process	1.0
GO:0006412	translation	1.0
GO:0043043	peptide biosynthetic process	1.0
GO:0031597	cytosolic proteasome complex	1.0
GO:0007163	establishment or maintenance of cell polarity	1.0
GO:0004185	serine-type carboxypeptidase activity	1.0
GO:0008540	proteasome regulatory particle, base subcomplex	1.0
GO:0051893	regulation of focal adhesion assembly	1.0
GO:0090109	regulation of cell-substrate junction assembly	1.0
GO:0030011	maintenance of cell polarity	1.0
GO:0007596	blood coagulation	1.0
GO:0050817	coagulation	1.0
GO:0007599	hemostasis	1.0
GO:1903391	regulation of adherens junction organization	1.0
GO:2001238	positive regulation of extrinsic apoptotic signaling pathway	1.0
GO:0070008	serine-type exopeptidase activity	1.0
GO:1902043	positive regulation of extrinsic apoptotic signaling pathway via death domain receptors	1.0
GO:0006518	peptide metabolic process	1.0
GO:0043603	cellular amide metabolic process	1.0
GO:0055002	striated muscle cell development	1.0
GO:0004004	ATP-dependent RNA helicase activity	1.0
GO:0008186	RNA-dependent ATPase activity	1.0
GO:1901566	organonitrogen compound biosynthetic process	1.0
GO:0003724	RNA helicase activity	1.0
GO:0033885	10-hydroxy-9-(phosphonooxy)octadecanoate phosphatase activity	1.0

vii. Phenotype: Hospitalization of COVID-19 patients (Hospitalized or not)

Table S-1.7: Hundred features on the study of hospitalization of COVID-19 patients. Feature importance in the selection of top hundred features while predicting the hospital status of COVID-19 patients

SEC61B	TYK2	GOT2	ZNF274	CDKN1C
MT-ND5	BMP8B	CDKN2C	TMBIM4	YLPM1
ARHGAP11A	TMEM92	CIDECP	MCU	CD177P1
SMARCC2	ENO1	AHDC1	TSC2	SGK223
TCEB1	RRM2	SELT	CHAC2	SEC11C
AAMP	TERF2	FUS	VPS11	UBL7 MT-TW
SEC61G	MYH9	15-Sep	TM7SF3	PSMB4
MRPL15	PAF1	DENND1B	KLHL3	CTPS1
TMED1	IGKV1-27	KAT7	ARHGAP35	CUL9
PRKCSH	SSR3	HK3	TMEM229B	DUSP28
DUT	LYL1	ESYT1	LINC01278	MPC2
DNAJB1	TROVE2	MAST3	GPI	GADD45GIP1
DNAJC2	PSMD6	ZNF362	SERP1	SLC9A8
ARPIN	HJURP	NLRP1	RP11-673C5.1	XBP1
ADM2	DDX10P1	RNF4	TMEM123	CDC37
SLC2A5	HNRNPU	CDCA5	IGF2BP2	ECI2
AZI2	CCNA2	ZNF267	LCN2	ANPEP
SCAP	IL10RA	PPP2CA	SPIB	PDXK
H2AFZ	FBRSL1	IGKV6D-21	DRAM1	ASMTL-AS1
SCAF1	MAP2K3	MTHFD2	METTL22	

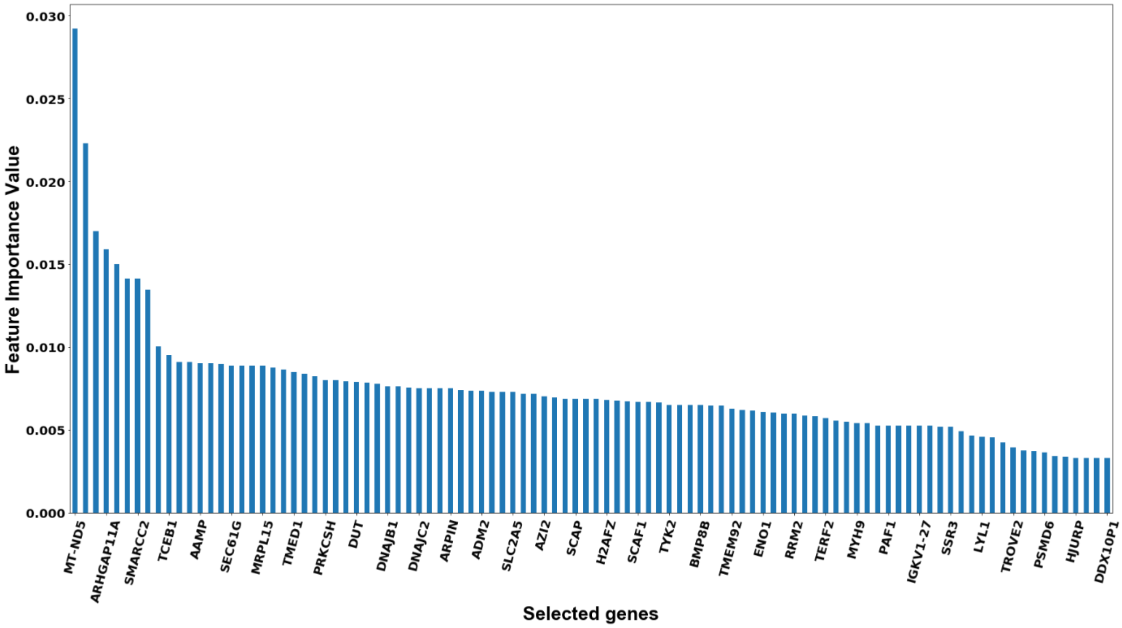


Figure S1-7. Bar plot of feature importance value of the features selected for predicting the hospital status of COVID-19 patients

Table S-1.71. GO terms of selected features in the study of hospitalization of COVID-19 patients

ID	Name	P Value
GO:0071332	cellular response to fructose stimulus	0.4380652830637410
GO:0006735	NADH regeneration	0.8950134293622740
GO:0061621	canonical glycolysis	0.8950134293622740
GO:0061718	glucose catabolic process to pyruvate	0.8950134293622740
GO:0061615	glycolytic process through fructose-6-phosphate	1.0
GO:0061620	glycolytic process through glucose-6-phosphate	1.0
GO:0036498	IRE1-mediated unfolded protein response	1.0
GO:0006007	glucose catabolic process	1.0
GO:0072524	pyridine-containing compound metabolic process	1.0
GO:0009141	nucleoside triphosphate metabolic process	1.0
GO:0006734	NADH metabolic process	1.0
GO:0009750	response to fructose	1.0
GO:0032781	positive regulation of ATPase activity	1.0
GO:0035326	enhancer binding	1.0
GO:0031491	nucleosome binding	1.0
GO:0004861	cyclin-dependent protein serine/threonine kinase inhibitor activity	1.0
GO:0009263	deoxyribonucleotide biosynthetic process	1.0
GO:0005829	cytosol	1.0
GO:0006986	response to unfolded protein	1.0
GO:0006090	pyruvate metabolic process	1.0
GO:0051082	unfolded protein binding	1.0
GO:0032968	positive regulation of transcription elongation from RNA polymerase II promoter	1.0
GO:0048029	monosaccharide binding	1.0
GO:0019320	hexose catabolic process	1.0
GO:0035966	response to topologically incorrect protein	1.0
GO:0033044	regulation of chromosome organization	1.0
GO:0009199	ribonucleoside triphosphate metabolic process	1.0
GO:0034641	cellular nitrogen compound metabolic process	1.0
GO:0009132	nucleoside diphosphate metabolic process	1.0
GO:0051726	regulation of cell cycle	1.0

viii. Phenotype: Age classification between COVID-19 patients (Patients with age greater or equal to 50 are considered as one group and other as another group)

Table S-1.8. Features selected in the grouping of COVID-19 patients into their age group. Here patients with age ≤50 are considered as group 1 and others are group 2.

INPP5B	RP5-1184F4.7	SRI	BCOR	SH3PXD2B
ENC1	ANKRD9	XXbac-BPG252P9.10	HTATSF1P2	KRT72
DGKQ	SPTBN1	CIRBP	ZC3H8	RBM15B
CLDN9	ACTBP8	PRKACB	SIGLEC17P	SPOCK2
CTSL	GNAO1	UBE2G2	AGAP1	JADE3
TIMM23B	SCART1	ZNF600	TRBV21-1	HLA-F-AS1
RNF144A	GOLGA8A	S100A10FGR	SFSWAP	RP11-658F2.8
ZNF135	PCNXL2	PAFAH1B1	ERBB2	NPM1P25
SLC25A45	SLC4A10	DYNLL1	TAS2R4	CTC-425O23.5
LBH	MS4A6A	RP11-416N2.4	SPATA1	C14orf79
TRDV2	PHYKPL	CYFIP2	RP11-23P13.6	REXO4
FAM102A	OLFM2	THBS4	ZNF514	SFXN2
PFAS	C2orf68	SGK223	EPSTI1	HMGCL
FBXL13	FAM151B	GALNT11	INPP5E	ZNF749
LINC00877	RBPMS2	MSRB2	AP000936.1	ZNF14
GPAM	RNF216	ST13	CLEC1A	TFG
ZNF264	ERO1L	ZNF850	TRMT10B	SDR42E1
GPC2	KLHL22	FBXO7	RAP1GAP	APBA2
NPAS2	LINC00894	GP6	EGLN3	ZC3H14
IGHD	EXD3		TOB2	TBC1D20

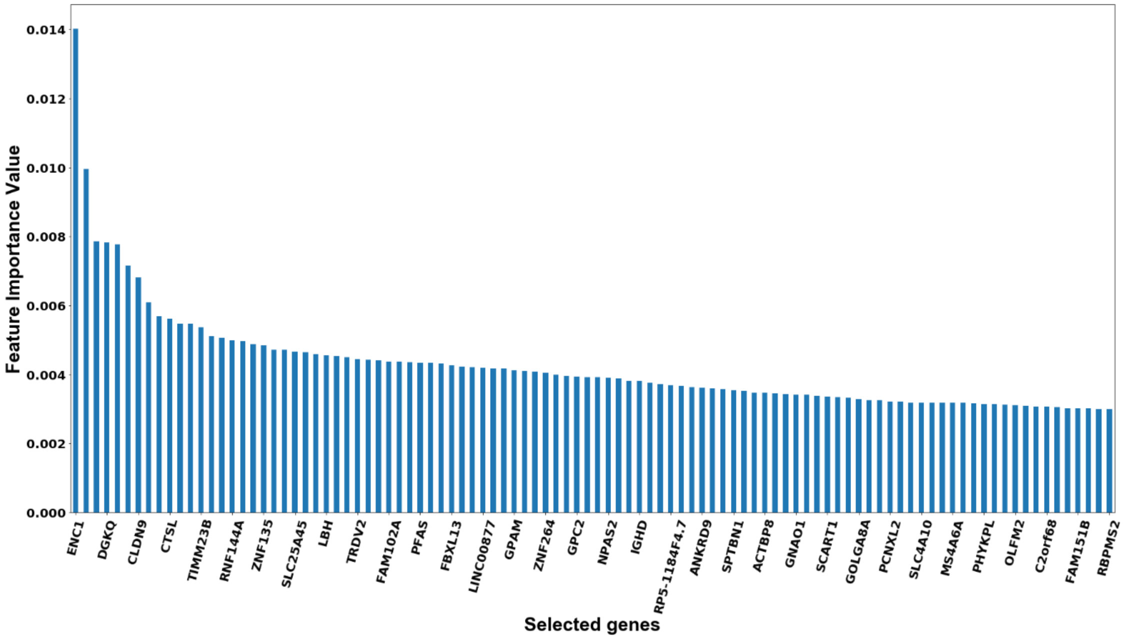


Figure S1-8. Selected features against feature importance values in the prediction of age group of COVID-19 patients

Table S-1.81. GO terms of the selected features in the prediction of age group of COVID-19 patients.

ID	Name	P Value
GO:0033085	negative regulation of T cell differentiation in thymus	1.0
GO:2000399	negative regulation of thymocyte aggregation	1.0
GO:0004439	phosphatidylinositol-4,5-bisphosphate 5-phosphatase activity	1.0
GO:1902106	negative regulation of leukocyte differentiation	1.0
GO:0045620	negative regulation of lymphocyte differentiation	1.0
GO:0045945	positive regulation of transcription from RNA polymerase III promoter	1.0
GO:0070970	interleukin-2 secretion	1.0
GO:0046030	inositol trisphosphate phosphatase activity	1.0
GO:0019471	4-hydroxyproline metabolic process	1.0
GO:0001675	acrosome assembly	1.0
GO:0034595	phosphatidylinositol phosphate 5-phosphatase activity	1.0
GO:0016671	oxidoreductase activity, acting on a sulfur group of donors, disulfide as acceptor	1.0
GO:0007626	locomotory behavior	1.0
GO:1903707	negative regulation of hemopoiesis	1.0
GO:0052745	inositol phosphate phosphatase activity	1.0
GO:0000415	negative regulation of histone H3-K36 methylation	1.0
GO:1903697	negative regulation of microvillus assembly	1.0
GO:1904425	negative regulation of GTP binding	1.0
GO:1904441	regulation of thyroid gland epithelial cell proliferation	1.0
GO:1904442	negative regulation of thyroid gland epithelial cell proliferation	1.0
GO:0090176	microtubule cytoskeleton organization involved in establishment of planar polarity	1.0
GO:1990789	thyroid gland epithelial cell proliferation	1.0
GO:1990790	response to glial cell derived neurotrophic factor	1.0
GO:1990792	cellular response to glial cell derived neurotrophic factor	1.0
GO:0097442	CA3 pyramidal cell dendrite	1.0
GO:0051150	regulation of smooth muscle cell differentiation	1.0
GO:0006359	regulation of transcription from RNA polymerase III promoter	1.0
GO:0046856	phosphatidylinositol dephosphorylation	1.0
GO:0033081	regulation of T cell differentiation in thymus	1.0

ix. Phenotype: COVID-19 VS non-COVID-19

Table S-1.9: Selected set of features while predicting a person is COVID-19 or non-COVID-19

GPR132	HMMR	SARNP	SRF	LAPTM5
BUB1	TRIP13	MELK	C1orf115	ARID5A
TOP2A	RAD51	SPRYD3	RNA28SN1	KIF11
TYMS	GPSM3	CKAP2L	VPS51	SYNGR2
BUB1B	ZNF385D	FCGRT	IGHG3	ZNF768
SGO1	PCLAF	H4C8	RPS6KA1	DHRS13
CDC34	VRK3	THEMIS2	NR1H2	RAB35
EXO1	COTL1	GGT1	CDC25C	TMEM259
SHCBP1	POLR2C	TXNDC5	SERTAD3	MDH2
PACS1	TMUB1	NCAPH	ARHGEF2	CMC4
TICRR	TMCO6	PSMD4	FUT7	B3GAT1
KLF16	MED11	ANLN	PARL	ZDHHC12
HASPIN	MAP2K3	SUGP1	TPPP3	TBC1D10A
GTSE1	MLF2	MMP17	CYTH4	GINS1
GPR108	SPAG5	RNA5S11	FAM20C	CDCA8
SLC12A9	ATP6V0D1	TMEM234	PITPNC1	ITIH1
SLC8B1	CHMP2A	RFNG	NCKAP5L	MIXL1
CATIP	WDR45	RNA5S1	PPM1M	ERCC6L
KIF14	GLDC	ELOB	G6PD	BORCS8
STIL	TRAPPC12	LSM6	RNA5S13	SLC15A3

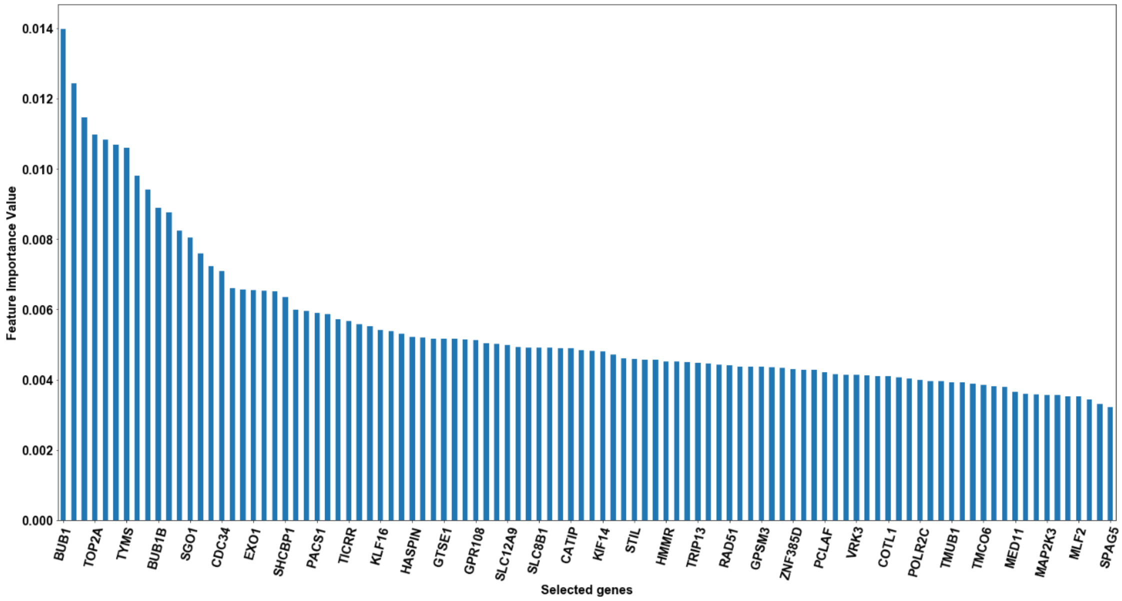


Figure S1-9. Feature importance of selected features in the classification of a patient into whether he is a COVID-19 or non-COVID-19

Table S-1.91. Results of GO analysis of the selected features in the prediction of a patient whether he is COVID-19 or non-COVID-19

ID	Name	P Value
GO:1903047	mitotic cell cycle process	2.291255656768042E-8
GO:0000278	mitotic cell cycle	1.215609077052528E-7
GO:0022402	cell cycle process	5.778255953915694E-7
GO:0007049	cell cycle	7.043651004160974E-6
GO:0000280	nuclear division	5.302995757998488E-5
GO:0048285	organelle fission	1.22624797618165E-4
GO:0007059	chromosome segregation	1.950654388870268E-4
GO:0098813	nuclear chromosome segregation	3.54283847458329E-4
GO:0000819	sister chromatid segregation	3.709058134237422E-4
GO:0051301	cell division	4.113402352321136E-4
GO:0044772	mitotic cell cycle phase transition	5.50670794958513E-4
GO:0007067	mitotic nuclear division	7.491281959433106E-4
GO:0044770	cell cycle phase transition	0.001033965593100000
GO:0015630	microtubule cytoskeleton	0.0056656736385850
GO:0007346	regulation of mitotic cell cycle	0.01021156756854700
GO:0000070	mitotic sister chromatid segregation	0.01062920987433700
GO:0010564	regulation of cell cycle process	0.01258466804823400
GO:0000793	condensed chromosome	0.02764611806951900
GO:1901990	regulation of mitotic cell cycle phase transition	0.05062667019206200
GO:1901987	regulation of cell cycle phase transition	0.08582874036807900
GO:0007088	regulation of mitotic nuclear division	0.1539796082234380
GO:0044839	cell cycle G2/M phase transition	0.1557103920581990
GO:0045930	negative regulation of mitotic cell cycle	0.2223371236211730
GO:1901991	negative regulation of mitotic cell cycle phase transition	0.2332205260679240
GO:0005815	microtubule organizing center	0.2534381583473620
GO:0005654	nucleoplasm	0.277005511812951
GO:0007093	mitotic cell cycle checkpoint	0.2834888023435440
GO:0000780	condensed nuclear chromosome, centromeric region	0.2863604445560390
GO:0005819	spindle	0.2891470672021870

x. Phenotype: ICU status of COVID-19 patients (admitted/not)

Table S-1.10. Hundred features used in the prediction of ICU status of COVID-19 patients

FGFR2	ZNF566	DCP1B	FCRL3	LIMCH1
GADD45A	ARG1	GALT	CXCL10	TAGLN2
P2RY6	MMAA	PRF1	IL32	SRPK1
TRAF3IP3	GPR68	FCER1A	ACVR2B	L3MBTL2
CKAP4	KLHL2	CSGALNACT1	CD8A	S1PR5
SEPTIN8	NFATC2	APOBEC3C	HACL1	FAM228B
LRRC70	SH2D1A	KIAA1671	SAMSN1	SERPINH1
KCNA6	PCOLCE2	CYB561	CD244	PYGL
ZNF683	MACROH2A2	SYTL2	RSPH14	DYRK2
APOBEC3H	PRKCQ	SH2D2A	AKIP1	BMP1
SARM1	FAM118A	DLG3	CD8B	PCSK9
MPP1	CIITA	PDGFRB	RASGRF2	TMEM255A
IL1R2	MCOLN2	TRAF1	TSR2	SAMD14
SLC4A8	ITGB7	AMOT	NUP93	ZBTB46
AMPH	ZNF528	SHFL	FBXO25	PRELID3B
APMAP	TBX21	SYN2	LRP10	ZNF662
ACSS1	TLR3	WFDC1	CLIC5	EDARADD
TRERF1	ASB2	CAMKK1	ABHD15	SKAP1
ADAMTS2	EOMES	PLIN5	ZSWIM5	ESYT1
USPL1	ALOX15	SIRT5	FLT3LG	MCCC1

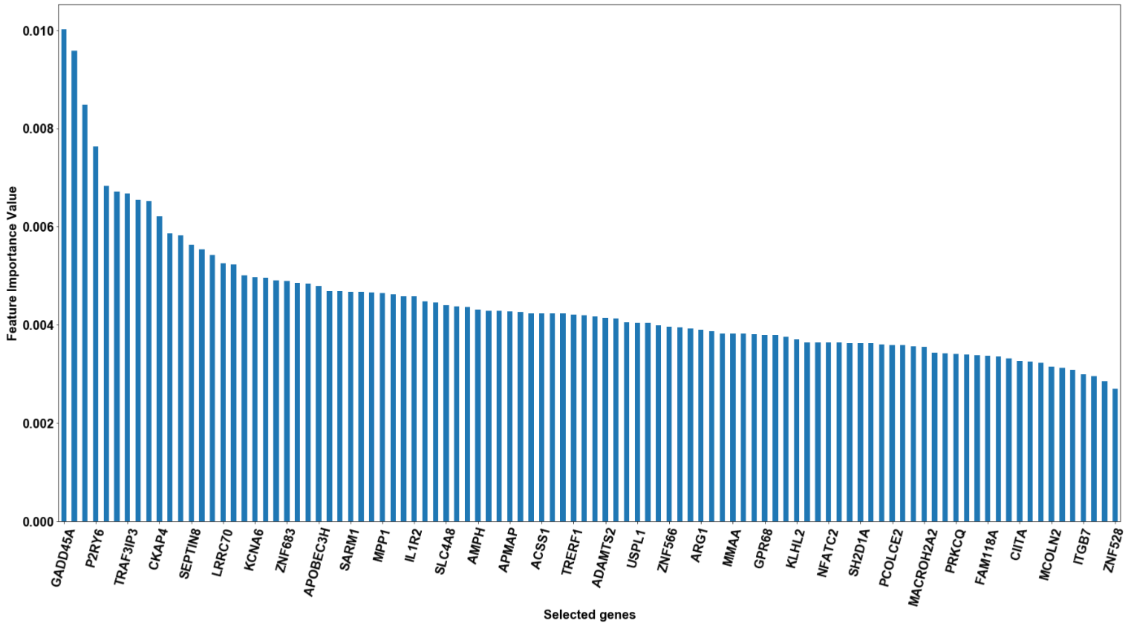


Figure S1-10. Feature importance of the selected features in the prediction of ICU status of COVID-19 patients

Table S2 1. For classifying the patients into their corresponding ‘time since onset’ stage, top 100 features with high mutual information values are selected

ID	Name	P Value
GO:0006955	immune response	0.006744047993928000
GO:0002376	immune system process	0.009119919404070000
GO:0002682	regulation of immune system process	0.02155510141119500
GO:0009897	external side of plasma membrane	0.09686090136948900
GO:0032964	collagen biosynthetic process	0.1202390473982450
GO:0002250	adaptive immune response	0.1219711348070980
GO:0009615	response to virus	0.148051231334928
GO:0051607	defense response to virus	0.1500502491206370
GO:0002252	immune effector process	0.1685917463816900
GO:0044236	multicellular organism metabolic process	0.1952742888645790
GO:0045321	leukocyte activation	0.3132403147279070
GO:0006952	defense response	0.334363957059843
GO:0042288	MHC class I protein binding	0.3536497821178470
GO:1901739	regulation of myoblast fusion	0.4183712368480110
GO:0042101	T cell receptor complex	0.4183712368480110
GO:0071865	regulation of apoptotic process in bone marrow	0.5350746155138010
GO:0071866	negative regulation of apoptotic process in bone marrow	0.5350746155138010
GO:0001775	cell activation	0.633071855606853
GO:0002697	regulation of immune effector process	0.874240843599701
GO:0032963	collagen metabolic process	0.8840451610616040
GO:0071839	apoptotic process in bone marrow	0.8887556545106520
GO:0060142	regulation of syncytium formation by plasma membrane fusion	0.970738455584199
GO:0044259	multicellular organismal macromolecule metabolic process	1.0
GO:0046649	lymphocyte activation	1.0

Table S-2.101. GO terms related to the selected features in the prediction of ICU status of COVID-19 patients

GO:0098552	side of membrane	1.0
GO:0071863	regulation of cell proliferation in bone marrow	1.0
GO:0071864	positive regulation of cell proliferation in bone marrow	1.0
GO:0002819	regulation of adaptive immune response	1.0
GO:0032814	regulation of natural killer cell activation	1.0
GO:0050776	regulation of immune response	1.0

II. Features selected using Mutual Information:

Previous ten phenotypes are again considered in the features selection using mutual information. Steps are same where feature importance is replaced by mutual information here. The output is described below.

i. Phenotype: Time since onset (Early, middle, late)

Table S-2.1. Mutual information selected features for the phenotype analysis of time onset. COVID-19 patients are classified into their level of infection. For this classification, 100 features are selected using mutual information.

ISG15	ZEB2	ZFYVE16	PIK3AP1	TAOK1
EIF4G3	SESTD1	VCAN	TCF7L2	RPL21P123
RSRP1	ANKRD44	DCP2	ETNK1	KPNB1
PUM1	STRADB	MZB1	FAR2	ATP5G1
PPIH	DNAJB2	HIVEP1	SCAF11	RNFT1
MIER1	ITM2C	UTRN	KMT2D	HELZ
GBP4	CHMP2B	KBTBD2	DIP2B	RP11-16C1.2
GBP5	NFKBIZ	LMTK2	SHMT2	RNF213
NOTCH2	KIAA2018	ZC3HAV1	OAS2	SMCHD1
MRPL9	PARP9	UBN2	RPL36AL	C18orf25
MRPL24	DTX3L	ENSG00000282939	SPTLC2	SLA2
UFC1	PARP14	FOXP3	DICER1	MYBL2
DISC1	AFF1	NBN	RCOR1	RPS15
RPS7	HERC5	VPS13B	DMXL2	C19orf66
ATAD2B	PDLIM5	CD274	TMOD2	TOMM40
ITSN2	DAPP1	RPS6	AKAP13	IGLV3-25
EIF2AK2	TET2	SIGMAR1	MRPS34	APOL6
SLC8A1	KIAA1109	ZBTB34	NFAT5	NHP2L1
REL	DDX60L	SBF2	ANKFY1	CSTB
PELI1	FYB	NEAT1	WSB1	SP110

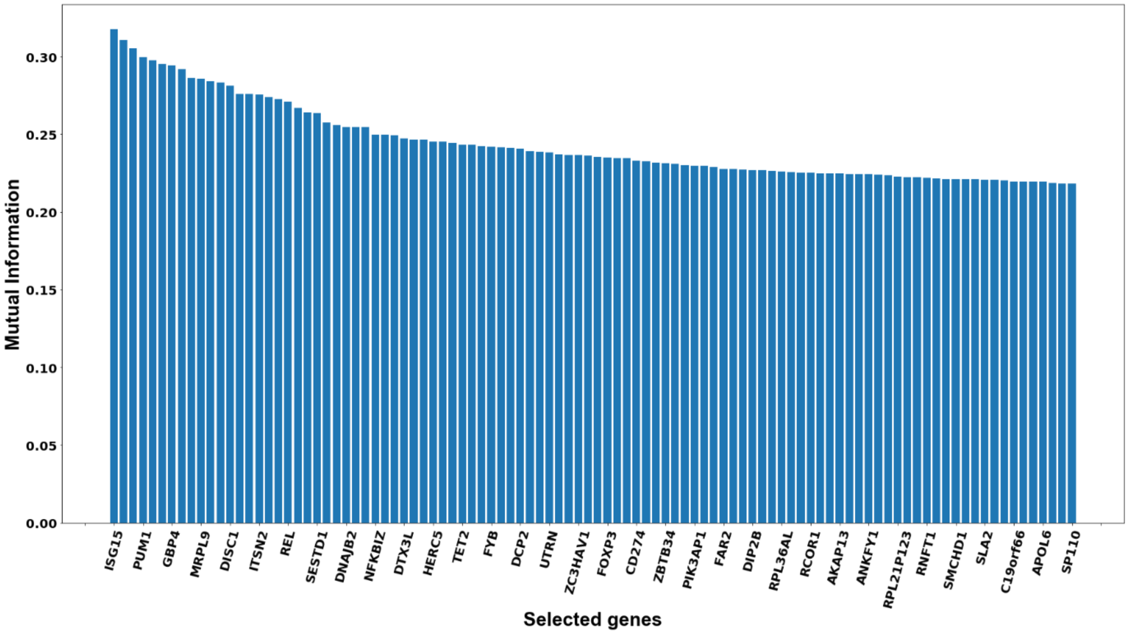


Figure S2-1. For the classification of patients into their 'time since onset' stage, features are selected using mutual information. Selected features are plotted against the corresponding mutual information.

Table S-2.11. Functional enrichment of the genes selected using mutual information in the classification of patients into their disease stage

ID	Name	P Value
GO:0005634	nucleus	0.03580704197451100
GO:0016032	viral process	0.0583963537641550
GO:0044764	multi-organism cellular process	0.06138909604830400
GO:0044403	symbiosis, encompassing mutualism through parasitism	0.08646966547667800
GO:0044419	interspecies interaction between organisms	0.08646966547667800
GO:0005840	ribosome	0.09053800031728300
GO:0045069	regulation of viral genome replication	0.1595310969607520
GO:0019058	viral life cycle	0.2722599084377580
GO:0045071	negative regulation of viral genome replication	0.3328768879086790
GO:1990904	ribonucleoprotein complex	0.3919407990370850
GO:0030529	intracellular ribonucleoprotein complex	0.3919407990370850
GO:0005829	cytosol	0.4505183752073360
GO:0003735	structural constituent of ribosome	0.4516150842793660
GO:0019079	viral genome replication	0.4957937012612440
GO:0043231	intracellular membrane-bounded organelle	0.5648320543260380
GO:0003950	NAD+ ADP-ribosyltransferase activity	0.5780018514390290
GO:0046700	heterocycle catabolic process	0.5810101144686640
GO:0044260	cellular macromolecule metabolic process	0.6153422104706410
GO:0006955	immune response	0.6378428931144960
GO:0044270	cellular nitrogen compound catabolic process	0.6508645202437450
GO:0045087	innate immune response	0.7572587153706160
GO:1903900	regulation of viral life cycle	0.7797355861232650
GO:0043933	macromolecular complex subunit organization	0.8789272958851340
GO:0001816	cytokine production	0.9985550691290680
GO:1901361	organic cyclic compound catabolic process	1.0
GO:0042274	ribosomal small subunit biogenesis	1.0
GO:0032020	ISG15-protein conjugation	1.0
GO:0043229	intracellular organelle	1.0
GO:0005515	protein binding	1.0

ii. Phenotype: Cohort (COVID-19 Bacterial Influenza Seasonal Covid and healthy)

Table S-2.2. List of selected features in the classification of patients into their disease cohort. Patients are classified into their disease cohort using transcriptome data. Mutual information selected features in this classification

AGRN	TBC1D8	MAPK14	PLBD1	SSH2
EIF4G3	GTDC1	TMEM30A	LDHB	RPL19
MTF1	ZEB2	RNF146	FGD4	ATP5G1
PPIE	STAT4	KDM7A	IRAK4	ICAM2
ZYG11B	AGFG1	KMT2C	GNS	HELZ
GBP4	HDAC4	CXCR3	ZFC3H1	LGALS3BP
SORT1	CTNNB1	PTP4A3	OAS2	SIGLEC1
MOV10	RPL29	LY6E	LATS2	AES
PHTF1	NFKBIZ	RIC1	TMCO3	RPS28
TRIM33	TMEM165	SIGMAR1	BAZ1A	TECR
CD2	WDFY3	TMEM2	NIN	FBL
PFKFB2	HERC6	AGTPBP1	TMEM229B	RPL13A
CD55	TET2	ZBTB34	DICER1	SIGLEC9
LPGAT1	CARD6	SBF2	ZNF106	LILRB3
PCNXL2	ZSWIM6	NEAT1	DPP8	USP18
KIDINS220	RPS23	CD3ELARP4	AKAP13	APOBEC3D
ITSN2	VCAN	BIFIT3	HMOX2	RPL3
NLRC4	GNB2L1	PIK3AP1	LPCAT2	EP300
PPP1R21	BTN3A3	TCF7L2	NFAT5	SAMSN1
REL	GNL1		WSB1	IL10RB

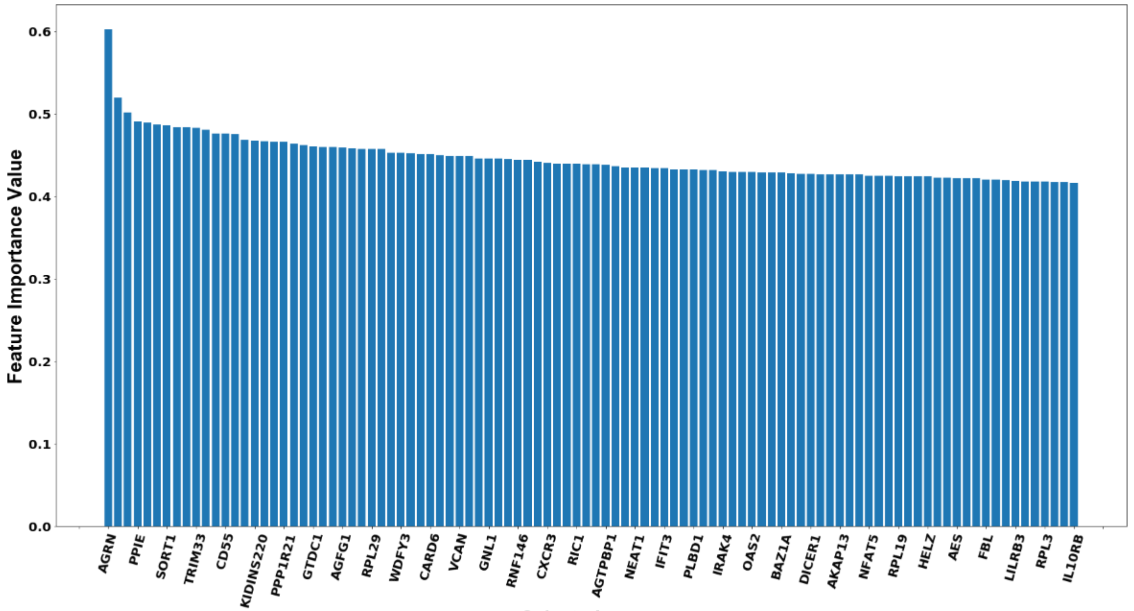


Figure S2-2. Feature importance is plotted against the features selected using mutual information in the classification of the patients into their disease cohort.

Table S-2.21. GO terms related to the selected features in the classification of patients into their disease cohort

ID	Name	P Value
GO:0005829	cytosol	0.01148598088644000
GO:0046700	heterocycle catabolic process	0.02004858982960100
GO:0044270	cellular nitrogen compound catabolic process	0.0231731143945240
GO:0006614	SRP-dependent cotranslational protein targeting to membrane	0.03333652699344400
GO:0044033	multi-organism metabolic process	0.03949948254984800
GO:1901361	organic cyclic compound catabolic process	0.04310332280881300
GO:0045047	protein targeting to ER	0.04759376090711800
GO:0006613	cotranslational protein targeting to membrane	0.05331330361258200
GO:0072599	establishment of protein localization to endoplasmic reticulum	0.05957548474544600
GO:0003723	RNA binding	0.0663686723194240
GO:0044391	ribosomal subunit	0.0766519934567760
GO:0022626	cytosolic ribosome	0.09083405385184400
GO:0070369	beta-catenin-TCF7L2 complex	0.0924803776913450
GO:0044334	canonical Wnt signaling pathway involved in positive regulation of epithelial to mesenchymal transition	0.0924803776913450
GO:0005840	ribosome	0.1060596820349360
GO:0019083	viral transcription	0.1304238853759470
GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	0.1405233410839900
GO:0019439	aromatic compound catabolic process	0.1590283523142160
GO:0070972	protein localization to endoplasmic reticulum	0.1612328945024340
GO:0019080	viral gene expression	0.1916733012290620
GO:0034097	response to cytokine	0.2099515327248330
GO:0006413	translational initiation	0.2264070364403890
GO:0016032	viral process	0.2547251437375530
GO:0044764	multi-organism cellular process	0.2666929361052830
GO:1902730	positive regulation of proteoglycan biosynthetic process	0.2764960781208460
GO:0071664	catenin-TCF7L2 complex	0.2764960781208460
GO:0010908	regulation of heparan sulfate proteoglycan biosynthetic process	0.2764960781208460
GO:0010909	positive regulation of heparan sulfate proteoglycan biosynthetic process	0.2764960781208460
GO:1990907	beta-catenin-TCF complex	0.2764960781208460
GO:0019058	viral life cycle	0.3189356868781780

iii. Phenotype: Healthy individuals VS all other patients

Table S-2.3. Hundred features selected in the stratification of patients into their healthy status. Here healthy people are classified against patients having any respiratory disease. Features are selected using mutual information.

RPL11	DBI	RPS14	TATDN1	VRK1
CD52	PPIL3	MRPL22	NSMCE2	C14orf2
NFYC	EEF1B2	DOK3	RPS6	RPS3AP47
RPS8	RP11-761N21.2	EEF1A1	ATP6V1G1	RPS3AP6
UQCRH	TMA7	HMGN3	RP11-	RPS27L
HSPB11	GNL3	TOMM7	466H18.1NUCB2	RPS17
RPL5	PRKCD	MRPL32	AP000936.1	RPS15A
RPL7P9	RP11-234A1.1	RPS3AP26,	ZNF22	ZNF646
DPH5	CCDC58	PMPCB	RPS24	RP11-51O6.1 NAE1
RPS27	RPL35A	LSM8	RP11-	SLC7A5
GAS5	RPL9	GCC1	543P15.1KLRB1	TRAPPC1
RP11-92K2.2	RP11-408P14.1	RP4-800G7.1	PFDN5	RPL26
RP11-175B9.3	RPL34	EEF1B2P3	RP11-244J10.1	RPS7P1
SNRPE	LARP7	RPL39	RPL41P5	RPL23
RPL21P28	SNHG8	RPL7	RPL6	RPL6P27
TOMM20	RPS3A	C8orf59	RPL21	SNORD17
RPS7	NDUFS4	UQCRB	TPT1	GTPBP1
RPS27A	RPL26P19	RPL30	LCP1	RPL41
SNRPG	COX7C	DCAF13	COMMD6	UBE2Q2P6
RPL31	HINT1	EIF3E	RPS29	

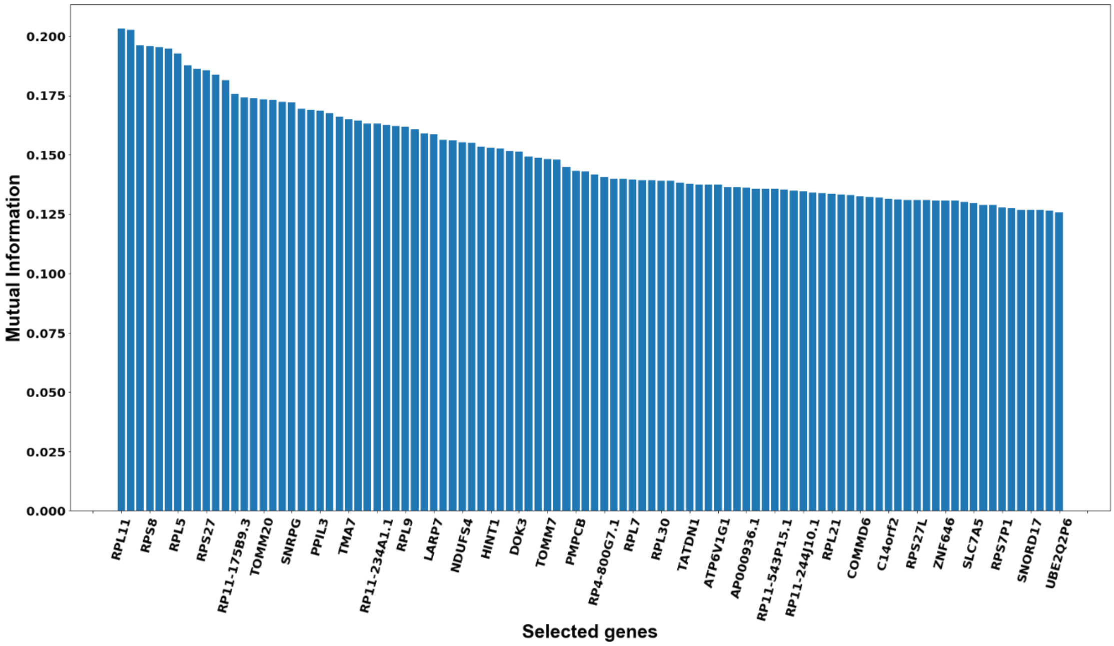


Figure S2-3. Mutual information selected features along with their mutual information value. These features are selected in the classification of the patients into their healthy status (are they healthy or having any respiratory disease)

Table S-2.31. Functional enrichment terms related to the selected features in the classification of the patients into their healthy status

ID	Name	P Value
GO:0006614	SRP-dependent cotranslational protein targeting to membrane	1.967674802335779E-33
GO:0022626	cytosolic ribosome	3.577081497989624E-33
GO:0045047	protein targeting to ER	1.138720613052256E-32
GO:0006613	cotranslational protein targeting to membrane	1.989515849812542E-32
GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	3.3252534623713E-32
GO:0072599	establishment of protein localization to endoplasmic reticulum	3.432487562317742E-32
GO:0044391	ribosomal subunit	4.459996000221644E-32
GO:0070972	protein localization to endoplasmic reticulum	4.460914896845869E-30
GO:0003735	structural constituent of ribosome	3.760095701871648E-28
GO:0005840	ribosome	2.247571750098905E-27
GO:0006402	mRNA catabolic process	5.016328611019078E-27
GO:0006413	translational initiation	1.282188554499871E-26
GO:0000956	nuclear-transcribed mRNA catabolic process	2.951625774389756E-26
GO:0019083	viral transcription	5.20588115283483E-26
GO:0006401	RNA catabolic process	1.58838621985432E-25
GO:0006612	protein targeting to membrane	2.537976941693086E-25
GO:0019080	viral gene expression	2.537976941693086E-25
GO:0044445	cytosolic part	7.226678224376706E-25
GO:0006364	rRNA processing	9.433339320269338E-25
GO:0016072	rRNA metabolic process	1.783486591466732E-24
GO:0044033	multi-organism metabolic process	2.42422904028114E-24
GO:0042254	ribosome biogenesis	1.670694955615947E-23
GO:0034655	nucleobase-containing compound catabolic process	1.965329166489828E-23
GO:0022613	ribonucleoprotein complex biogenesis	6.966634004326972E-23
GO:0046700	heterocycle catabolic process	1.826986920937351E-22
GO:0044270	cellular nitrogen compound catabolic process	2.83293603096527E-22
GO:0019439	aromatic compound catabolic process	4.361210548360289E-22
GO:1901361	organic cyclic compound catabolic process	1.871108252225099E-21
GO:1990904	ribonucleoprotein complex	4.084997933676512E-21

iv. Phenotype: Healthy individuals VS COVID-19 patients

Table S-2.4. Top hundred features are selected while classifying the COVID-19 patients against healthy individuals. Patients with other diseases are not considered in this section

ATAD3A	TMA7	RPL39	TPI1	SLC7A5
RPL11	CCDC58	RPL7	PFDN5	PFN1
SH3BGRL3	EIF4A2	C8orf59	RP11-244J10.1	TRAPPC1
CD52	RPL35A	UQCRB	RPL6	RPL26
RABGGTB	RPL9	COX6C	RSRC2	RPS7P1
RPL5	COMMD8	POLR2K	RPL21	RPL23
RPL7P9	RPL34	TATDN1	UFM1	MAP3K14
TRIM33	SNHG8	RANBP6	TPT1	DDX5
RPS27	RPS3A	RPS6	COMMD6	EXOC7
GAS5	COX7C	TOLLIP	DCAF11	ARHGDIA
RPL21P28	HINT1	LSP1	RPS29	RPS15AP1
TOMM20	LMAN2	RP11-466H18.1	RPL41P2	RALY
RPS7	NDUFA4	PPP1CA	RPS3AP47	DYNLRB1
NDUFAF7	TOMM7	CLNS1A	RSL24D1	ADRM1
RPS27A	RPS3AP26	AP000936.1	RPS3AP6	ASNA1
MITD1	LSM8	RPS24	FAM96A	CAPNS1
RPL31	NDUFA5	BTAF1	RPS17	AKT2
CLK1	RP4-800G7.1	FAM204A	ZNF213	DEDD2
EEF1B2	EEF1B2P3	RP11-572P18.1	RPS15A	RPL41
RP11-761N21.2	COX7B	RP11-543P15.1	ZNF646	UBE2Q2P6

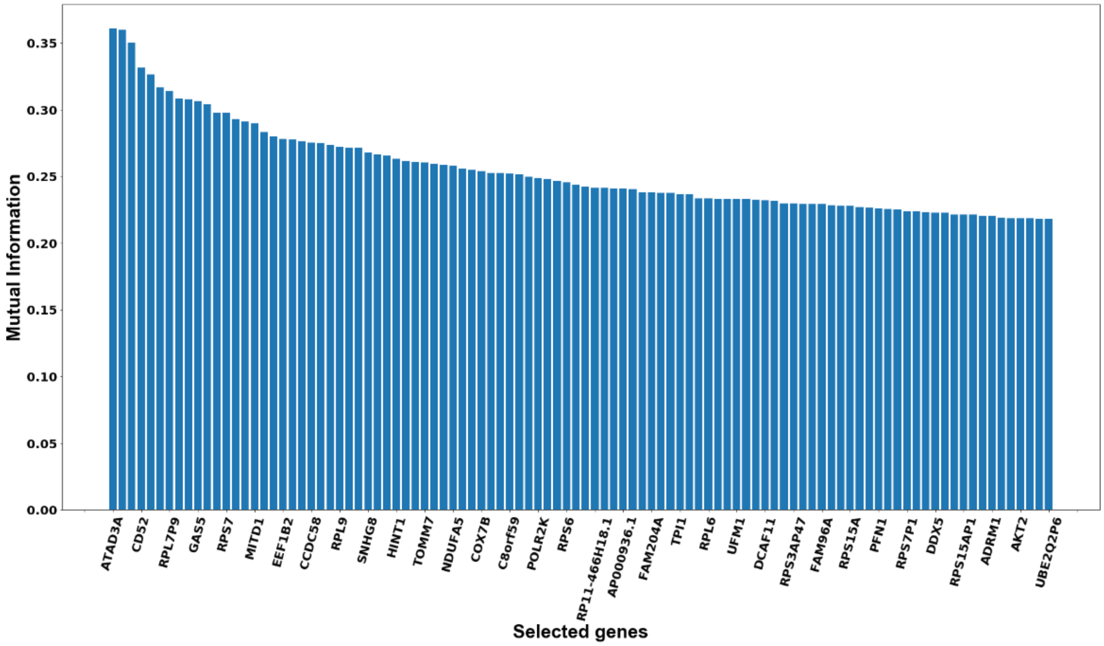


Figure S2-4. Bar plot between selected features and their corresponding mutual information value. This is on the classification of patients into, whether they are COVID-19 or healthy

Table S-2.41. Selected genes are tested for gene enrichment analysis. List of GO terms on the features selected while classifying them into healthy or COVID-19

ID	Name	P Value
GO:0022626	cytosolic ribosome	8.252735198284982E-26
GO:0006614	SRP-dependent cotranslational protein targeting to membrane	8.588748262121286E-26
GO:0045047	protein targeting to ER	3.59692117279165E-25
GO:0006613	cotranslational protein targeting to membrane	5.673367979155382E-25
GO:0072599	establishment of protein localization to endoplasmic reticulum	8.85896078377486E-25
GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	2.751965321499435E-23
GO:0070972	protein localization to endoplasmic reticulum	4.768825544113978E-23
GO:0019083	viral transcription	8.523479937204124E-23
GO:0019080	viral gene expression	3.611566365389238E-22
GO:0044391	ribosomal subunit	4.662576117612244E-22
GO:0044033	multi-organism metabolic process	2.82628268008268E-21
GO:0006612	protein targeting to membrane	1.223226234732029E-20
GO:0006413	translational initiation	2.219282167851668E-20
GO:0000956	nuclear-transcribed mRNA catabolic process	4.436918675276892E-20
GO:0006401	RNA catabolic process	1.260289897766749E-19
GO:0006402	mRNA catabolic process	2.050843301972785E-19
GO:0003735	structural constituent of ribosome	4.684009361503728E-19
GO:0044445	cytosolic part	6.330852417641404E-19
GO:0005840	ribosome	1.841949749303713E-18
GO:0034655	nucleobase-containing compound catabolic process	4.004774395140669E-18
GO:0016072	rRNA metabolic process	2.217771661936978E-17
GO:0046700	heterocycle catabolic process	2.652941902774213E-17
GO:0044270	cellular nitrogen compound catabolic process	3.847999066761625E-17
GO:0019439	aromatic compound catabolic process	5.547173524433062E-17
GO:0090150	establishment of protein localization to membrane	6.060822073827918E-17
GO:0022625	cytosolic large ribosomal subunit	1.2067789922562E-16
GO:1901361	organic cyclic compound catabolic process	1.90598044792777E-16
GO:0006364	rRNA processing	2.989030349094696E-16
GO:0042254	ribosome biogenesis	1.699196239290036E-15

v. Phenotype: COVID-19 patients VS all other respiratory diseases

Table S-2.5. Features selected in the classification between COVID-19 and all other respiratory diseases. Healthy individuals are omitted in this section. COVID-19 is studies against all other respiratory diseases. Mutual information is used in the feature selection.

SRM	RRP9	ARHGAP26	PTP4A3	DICER1
MTF1	ATXN7	HIVEP1	DOCK8	ZNF106
GBP4	NFKBIZ	KDM1B	CD274	SPG11
SORT1	PARP9	HSPA1B	SIGMAR1	DMXL2
HIPK1	DTX3L	ELOVL5	ZCCHC6	CSNK1G1
CD2	PLSCR1	FAM46A	HIATL1	AKAP13
NOTCH2	IL1RAP	PHACTR2	ZBTB34	HMOX2
MRPL9	CPEB2	PLAGL1	TRIM22	ADCY7
MRPL24	SCARB2	STX11	SBF2	LPCAT2
SFT2D2	WDFY3	UTRN	RP11-750H9.5	NUFIP2
PCNXL2	AFF1	DOCK4	NEAT1	RHOT1
BIRC6	PDLIM5	ZC3HAV1	LARP4B	ATP5G1
HEATR5B	TET2	UBN2	PIK3AP1	HELZ
EIF2AK2	KIAA1109 DDX60L	KMT2C	TCF7L2	RHBDF2
REL	ZSWIM6	CYBB	LDHB	SMCHD1
TET3	LHFPL2	FOXP3	SCAF11	CARD8
ZEB2	ZFYVE16	R3HCC1	DIP2B	LILRB1
GPD2	RPS23	NCOA2	GNS	EP300
ANKRD44	VCAN	PDP1	RPL36AL	RRP1
HDAC4		VPS13B	SPTLC2	SP110

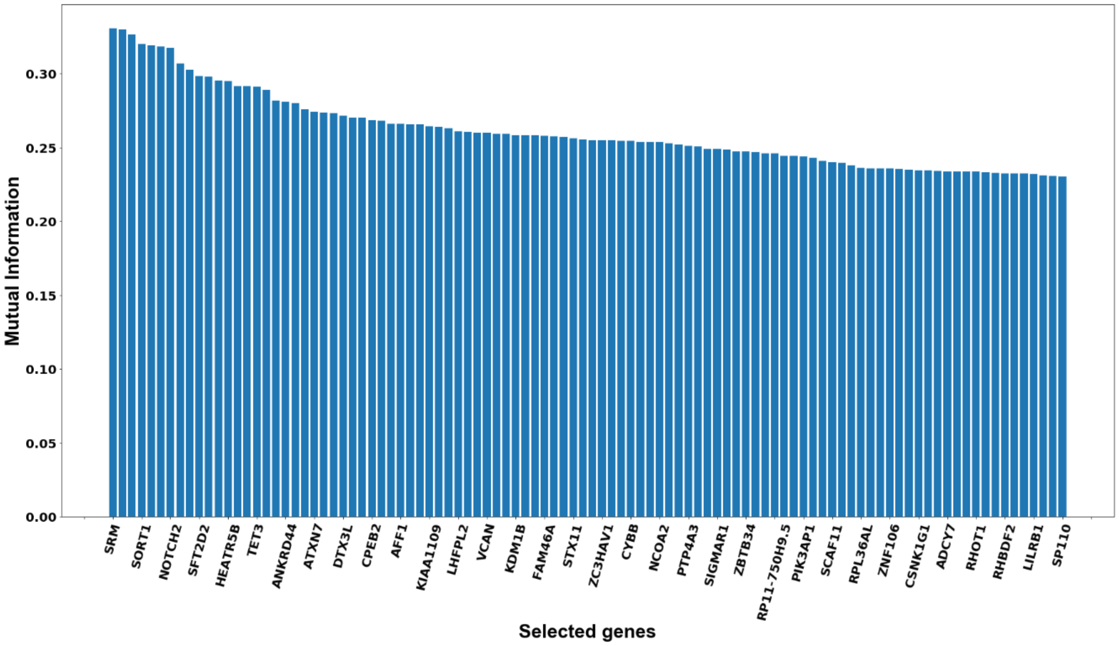


Figure S2-5. Mutual information value is plotted for each of the selected feature. This is for the classification of patients into whether they are COVID-19 patients or any other respiratory disease patient.

Table S-2.51. GO terms of the selected features while predicting whether patients are COVID-19 or having any other respiratory disease

ID	Name	P Value
GO:0016570	histone modification	0.2390918807461020
GO:0032693	negative regulation of interleukin-10 production	0.2424952758960960
GO:0046872	metal ion binding	0.4514612710062230
GO:0044428	nuclear part	0.497580504141544
GO:0070579	methylcytosine dioxygenase activity	0.5305849274004100
GO:0043169	cation binding	0.5569443210556680
GO:0043167	ion binding	0.563505343395928
GO:0043231	intracellular membrane-bounded organelle	0.6370242903084050
GO:0005840	ribosome	0.7436884792230050
GO:0045077	negative regulation of interferon-gamma biosynthetic process	0.8812983100916230
GO:0005634	nucleus	1.0
GO:0042788	polysomal ribosome	1.0
GO:0016569	covalent chromatin modification	1.0
GO:0031981	nuclear lumen	1.0
GO:0032689	negative regulation of interferon-gamma production	1.0
GO:0001817	regulation of cytokine production	1.0
GO:0045944	positive regulation of transcription from RNA polymerase II promoter	1.0
GO:0070013	intracellular organelle lumen	1.0
GO:0043233	organelle lumen	1.0
GO:0031974	membrane-enclosed lumen	1.0
GO:2001179	regulation of interleukin-10 secretion	1.0
GO:0060297	regulation of sarcomere organization	1.0
GO:1901566	organonitrogen compound biosynthetic process	1.0
GO:1901564	organonitrogen compound metabolic process	1.0
GO:0001819	positive regulation of cytokine production	1.0
GO:0044422	organelle part	1.0
GO:0044446	intracellular organelle part	1.0
GO:0050708	regulation of protein secretion	1.0
GO:0045087	innate immune response	1.0
GO:0050707	regulation of cytokine secretion	1.0

vi. Phenotype: Gender classification among COVID-19 patients

Table S-2.6. Selected features in the study of gender specification of COVID-19 patients. Only COVID-19 patients are considered in this study to predict their gender. Mutual information selected features in this study

ENO1	GCNT4	ALG13	FBLN5	ZNF525
VPS13D	CAMK4	6-Sep	IGHV4-31	ZNF416
CAP1	ATG12	CD40LG	GOLGA8B	RPS4Y1
FAM159A	TCF7	DNASE1L1	THBS1	ZFY
FCGR2C	FIG4	CLU	RMDN3	PRKY
CHRM3-AS2	CLIP2	LINC00861	ADAMTS7P1	TTTY15
LBH	HIPK2	RP11-213G2.3	PKD1P5	USP9Y
TRABD2A	GIMAP1	RHOG	GP1BA	DDX3Y
PTPN18	PRKX	ILK	ZNF594	UTY
METAP1D	RP11-706O15.1	TPP1	ALOX12	TTTY14
ORC2	PIGA	CAPN1	ASGR2	BCORP1
MTERF4	CA5B	ALDH3B1	TRAPPC1	TXLNGY
HDLBP	TXLNG	VWA5A	CTC1	KDM5D
IFRD2	EIF1AX	ESAM	CEP95	EIF1AY
CD96	EIF2S3	PSAP	RAB40B	5-Sep
MGLL	ZFX	RP11-705C15.3	UBOX5	TRMT2A
SENP5	KDM6A	HELB	C20orf27	GAS2L1
LEF1	KDM5C	NR2C1	ZNF337	NAGA
INPP4B	XIST	DDHD1	CEBPA	PARVB
IL7R	VSIG1	CATSPERB	ZNF382	KIAA0930

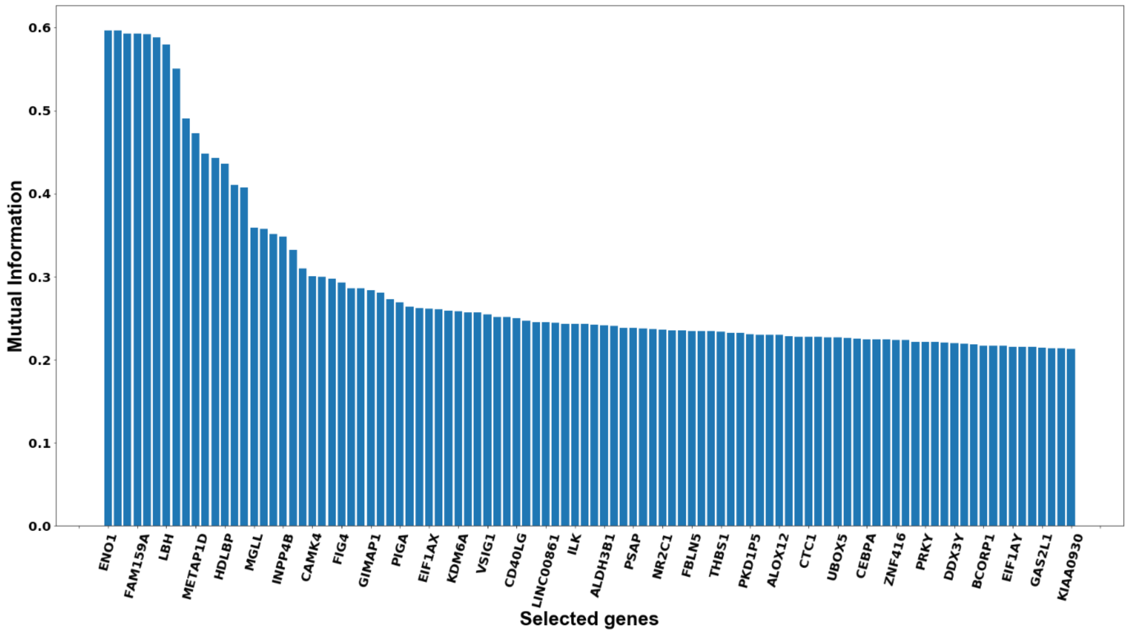


Figure S2-6. Mutual information value VS features. This is in the classification of COVID-19 patients into male or female

Table S-2.61. GO analysis on the selected features while predicting the gender of COVID-19 patients. GO terms of this study.

ID	Name	P Value
GO:0070076	histone lysine demethylation	0.01557084526202000
GO:0032452	histone demethylase activity	0.02206669991346600
GO:0016577	histone demethylation	0.03037023248485700
GO:0006482	protein demethylation	0.04077349333323800
GO:0008214	protein dealkylation	0.04077349333323800
GO:0032451	demethylase activity	0.06913283931456400
GO:0051213	dioxygenase activity	0.2788465726964090
GO:0070988	demethylation	0.6063869170622100
GO:0033153	T cell receptor V(D)J recombination	0.7268247471260320
GO:0002568	somatic diversification of T cell receptor genes	0.7268247471260320
GO:0002681	somatic recombination of T cell receptor gene segments	0.7268247471260320
GO:0071557	histone H3-K27 demethylation	0.7268247471260320
GO:0071558	histone demethylase activity (H3-K27 specific)	0.7268247471260320
GO:0032453	histone demethylase activity (H3-K4 specific)	1.0
GO:0034720	histone H3-K4 demethylation	1.0
GO:0034596	phosphatidylinositol phosphate 4-phosphatase activity	1.0
GO:0030193	regulation of blood coagulation	1.0
GO:1900046	regulation of hemostasis	1.0
GO:0050818	regulation of coagulation	1.0
GO:0030168	platelet activation	1.0
GO:0007596	blood coagulation	1.0
GO:0008375	acetylglucosaminyltransferase activity	1.0
GO:0050817	coagulation	1.0
GO:0007599	hemostasis	1.0
GO:0030195	negative regulation of blood coagulation	1.0
GO:1900047	negative regulation of hemostasis	1.0
GO:0033151	V(D)J recombination	1.0
GO:2000353	positive regulation of endothelial cell apoptotic process	1.0
GO:0003743	translation initiation factor activity	1.0
GO:0050819	negative regulation of coagulation	1.0

vii. Phenotype: Hospitalization of COVID-19 patients (Hospitalized or not)

Table S-2.7. Hundred features in the study of hospitalization of COVID-19 patients. Mutual information in the selection of top hundred features while predicting the hospital status of COVID-19 patients

SKI	SELT	SEC61G	ESYT1	TMEM92
MIIP	SSR3	AP1S1	SMARCC2	NME1
HDAC1	ECE2	MSN	R3HDM2	DCAF7
C1orf43	IGJ	MRPL15	BTG1	PSMD12
GON4L	OSTC	GGH	ALG5	BIRC5
HNRNPU	MAD2L1	TCEB1	PLD4	PYCR1
RRM2	MRPL36	ST3GAL1	WDR76	SEC11C
TP53I3	REEP5	CKS2	DUT	NFATC1
HEATR5B	TCF7	SEC61B	PIF1	PCNA
PRADC1	SEC24A	FNBP1	RP11-673C5.1	GINS1
MTHFD2	TXNDC15	VAV2	TICRR	UBE2C
MGAT5	HNRNPA0	BRD3	PLK1	SLCO4A1
PSMD14	CYFIP2	LDHA	ARHGAP17	MYDGF
AAMP	GMNN	GYLTL1B	FUS	KLF2
HJURP	HLA-DOA	HNRNPUL2	MT1F	ZC3H4
CIDECP	COX7A2	PACS1	FAM96B	SPIB
MANF	UBE2J1	SPCS2	AURKB	ISOC2
ARF4	EZR	CWC15	CDC6	ATF4
PSMD6	RPA3	POLL	UBTF	MT-ND4
MRPS22	CYCS	ABLIM1	KAT7	MT-ND5

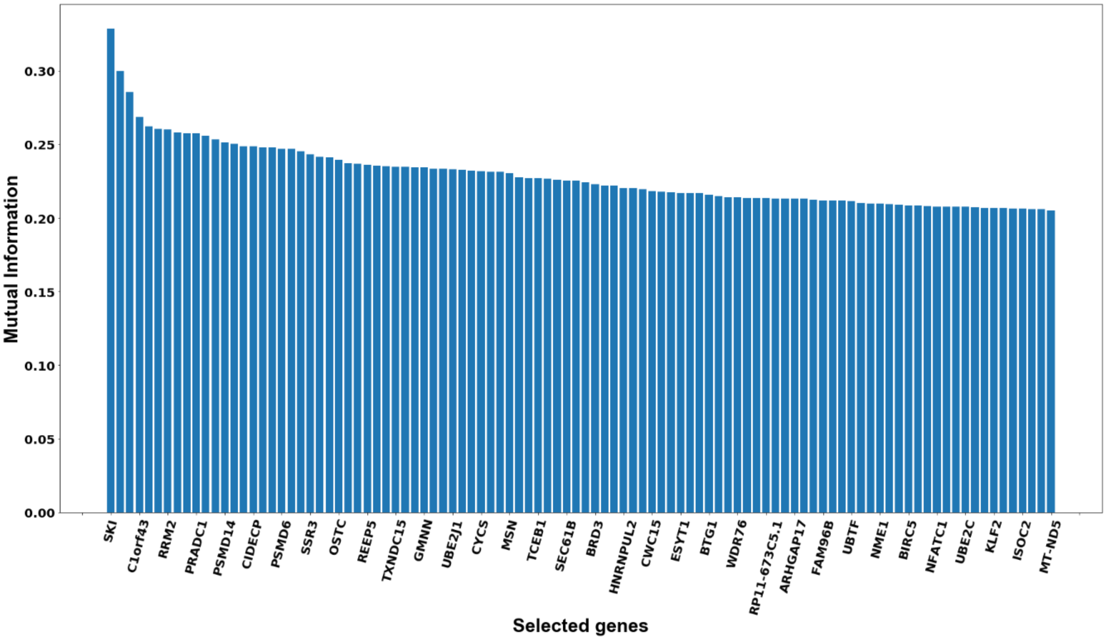


Figure S2-7. Bar plot of mutual information value of the features selected for predicting the hospital status of COVID-19 patients

Table S-2.71. GO terms of selected features in the study of hospitalization of COVID-19 patients

ID	Name	P Value
GO:0031145	anaphase-promoting complex-dependent catabolic process	4.392990575271873E-4
GO:0044446	intracellular organelle part	0.001013316239810000
GO:0043231	intracellular membrane-bounded organelle	0.001599789553094000
GO:0043227	membrane-bounded organelle	0.002051632872525000
GO:0044422	organelle part	0.002745780930395000
GO:0051439	regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle	0.00593979719074800
GO:0006260	DNA replication	0.006949249769043000
GO:1901990	regulation of mitotic cell cycle phase transition	0.007630624791075000
GO:0051437	positive regulation of ubiquitin-protein ligase activity involved in regulation of mitotic cell cycle transition	0.008172936721043000
GO:1904668	positive regulation of ubiquitin protein ligase activity	0.01188080271323300
GO:1901987	regulation of cell cycle phase transition	0.01384881810004100
GO:1903364	positive regulation of cellular protein catabolic process	0.01391618884241200
GO:1904666	regulation of ubiquitin protein ligase activity	0.01574421711293900
GO:0043226	organelle	0.01832570452476800
GO:0005654	nucleoplasm	0.01841662125209700
GO:2000060	positive regulation of protein ubiquitination involved in ubiquitin-dependent protein catabolic process	0.02194818861934500
GO:0006259	DNA metabolic process	0.02245947956347800
GO:0044772	mitotic cell cycle phase transition	0.02476576454461200
GO:0043229	intracellular organelle	0.02937907882045700
GO:0070013	intracellular organelle lumen	0.030132166817850000
GO:0043233	organelle lumen	0.0303146765433350
GO:0031974	membrane-enclosed lumen	0.0303146765433350
GO:2000058	regulation of protein ubiquitination involved in ubiquitin-dependent protein catabolic process	0.035882326798220000
GO:0051443	positive regulation of ubiquitin-protein transferase activity	0.04029614321783900
GO:0044770	cell cycle phase transition	0.04169781731532400
GO:0042787	protein ubiquitination involved in ubiquitin-dependent protein catabolic process	0.04420176259956900
GO:0007346	regulation of mitotic cell cycle	0.06890482324075000
GO:0051438	regulation of ubiquitin-protein transferase activity	0.08932048675746100
GO:0051436	negative regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle	0.1004346741629930

viii. Phenotype: Age classification between COVID-19 patients

Table S-2.8. Features selected in the grouping of COVID-19 patients into their age group. Here patients with age ≤50 are considered as group 1 and others are group 2.

MTATP6P1	NCAPG	CD40LG	NFE2,	DHRS7B
PGD	TXK	SGK223	TESPA1	SYNRG
ATPIF1	OTUD4	E2F5	ITGA7	GSDMB
AZIN2	SMAD1	DNAJA1	TDG	NMT1
RWDD3	CCL28	PAX5	TMEM119	RP13-890H12.2
SMG5	ERCC8	CDK20	GPR133	CRHR1-IT1
1-Mar	ENC1	POMT1	RNASE6	COX11
PCGF1	AC116366.6	APBB1	RNASE1	UBE2O
HK2	UBE2D2	PPP6R3	TRDV2	DNTTIP1
C2orf68	HNRNPH1	C2CD3	RBM23	UCKL1
FZD5	NQO2	ADARB2	APBA2	GADD45B
SNED1	HIST1H2BG	VIM	CCNB2	TGFB1
TMEM40	NUDT3	PSAP	UBAP1L	MEGF8
CDC25A	YIPF3	MICU1	ZSCAN2	GP6
RP13-131K19.7	ACTBP8	C10orf2	NUBP1	PI4KA
RFT1	RAB32	LZTS2	SPG7	MIAT
TIMMDC1	CARD11	RPARP-AS1	RP11-104N10.2	TRIOBP
PARP9	P2RY8	TULP3	PRPF8	GTPBP1
RASA2	TIMP1	SSPN	SRR	NPTXR
DGKQ	LINC01278	ITGB7	MPRIP	TOB2

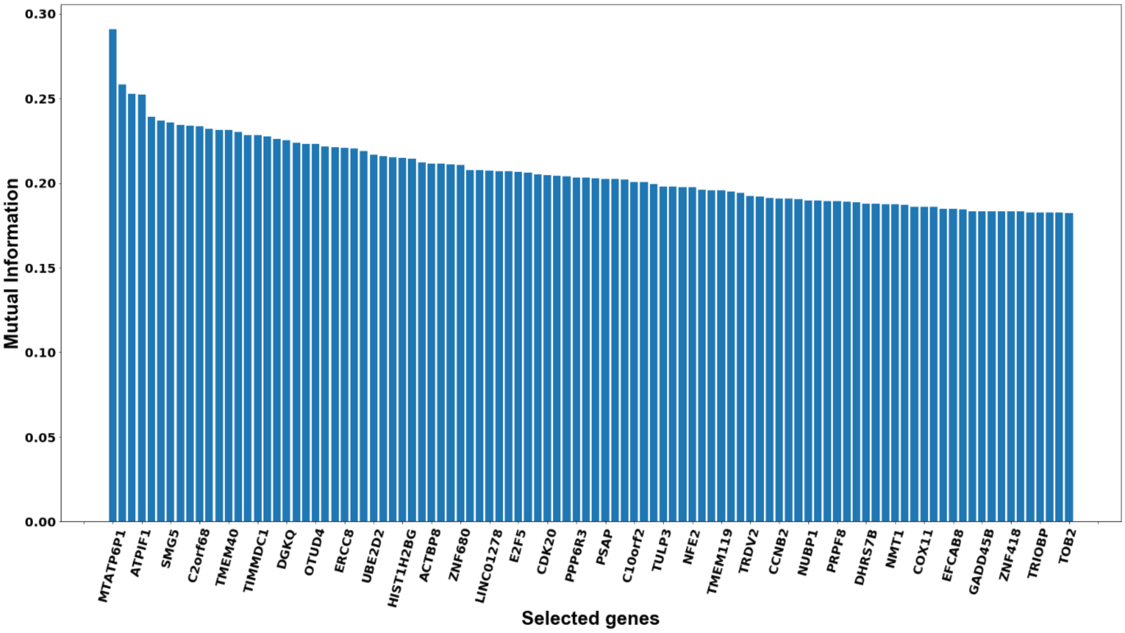


Figure S2-8. Selected features against mutual information values in the prediction of age group of COVID-19 patients

Table S-2.81. GO terms of the selected features in the prediction of age group of COVID-19 patients.

ID	Name	P Value
GO:1904888	cranial skeletal system development	0.6573558252687690
GO:0001540	beta-amyloid binding	1.0
GO:0045778	positive regulation of ossification	1.0
GO:0007229	integrin-mediated signaling pathway	1.0
GO:0007183	SMAD protein complex assembly	1.0
GO:0060348	bone development	1.0
GO:0097094	craniofacial suture morphogenesis	1.0
GO:0080135	regulation of cellular response to stress	1.0
GO:0072655	establishment of protein localization to mitochondrion	1.0
GO:0070585	protein localization to mitochondrion	1.0
GO:0060395	SMAD protein signal transduction	1.0
GO:0009880	embryonic pattern specification	1.0
GO:0050901	leukocyte tethering or rolling	1.0
GO:0046902	regulation of mitochondrial membrane permeability	1.0
GO:1903747	regulation of establishment of protein localization to mitochondrion	1.0
GO:0098629	trans-Golgi network membrane organization	1.0
GO:1901664	regulation of NAD+ ADP-ribosyltransferase activity	1.0
GO:1901666	positive regulation of NAD+ ADP-ribosyltransferase activity	1.0
GO:1902071	regulation of hypoxia-inducible factor-1alpha signaling pathway	1.0
GO:1902073	positive regulation of hypoxia-inducible factor-1alpha signaling pathway	1.0
GO:0036361	racemase activity, acting on amino acids and derivatives	1.0
GO:0070179	D-serine biosynthetic process	1.0
GO:0071543	diphosphoinositol polyphosphate metabolic process	1.0
GO:0071544	diphosphoinositol polyphosphate catabolic process	1.0
GO:0008721	D-serine ammonia-lyase activity	1.0
GO:0008792	arginine decarboxylase activity	1.0
GO:0009814	defense response, incompatible interaction	1.0
GO:0009817	defense response to fungus, incompatible interaction	1.0
GO:0047661	amino-acid racemase activity	1.0
GO:0015961	diadenosine polyphosphate catabolic process	1.0

ix. Phenotype: COVID-19 positive VS COVID-19 negative

Table S-2.9. Table S-1.9: Selected set of features while predicting whether a person is COVID-19 or non-COVID-19

ABHD11	CHMP2A	GLDC	MELK	SGO1
ACAA1	CKAP2L	GNB2	MKI67	SIPA1
ACTB	CLTB	GPR132	NCAPH	SKA3
ALDOA	COTL1	GRK6	NDC80	SLC12A9
ASPM	CSK	GTSE1	NUDT22	SLC15A3
BCAP31	CYBA	HJURP	NUF2	SLC39A4
BUB1	CYTH4	HMMR	NUSAP1	SLC8B1
BUB1B	DBNL	HPS6	OSGIN1	STIL
C19orf38	DEPDC1B	IGF1	PBK	SUMF1
CAPZB	DHRSX	IGHG1	PCLAF	SYNGR2
CCNA2	DIAPH3	JCHAIN	PHKA1	SYTL1
CDC25A	DLGAP5	KCNN3	POLQ	TALDO1
CDC45	DTL	KIF14	RAB1B	TBC1D10B
CDC6	E2F8	KIF20A	RALY	TNFRSF6B
CDCA2	ESCO2	KLF16	RCN3	TOP2A
CEBPA	EXO1	LAPTM5	RNA18SN2	TPX2
CENPE	FTL	LSP1	RNA18SN3	TRIP13
CEP55	FUT7	MAP1S	RNA18SN4	TWF2
CHEK1	GBGT1	MCM10	RNA28SN4	UCHL1
CHMP1A	GINS1	MED11	RRM2	ZNF622

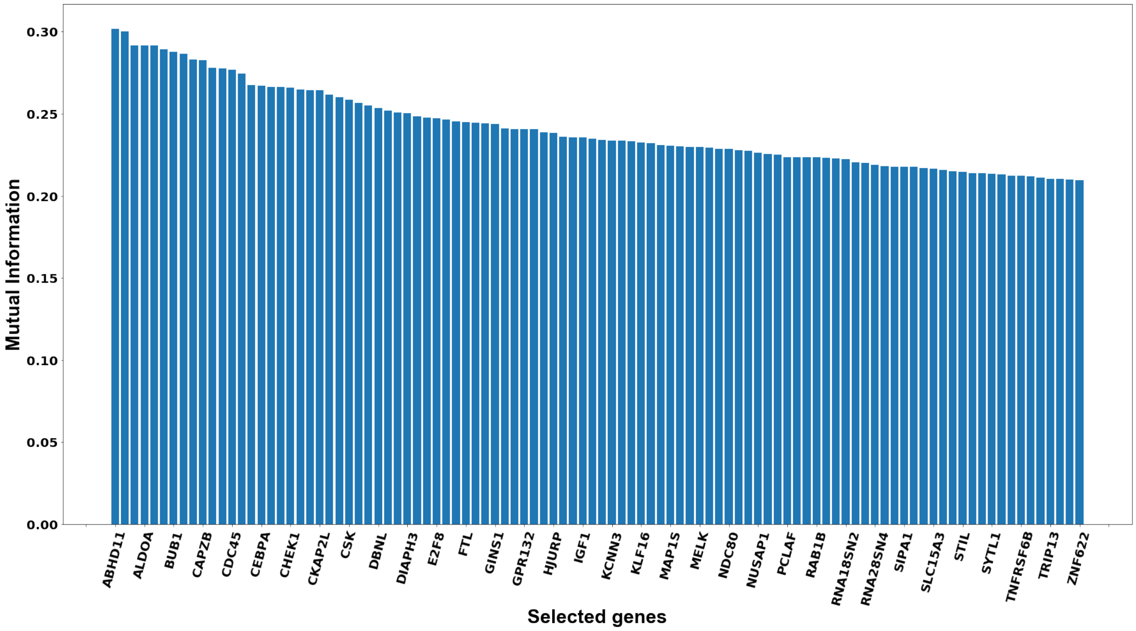


Figure S1 9. Mutual information value of selected features in the classification of a patient into whether he is a COVID-19 or non-COVID-19

Table S-2.91. Results of GO analysis of the selected features in the prediction of a patient whether he is COVID-19 or non-COVID-19

ID	Name	P Value
GO:1903047	mitotic cell cycle process	8.491550203267356E-18
GO:0000278	mitotic cell cycle	9.87694590762079E-18
GO:0007049	cell cycle	9.561717106848294E-17
GO:0022402	cell cycle process	4.648073315565866E-16
GO:0000280	nuclear division	1.32983273267192E-13
GO:0048285	organelle fission	5.458139254549074E-13
GO:0007067	mitotic nuclear division	9.553207259569078E-13
GO:0007059	chromosome segregation	6.864549944935464E-12
GO:0051301	cell division	2.277895729673241E-10
GO:0000819	sister chromatid segregation	1.342018620288827E-8
GO:0098813	nuclear chromosome segregation	2.430652928126043E-8
GO:0000793	condensed chromosome	1.280193571490309E-7
GO:0051726	regulation of cell cycle	1.484258716913475E-7
GO:0000070	mitotic sister chromatid segregation	1.510045231262662E-7
GO:0044772	mitotic cell cycle phase transition	1.847691142512593E-7
GO:0044770	cell cycle phase transition	4.309751925471474E-7
GO:0007346	regulation of mitotic cell cycle	5.270317747189232E-7
GO:0006260	DNA replication	7.91331001711496E-7
GO:0010564	regulation of cell cycle process	1.178796980522714E-6
GO:0005694	chromosome	1.58030920465824E-6
GO:0007088	regulation of mitotic nuclear division	4.170526096025865E-6
GO:0005819	spindle	7.482309225928004E-6
GO:0000940	condensed chromosome outer kinetochore	8.095643396515286E-6
GO:0051783	regulation of nuclear division	2.195173797550326E-5
GO:0006996	organelle organization	2.615551316409474E-5
GO:0007017	microtubule-based process	3.027058234819577E-5
GO:0000775	chromosome, centromeric region	1.154740592005985E-4
GO:0000777	condensed chromosome kinetochore	1.556560761278361E-4
GO:0000779	condensed chromosome, centromeric region	3.057985661241495E-4
GO:0015630	microtubule cytoskeleton	3.662100090529519E-4

x. Phenotype: ICU status of COVID-19 patients (admitted/not)

Table S-2.10. Hundred features used in the prediction of ICU status of COVID-19 patients

ABHD15	CCR5	GRB10	OAS2	SMAD3
ADARB1	CD2	GZMK	OLAH	ST6GALNAC3
ADRB1	CD3G	HSH2D	OTUD3	ST8SIA1
AKAP9	CD4	IGSF9B	P2RY10	SYTL2
APOBEC3D	CD8A	IKZF3	PARP3	TARP
APOBEC3H	CEP41	IL12RB2	PDE4A	TGFBR3
APOL3	COL17A1	IL1R2	PFKFB2	TIFAB
ARG1	CRTAM	KCNA6	PGD	TLR3
ARMC12	CSGALNACT1	KIAA1671	PHF10	TMEM229B
ARMCX6	CYB561	KLRG1	PNPT1	TMIGD3
ATP1B1	CYP4V2	LGR6	PRR5L	TNIP3
BTN3A3	DAAM2	LRFN3	RHAG	TRAF3IP3
C3orf18	DCP1B	LRRC70	ROPN1L	TTC39B
CA4	EOMES	MINDY4B	S100P	TULP4
CACNA2D2	EPCAM	MSL3	S1PR5	TVP23A
CAMK2D	EVL	MYBL1	SAMD3	UBFD1
CCDC136	FGFR2	NCALD	SAMD4A	USPL1
CCDC65	mFLT3LG	NFATC2	SEPTIN8	ZNF510
CCL2	GADD45A	NSUN7	SETDB2	ZNF683
CCL4	GPR68	NUP205	SLC4A8	ZNF829

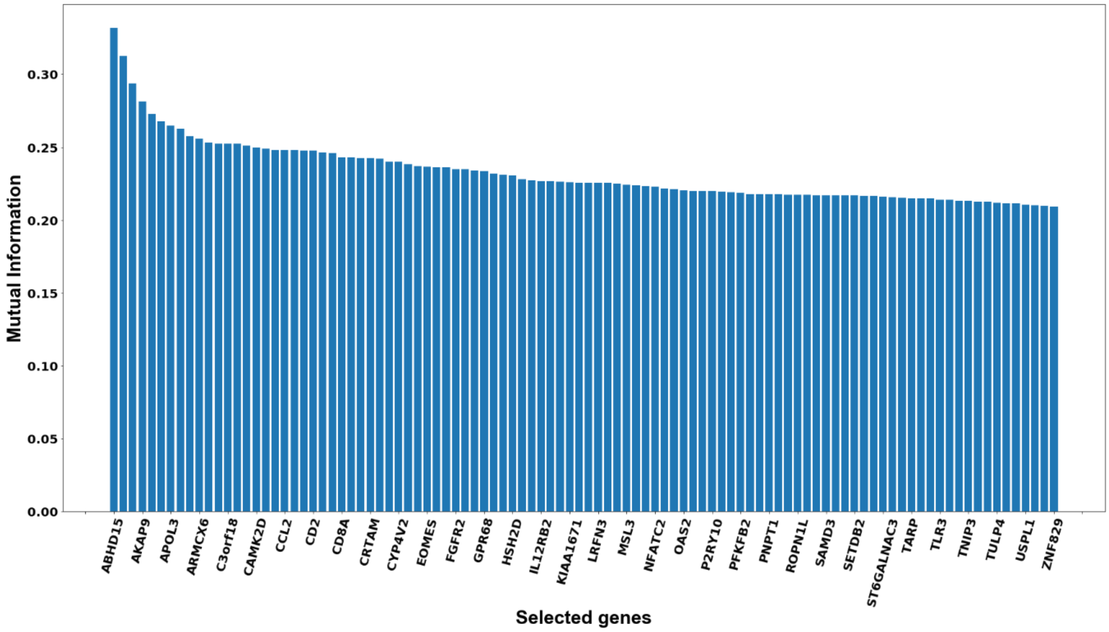


Figure S2-10. Feature importance of the selected features in the prediction of ICU status of COVID-19 patients

Table S-2.101. GO terms related to the selected features in the prediction of ICU status of COVID-19 patients

ID	Name	P Value
GO:0001071	nucleic acid binding transcription factor activity	1.0
GO:0003700	transcription factor activity, sequence-specific DNA binding	1.0
GO:0003677	DNA binding	1.0
GO:0006355	regulation of transcription, DNA-templated	1.0
GO:1903506	regulation of nucleic acid-templated transcription	1.0
GO:2001141	regulation of RNA biosynthetic process	1.0
GO:0051252	regulation of RNA metabolic process	1.0
GO:0006351	transcription, DNA-templated	1.0
GO:0097659	nucleic acid-templated transcription	1.0
GO:0044464	cell part	1.0
GO:0005623	cell	1.0
GO:0032774	RNA biosynthetic process	1.0
GO:2000112	regulation of cellular macromolecule biosynthetic process	1.0
GO:0003674	molecular_function	1.0
GO:0003676	nucleic acid binding1	1.0
GO:0008152	metabolic process	1.0
GO:0019219	regulation of nucleobase-containing compound metabolic process	1.0
GO:0008150	biological_process	1.0
GO:0010556	regulation of macromolecule biosynthetic process	1.0
GO:0005622	intracellular	1.0
GO:0005575	cellular_component	1.0
GO:0046872	metal ion binding	1.0
GO:0044260	cellular macromolecule metabolic process	1.0
GO:0005488	binding	1.0
GO:0006139	nucleobase-containing compound metabolic process	1.0
GO:0043229	intracellular organelle	1.0
GO:0043226	organelle	1.0
GO:0043169	cation binding	1.0
GO:0080090	regulation of primary metabolic process	1.0

تحليل التعلم الآلي الشامل للأنماط الظاهرية لمرضى COVID-19 باستخدام بيانات النسخ

براتيبا جياناثان

قسم هندسة الحاسوب ، جامعة جافنا ، سريلانكا

بريد الكتروني : pratheeba@eng.jfn.ac.lk

المُستَخلَص

تاريخ استلام البحث: 2021/11/08
تاريخ تعديل البحث: 2022/02/08
تاريخ قبول البحث: 2022/02/28

الهدف: تتيح لنا التقنيات المتطورة قياس البيانات الجزيئية البشرية على نطاق واسع. يتم استخدام هذه البيانات على نطاق واسع من قبل الباحثين في العديد من الدراسات وتساعد في تقدم المجال الطبي. تعتبر بيانات النسخ والبروتيوم والمستقلب والإيبيجينوم قليلة من هذه البيانات الجزيئية. تستخدم هذه الدراسة بيانات النسخ لمرضى COVID-19 للكشف عن الجينات غير المنظمة في SARS-COV-2.

الطريقة: تُستخدم جينات مختارة في نماذج التعلم الآلي للتنبؤ بالأنماط الظاهرية المختلفة لهؤلاء المرضى. تمت دراسة عشرة أنماط ظاهرية مختلفة هنا، مثل الوقت منذ البداية، وحالة COVID-19، والاتصال بين العمر وCOVID-19، وحالة المستشفى وحالة وحدة العناية المركزة، باستخدام نماذج التصنيف. علاوة على ذلك، تقارن هذه الدراسة التوصيف الجزيئي لمرضى COVID-19 بأمراض الجهاز التنفسي الأخرى.

النتائج: يُظهر تحليل الأنطولوجيا الجينية على السمات المختارة أنها مرتبطة بشكل كبير بالعدوى الفيروسية. يتم اختيار الميزات باستخدام طريقتين ويتم استخدام الميزات المحددة بشكل فردي في تصنيف المرضى باستخدام ست خوارزميات مختلفة للتعلم الآلي. لكل صفة مختارة ، تتم مقارنة النتائج للعثور على أفضل نموذج تنبؤ.

الخلاصة: على الرغم من عدم وجود أي فروق ذات دلالة إحصائية بين طرق اختيار الميزة، فإن Random Forest وSVM يعملان بشكل جيد للغاية في جميع دراسات النمط الظاهري.

الكلمات المفتاحية: COVID-19، بيانات النسخ، تحليل النمط الظاهري، نماذج التعلم الآلي، أمراض الجهاز التنفسي، الجينات غير المنظمة.