Available online at www.medicinescience.org

ORIGINAL RESEARCH

Medicine Science International Medical Journal

Medicine Science 2021;10(4):1524-33

Classification of healthy controls and Covid-19 cases established on transcriptomic analysis using proposed ensemble model

Zeynep Kucukakcali, Seyma Yasar, Cemil Colak

Inonu University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Malatya, Turkey

Received 09 Septembe 2021; Accepted 04 November 2021
Available online 01.12.2021 with doi: 10.5455/medscience.2021.09.284

Copyright@Author(s) - Available online at www.medicinescience.org
Content of this journal is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Abstract

COVID-19, which is a highly contagious disease, has different symptoms in humans. Therefore, the scientific and genetic status of the virus should be clarified as soon as possible. This study aims to classify COVID-19 and determine the important genes related to the disease by applying the ensemble learning techniques on the public COVID-19 dataset. The data set consists of 579 genes belonging to 32 individuals. While 10 of these people are not COVID-19, 22 are people with COVID-19. In this study Lasso, one of the feature selection methods was used. The ensemble learning methods (Bagging, Boosting, and Stacking) were applied to the public dataset. The performance of the models used was evaluated with accuracy, sensitivity, specificity, positive predictive value, and negative predictive value. Of the constructed ensemble models, the Stacking technique produced the best classification performance compared to the Bagging and Boosting methods. Accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1 score obtained from the Stacking technique were 99.85%, 99.91%, 99.82%, 99.64%, 99.95%, and 99.89respectively. CD22, CD19, C4BPA, ARHGDIB, AICDA, CCR5, CCL7, CCL26, CCL22 and CCL16 genes calculated from the Stacking method were the most important genes related to COVID-19. The genes determined from the model may be determinants for early diagnosis and treatment of the COVID-19 disease.

Keywords: Machine learning, ensemble learning, COVID-19, classification

Introduction

In late 2019, cases of pneumonia of unknown etiology were reported by the World Health Organization (WHO) Country Office China in Wuhan, China's Hubei province. On January 7, 2020, the agent was identified as a new Coronavirus (2019-nCoV) that was not previously detected in humans. Later, the name of the 2019-nCoV disease was named as Coronavirus Disease-2019 (COVID-19), and the virus was named as SARS-CoV-2 due to its close similarity to the Severe Acute Respiratory Syndrome-related Coronavirus (SARS CoV) [1].

SARS-CoV-2 is a disease that can cause serious acute respiratory problems. The disease that begins 2-14 days after exposure to the virus, fever or chills, cough, shortness of breath or a feeling of pressure in the chest, tiredness, muscle or body aches, headache, decreased sense of taste or smell, sore throat, nasal congestion or runny nose, nausea or with symptoms such as vomiting, diarrhea as it may occur with symptoms such as or without symptoms.

The symptom list is expanding day by day. In more severe cases, causes pneumonia, Acute Respiratory Distress Syndrome (ARDS), multiple organ failure, and death [2]. In a study describing 138 patients hospitalized with COVID-19 pneumonia in Wuhan, fever was observed in 99% of the patients. Also, fatigue was observed in 70% of the patients, dry cough in 59%, anorexia in 40%, myalgia in 35%, dyspnea in 31%, and cough with sputum in 27%. The COVID-19 pandemic, which emerged in Wuhan, People's Republic of China, in December has started to spread all over the world since March, and as of December 2020, 65 million people have been reported to get sick in the world. It caused a total of 1,513,179 deaths on six continents around the World. The epidemic level of the disease has led to a strain on health resources in many countries, and this situation has made it necessary to evaluate all methods that can guide diagnosis and treatment [3]. Although many cases have been reported for COVID-19 infection, what needs to be done was to clarify the scientific and genetic status of this

***Corresponding Author:** Zeynep Kucukakcalı, Inonu University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Malatya, Turkey
E-mail: zeynep.tunc@inonu.edu.t

virus as soon as possible. It was necessary to examine and reveal the genomic structure of the virus, gene regions, protein binding points, attachment, and neutralizing structures, etc. For this, first of all, it was necessary to isolate the virus, to perform simple gene regions and sequencing of the whole genome at advanced level, and then to perform bioinformatics analysis. Therefore, in many countries, health organizations, universities or private sectors have made efforts to reveal this virological and epidemiological information [4].

Methods that enable inference from data stacks and generation of information are included in the data mining discipline. Briefly, data mining is defined as the process of generating information by discovering patterns in data. In data mining, information can be extracted from data stacks automatically. Data mining includes a combination of techniques from different disciplines such as database technology, statistics, machine learning, pattern recognition, neural networks, data visualization, and spatial data analysis [5]. Machine learning, one of these techniques, is a subfield of artificial intelligence that aims to make predictions about new data when they are exposed to new data by performing data-based learning. Machine learning systems aim to completely eliminate the need for human intuition or to gain the ability to make decisions with human-machine cooperation. Machine learning methods are increasingly used in the diagnosis and prediction of diseases in the healthcare field. Machine learning methods in the prediction of diseases generally perform the classification process [6]. Although machine learning methods mostly have high accuracy performance in classification processes, they cannot give the desired performance level in some data sets. There are many reasons why the desired performance level is not achieved. Reasons such as erroneous imputation in place of observations in the data set, class imbalance in the data set, presence of noisy data, an insufficient sample size cause serious losses in the performance of machine learning methods. Different solutions have been proposed to improve these performance losses during learning. One of these methods is community learning methods. Unlike the classification of a single machine learning algorithm, ensemble learning methods classify the data of multiple machine learning algorithms separately, providing a common classification result from the estimates of each classifier. Thus, according to the prediction results of a machine learning method, the common prediction results obtained from more than one machine learning method offer more accurate, more reliable, and higher performance [7].

The working principle of ensemble learning methods is based on the principle that multiple classifiers can perform classification with higher accuracy than a single classifier predicts. Ensemble learning methods have found a wide range of applications in recent years with their successful results. Commonly used ensemble learning methods have been successfully applied in the diagnosis and diagnosis of many diseases [8].

This study aims to classify the COVID-19 transmission status by applying the ensemble learning method, which is an important sub-field of machine learning, on the gene data set of patients with and without public COVID-19, and to determine the important genes that cause the disease.

Materials and Methods

Dataset

In the study, the ensemble learning method, which is an important sub-field of machine learning, was applied to the gene data set of patients with and without open access covid19 . 579 genes are belonging to 32 individuals in the data set used. While 10 (%31,3) of these people are not COVID-19, 22 (%68.7) are people with COVID-19. Daily transcriptomic profiling was performed on whole blood collected from COVID-19 cases. Whole blood was collected in Tempus Blood RNA tubes and RNA was extracted from whole blood using the Tempus Spin RNA Isolation Kit. It was processed and analyzed as above for healthy control [9].

Feature selection

Variable Selection is an important step in a predictive modeling project. This is also called 'Feature Selection'. One of the most important steps in building a statistical model is deciding what data to include. High efficiency can be achieved by identifying the most useful properties of a data set before working with very large data sets and models with high computational costs. Feature selection is the process of defining features in a data set that has an impact on the dependent variable The high dimensionality of the explanatory variables can cause both high computation time and the risk of over-learning of the data. Moreover, it is difficult to interpret models with many features. Ideally, important features should be selected before performing statistical modeling. The methods used in feature selection are generally grouped into three groups as filter methods based on statistical information only, wrapper methods that perform search operations on properties, and embedded methods based on finding the best divisor criterion [10]. In this study, the Lasso feature selection method was used as the feature selection method.

Most machine learning and data mining methods may not be effective for high dimensional data. For this reason, more effective results can be obtained with these methods when the dimensionality is reduced [11]. Gene expression data sets are quite large. Modeling analyzes with gene expression data sets take a long time due to their large size, and therefore these data sets may lead to computational inefficiency in the analysis. The high dimensionality problem can cause the performance of the model to decrease. Also, a large number of genes in gene expression data sets can cause a classification algorithm to fit the training examples and to generalize new samples poorly. To solve these problems, Lasso, one of the feature selection methods, was used in this study. The LASSO method was first used in 1996 by Robert Tibshirani. The LASSO method puts a constraint on the sum of the absolute values of the model parameters, the sum must be less than a fixed value (upper limit). To do this, the method applies a throttling process in which it penalizes the coefficients of the regression variables, some of which drop to zero.

During the feature selection process, variables that still have a non-zero coefficient after narrowing are selected for the model. The purpose of this process is to minimize the guessing error. It is especially useful when there are few observations and a large number of variables in the data set. Also, LASSO helps increase the interpretability of the model by eliminating irrelevant variables

that are not associated with the response variable, thus eliminating the problem of over-learning [12].

Classification Methods

Simple Logistic Regression Analysis

Logistic regression is a method used to determine the cause-effect relationship with the independent variables when the dependent variable is observed as binary or multiple categorically. Logistic regression, which is a method of determining the probability of the expected value of the dependent variable according to the values of the independent variables, can also be used to classify data based on the effects of the independent variables. The effects of independent variables on the dependent variable are obtained as probabilities and the risk factors are determined as probabilities. In the applications of logistic regression models in the field of medicine, independent variables are risk variables or variables that determine the occurrence of a disease or not. In short, logistic regression is a regression method that helps to assign and classify the expected value of the dependent variable according to the independent variables [13].

Artificial neural networks

Artificial Neural Networks (ANNs) are computer systems developed to directly realize the features of the human brain such as learning, generating, creating, and discovering new information without any assistance. ANN are physically cellular systems that receive, store, and use experimental information. ANN can provide nonlinear modeling without any prior knowledge between input and output variables, without any assumptions [14]. ANNs are successfully applied in many different fields due to their learning ability, adaptability to different problems easily, needing less information after the learning process, ability to make generalizations, fast processing due to their parallel structures, and ability to solve difficult mathematical models very quickly. Artificial neural networks are a successful method in solving many daily life problems such as classification, modeling, and prediction [15, 16].

Support Vector Machine

SVM methods have taken their place among the popular algorithms of recent years. SVM, developed by Vapnik Chervonenkis, is a machine learning model that is used in regression problems in addition to classification problems. SVM uses a technique called kernel trick to transform data. Kernel trick methods determine the optimal boundary among possible outcomes based on data transformation models. That is, kernel trick methods first perform complex data transformations and then determine how these data will be separated based on defined tags or results. The main purpose of SVM is to obtain a hyperplane that will distinguish the classes belonging to the target variable in the most appropriate way. Two situations can be encountered in SVM. These are the cases where the data are in a structure that can be separated linearly or in a structure that cannot be separated linearly. Whether classification or regression with SVM will be done, kernel functions are used to solve nonlinear situations[17]. The nonlinear SVM method aims to obtain nonlinear classifiers by applying kernel functions with different structures to the maximum margin hyperplane. The algorithm obtained is similar to linear SVM. However, every inner product is replaced by a nonlinear kernel function In summary, using kernel functions, instead of calculating the product values of all values over and over again, it is provided to find the value in the property space by directly substituting the value in the kernel function. The algorithm thus obtained allows the maximum spacing hyperplane to be placed in the transformed sample space. In this way, the problem of dealing with a high dimensional property space is eliminated. Another advantage of the kernel functions is that after the function is set up and the values are found for any training example during the training phase, it is much easier to calculate the mold values for other samples as they are completely ready except for the training example [18].

Random Forest (RF)

RF method is a classification method developed by Leo Breiman and Adele Cutler and includes the voting method. It consists of many decision trees coming together and the winning class is determined by voting by individual trees. The decision trees in the forest are independent of each other and are created from the samples taken from the data set with the bootstrap technique [19]. The RF method is a forest classifier consisting of many decision trees, and classification or regression trees can be established with this method. If the "class variable" in the data set is categorical, classification trees are established, and if continuous, regression trees are established [6]. Determining branching criteria and choosing an appropriate pruning method in the RF method is a very important issue. The Gini index method is used to determine the branching criteria of the random forest classifier. The Gini index measures the weakness of class features. In the RF method, as in other classification methods, some parameters must be determined by the user. These parameters are the number of instances to be used in each node and the number of trees to be created, which are required to build the tree structure. In other words, during a classification process, the random forest is created from specified K trees by the user [20].

Decision Trees

Decision trees, one of the popular and powerful methods of information discovery and data mining, are a hierarchical and sequential method of displaying the rules within the data. Decision trees are a visual modeling method that displays the existing information mass more understandably and presents decision options and probabilistic situations in a certain order. In summary, it can be said that decision trees represent a hierarchical model that includes decisions and their results. Thanks to its easy-to-understand graphical structure and rules, it is widely used in many areas. The decision trees model, which is among the classification models, is a model with predictive value. Decision trees ask questions starting from the first stage to the final decision options and form their structure with the answers they receive to these questions, and rules (if-then rule) can be written with this tree structure [21].

Ensemble Learning Method

Machine learning methods have reached a very common area of use in artificial intelligence and applied sciences recently. This success of machine learning methods depends on the algorithms

used to achieve successful predictions and perform classification operations with high accuracy. While machine learning methods provide high accuracy performance in many complex data sets with powerful algorithms, they perform classifications with high variance and low accuracy values in some data sets. Different methods have been proposed to prevent this performance loss in classification and estimation processes. One of these methods is ensemble learning methods. When ensemble learning methods first emerged, they were applied to reduce the high variance in machine learning methods and to increase the obtained accuracy rate. However, he also achieved successful results in solving problems such as feature selection, missing features, error correction, confidence interval estimation, unbalanced data, and classes, which are frequently encountered in machine learning methods [7]. The basic working logic of the ensemble learning method is based on the principle that many decisions will contain healthier and more accurate results than a single decision. The likelihood that an expert's decision will be wrong is more likely than the joint decision made by more than one expert to be wrong. In other words, a decision taken by more than one expert will be more reliable and more accurate than the decision of an expert [7]. Based on this, ensemble learning methods are generally learning methods with higher accuracy and performance obtained by combining the predictions of more than one machine learning method, rather than the performance obtained as a result of a single machine learning method. An important factor affecting the classification performance in the ensemble learning method is the selection of the joining method appropriate to the data. In studies, attention should be paid to the selection of the appropriate joining technique for classifiers. There is different ensemble learning methods according to the joining techniques, the sample selection for the training data set, and the process steps. These methods are the bagging ensemble learning method, the boosting ensemble learning method, and the stacking ensemble learning method [7, 22].

Bagging (Bootstrap aggregating)

The bagging ensemble learning method, which is referred to as Bootstrap aggregating, is based on the bootstrap sampling method. The bagging method is a method that aims to retrain the basic learner by creating new training data sets by random selection with substitution from a known training data set. In summary, the main purpose of the Bagging method is to obtain new data sets randomly using training data and to increase the success of classification by creating differences. In the bagging method, first, the data set is divided into training and test data. One or more new training sets consisting of n samples is obtained by random selection method by substituting it from the training set containing N samples. Each basic classifier in the community obtained by the bagging method is trained with training sets containing different examples obtained in this way. Finally, the result of each major classifier is combined with the majority vote [23]. In this study, the decision trees method is used as a classifier for the Bagging community learning method.

Boosting

The boosting method is an ensemble learning method that was introduced by Schapire in 1990 and developed until the 2000s. The term "boost" refers to a family of algorithms that transform poor learning methods into powerful learning techniques. Boosting is an ensemble method to improve the model estimates of any learning algorithm, and unlike the Bagging method, the estimators are created sequentially, although they are not independent of each other. This method aims to combine weak estimators to obtain strong estimator (s). Models are created by assigning weight to observations. In the Boosting method, as in the bagging method, N training sets are created. In this method, models with low variance and bias are obtained by the presence of both the bagging method and the assignment of weight to the observations [24]. In this study, the decision trees method was used as a classifier for the Boosting ensemble learning method.

Stacking

The stacking method is a simple ensemble learning technique that creates a meta classifier by combining two or more basic multiple classification models. It is an ensemble model that is trained by combining the estimates of the classification models used. Predictions made from models created by the basic classifier are used as input for each ordered layer and are combined to create a new set of predictions. In the stacking method, basic classification models are trained on the original training data set and then created based on the outputs (estimates) of the basic classification models in the meta-classifier ensemble. The meta-classifier performs classification by training on the predicted class labels [25]. In this study, logistic regression, artificial neural networks, Random Forest, and support vector machines method were used as classifiers for the Stacking ensemble learning method. The Random Forest method was used as a meta classifier.

Performance evaluation criteria

1000 repetitive bootstrap validation method was used for model validity. Bootstrap validation performs validation after bootstrapping a sampling of the training dataset to estimate the statistical performance of a learning operator. It is mainly used to predict how accurately a model will perform in practice. The Bootstrap Validation method has 2 sub-stages: training and test sub-process. The training sub-process is used to train a model. The trained model is then applied in the test sub-process. The performance of the model is also measured during the testing phase [26].

The classification matrix for the calculation of performance metrics is given in Table 1.

Data analysis

Quantitative data are summarized by mean ± standard deviation and median (minimum-maximum). Normal distribution was evaluated with the Shapiro Wilks test. In terms of input variables, the existence of a statistically significant difference and the relationship between the categories of the output variable, "healthy control (hc)" and "ncov" groups, were examined using independent sample t-test and Mann-Whitney U test. $p<0.05$ values were considered statistically significant. RapidMiner Studio software and IBM SPSS Statistics 26.0 for the Windows package program were used in all modeling and analysis [26].

Table 1. Confusion matrix for calculating performance metrics

		Real		
		Positive	**Negative**	**Total**
Predicted	Positive	True positive (TP)	False positive (FP)	TP+FP
	Negative	False negative (FN)	True negative (TN)	FN+TN
	Total	**TP+FN**	**FP+TN**	**TP+TN+FP+FN**

Accuracy = (TP+TN)/(TP+TN+FP+FN) Sensitivity = TP/(TP+FN) Specificity = TN/(FP+TN) Positive predictive value = TP/(TP+FP)
Negative predictive value =TN/(TN+FN) F1-score = (2*TP)/(2*TP+FP+FN)

Results

A 42 genes remained in the data set obtained by applying the LASSO feature selection method to the data set consisting of 579 genes. These genes obtained as a result of the Lasso variable selection are given in Table 2.

Descriptive statistics related to the target variable examined in this study are presented in Table 3. There is a statistically significant difference between the dependent variable classes in terms of other variables.

Classification matrices of Bagging, Boosting, and Stacking models, which are among the ensemble learning methods used to classify the dataset in this study, are given in Table 4 below.

The values for the metrics of the classification performance of Bagging, Boosting, and Stacking models are given in Table 5. Accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1 score obtained from the Bagging model were 97.70%, 98.34%, 96.39%, 98.23%, 96.60%, and 98.29% respectively. Accuracy, sensitivity, specificity, positive predictive value, negative predictive, and F1 score value obtained from the Boosting model were 97.71%, 98.35%, 96.39%, 98.23%, 96.63%, and 98.29% respectively. Accuracy, sensitivity, specificity, positive predictive value, negative predictive value and F1 score obtained from the Stacking model were 99.86%, 99.82%, 99.92%, 99.95%, 99.64%, and 99.89% respectively.

In Figure 1, the values of performance criteria obtained from Bagging, Boosting, and Stacking models are plotted.

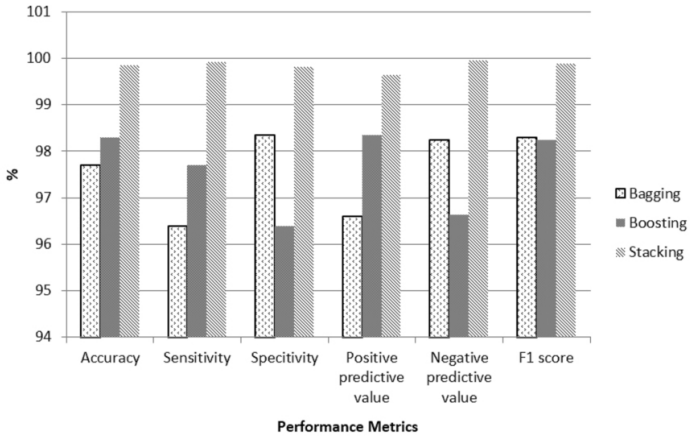


Figure 1. Graph of values for performance criteria obtained from Bagging, Boosting, and Stacking models

Table 6 and Figure 2 show the significance levels of genes that are important for COVID-19.

Table 2. Genes obtained as a result of the Lasso variable selection

Genes					
ABL1	C4BPA	CCL13	CCR10	ATG10	CCL11
AICDA	C6	CCL15	CCR5	BCL2L11	CCL2
AIRE	C7	CCL16	CCR6	BID	CCL24
ARHGDIB	C8B	CCL22	CD19	BST2	CCND3
BCL2	C9	CCL26	CD1A	C1QBP	CCR8
BTK	CARD9	CCL7	CD209	C1S	CCRL2
C1R	CCBP2	CCR1	CD22	C2	CD247

Table 3. Descriptive statistics for Quantitative Input variables

Genes	Groups				p value
	hc		nCov		
	Mean± Standard deviation	Median (min-max)	Mean± Standard deviation	Median (min-max)	
ABL1	6.36±0.18	-	6.82±0.22	-	**<0.001***
AIRE	4.11±0.32	-	3.33±0.31	-	**<0.001***
ARHGDIB	12.68±0.09	-	12.03±0.14	-	**<0.001***
BCL2	8.4±0.23	-	7.66±0.33	-	**<0.001***
BTK	8.27±0.17	-	7.82±0.28	-	**<0.001***
C1R	3.79±0.12	-	3.37±0.33	-	**0.001***
C6	4.83±0.19	-	4.17±0.44	-	**<0.001***
C8B	3.77±0.16	-	3.12±0.27	-	**<0.001***
C9	3.74±0.14	-	3.22±0.26	-	**<0.001***
CARD9	5.78±0.29	-	5.22±0.38	-	**<0.001***
CCBP2	4.85±0.43	-	3.51±0.48	-	**<0.001***
CCL13	3.88±0.25	-	3.31±0.3	-	**<0.001***
CCL15	3.9±0.2	-	3.19±0.25	-	**<0.001***
CCL16	3.74±0.14	-	3.12±0.27	-	**<0.001***
CCL22	3.82±0.17	-	3.13±0.24	-	**<0.001***
CCL26	3.74±0.14	-	3.16±0.22	-	**<0.001***
CCL7	3.74±0.14	-	3.12±0.27	-	**<0.001***
CCR1	8.39±0.51	-	10.01±1.01	-	**<0.001***
CCR10	3.78±0.11	-	3.18±0.31	-	**<0.001***
CCR5	7.34±0.44	-	8.58±0.33	-	**<0.001***
CD19	7.91±0.35	-	6.19±0.28	-	**<0.001***
CD1A	4.51±0.33	-	3.43±0.31	-	**<0.001***
CD209	4.08±0.29	-	3.35±0.32	-	**<0.001***
CD22	8.54±0.35	-	6.78±0.33	-	**<0.001***
ATG10	5.11±0.15	-	5.6±0.35	-	**<0.001***
BCL2L11	7.07±0.2	-	6.66±0.34	-	**0.001***
BID	5.55±0.2	-	5.08±0.39	-	**0.001***
C2	4.83±0.51	-	6.11±0.91	-	**<0.001***
CCL24	3.92±0.21	-	3.46±0.38	-	**0.001***
CCND3	9.16±0.12	-	9.49±0.31	-	**0.003***
CCR8	4.18±0.39	-	3.66±0.5	-	**0.007***
CCRL2	5.97±0.3	-	6.59±0.5	-	**0.001***
CD247	9.66±0.32	-	10.17±0.27	-	**<0.001***
AICDA		3.94 (3.73-4.88)		3.13 (2.45-3.56)	**<0.001****
C4BPA		4.72 (3.74-6.35)		3.13 (2.45-3.56)	**<0.001****
C7		3.79 (3.47-3.88)		3.36 (2.53-4.71)	**<0.001****
CCR6		7.49 (7.17-7.93)		6.45 (6.05-7.37)	**<0.001****
BST2		7.62 (7.18-8.24)		8.28 (7.63-9.84)	**0.001****
C1QBP		9.3 (9.14-9.45)		9.61 (8.89-9.87)	**0.002****
C1S		3.89 (3.47-4.17)		3.39 (2.53-4.97)	**<0.001****
CCL11		4.29 (3.97-4.93)		3.65 (3.23-4.79)	**0.001****
CCL2		4.02 (3.73-4.78)		3.39 (2.45-6.03)	**0.046****

Table 4. Classification matrices of Bagging, Boosting, and Stacking models

Classification Matrix of the Bagging Model			
Prediction	**Reference**		
	nCov	hc	Total
nCov	7287	131	7418
hc	123	3499	3622
Total	7410	3630	11040
Classification Matrix of the Boosting Model			
Prediction	**Reference**		
	nCov	**hc**	**Total**
nCov	7287	131	7418
hc	123	3499	3622
Total	7410	3630	11040
Classification Matrix of the Stacking Model			
Prediction	**Reference**		
	nCov	**hc**	**Total**
nCov	7397	3	7400
hc	13	3627	3640
Total	7410	3630	11040

Table 5. Values for the metrics of the classification performance of Bagging, Boosting, and Stacking models

Models	**Metric**	**Value (%)**
Bagging	Accuracy	97.69
	Sensitivity	96.39
	Specificity	98.34
	Positive predictive value	96.60
	Negative predictive value	98.23
	F1 score	98.29
Boosting	Accuracy	97.70
	Sensitivity	96.39
	Specificity	98.35
	Positive predictive value	96.63
	Negative predictive value	98.23
	F1 score	98.29
Stacking	Accuracy	99.85
	Sensitivity	99.91
	Specificity	99.82
	Positive predictive value	99.64
	Negative predictive value	99.95
	F1 score	99.89

Tablo 6. Variable importance values for the Stacking ensemble learning model

Genes	IGR	Genes	IGR	Genes	IGR
CD22	0.90	CCL15	0.71	CCL13	0.47
CD19	0.90	BCL2	0.71	CARD9	0.47
C4BPA	0.90	CD209	0.65	CCL24	0.42
ARHGDIB	0.90	C9	0.65	CCL11	0.42
AICDA	0.90	CCR6	0.58	C1S	0.42
CCR5	0.74	CCR10	0.58	C1QBP	0.42
CCL7	0.74	C7	0.58	CCR8	0.38
CCL26	0.74	C6	0.58	CCL2	0.38
CCL22	0.74	C1R	0.52	BID	0.38
CCL16	0.74	ABL1	0.52	BCL2L11	0.38
CCBP2	0.74	CCR1	0.48	CCND3	0.37
C8B	0.74	BTK	0.48	BST2	0.31
AIRE	0.74	C2	0.47	CD247	0.31
CD1A	0.71	ATG10	0.47	CCRL2	0.31

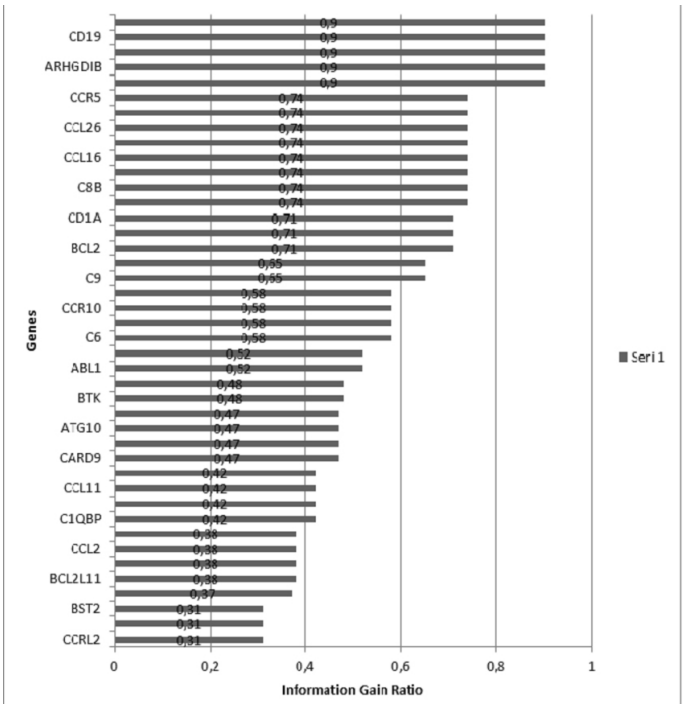


Figure 2. The graphic of variable importance values for the Stacking ensemble learning model

Discussion

SARS-CoV-2 is still a virus that threatens the whole world and can spread very quickly. The COVID-19 it causes can lead to serious pathologies in many tissues, especially the respiratory tract, and can cause common systemic complications that can progress to multiple organ damage. SARS-CoV-2 can be asymptomatic, or it can result in serious conditions such as ARDS, respiratory failure, diffuse thromboembolism, and even death. Although the majority of patients can be cured with symptomatic treatment and intensive care support, no specific treatment has yet been found. High transmission rate and mortality rates, lack of vaccine, unclear treatment protocols and side effects, and the inability to know the risk of recurrence, prognosis, and long-term consequences of the disease increase concerns and fears about COVID-19. The fight against the disease has led to a strain on health resources in many countries, and this situation has made it necessary to evaluate all methods that can guide diagnosis and treatment [3]. Although many cases have been reported for COVID-19 infection, what needs to be done is to clarify the scientific and genetic status of this virus as soon as possible.

Genomics, which processes and stores its outputs through information technologies, is a science developed by advances in automation and bioinformatics [27]. Genomics is a discipline suitable for research to evaluate almost any subject and situation. With good fiction and a correct comparison, research can be done in almost every field of medicine (Oncology, Pharmacology, Immunology, Biochemistry, Microbiology, etc.). Through comparative studies, studies such as cancer and prognosis prediction, drug response prediction, and individualized drug development, the nature of the immune response and prediction of transplantation outcome can be conducted [28].

Machine learning methods are one of the technologies that are widely used in the diagnosis of diseases and clinical decision support systems in recent years and have a wide application area. Machine learning, which has a wide application area in the field of health, constitutes the basic infrastructure of the applications of identifying patterns in the detection of genetic diseases, early diagnosis of cancer diseases, and medical imaging [29].

Ensemble learning methods, one of the machine learning methods,

train more than one classifier by modeling instead of training the training data set with a classifier. In the aggregation stage, the estimates of the classifiers are combined after the estimates of each classifier are obtained. Since the error value in predictions decreases, higher performance can be obtained in ensemble learning compared to basic classifiers. Ensemble learning methods have found a wide range of applications in recent years with their successful results [7].

In this study, ensemble learning models, one of the machine learning methods, were applied to the gene data set of patients with and without open access COVID-19. According to the results of 3 different models used, the method for the best classification performance is the Stacking method. Accuracy, sensitivity, specificity, positive predictive value, and negative predictive value and F1 score obtained from this model were 99.85%, 99.91%, 99.82%, 99.64%, 99.95, and 99.89% respectively. CD22, CD19, C4BPA, ARHGDIB, AICDA, CCR5, CCL7, CCL26, CCL22 and CCL16 genes calculated from the best performing Stacking method can be used as biomarkers for COVID-19.

There is a lot of studies to reveal the relationships of the COVID-19 epidemic with genes. In a study following age-dependent gene expression response in blood upon infection with an influenza virus ex-vivo26, revealed 13 genes (CSF3R, S100A9, S100A8, FCGR2A, CR1, CLEC7A, ARHGDIB, CEACAM6, LILRB3, LILRA5, LILRA1, NCF4, TLR1, LY96) that were associated with age-dependent response and were upregulated in Covid-19 patients [30]. CCR5 is known to be responsible for the induction of inflammation in a wide range of infectious diseases and recruit leukocytes towards inflammation sites [31]. In another study, although highlighted a significant association of CCR5 Δ32 variant with susceptibility and mortality from SARS-CoV-2 infection, it has set the stage for in-depth analysis by factoring in various other aspects [32]. In another study points to CCR5 as a promising target for the treatment of COVID-19, but requires validation in additional large cohorts [33]. In another study, CCL7 was associated with increased viral load, loss of lung function, lung injury, and a fatal outcome [34]. CCL26 gene is the SARS-CoV-2 host response genes. In a study, the CCL22 gene was significantly down-regulated after 24 hours of HCQ (hydroxychloroquine) treatment as compared to the untreated condition [35].

Conclusion

As a result, the proposed community learning methods achieved very high performances in the classification of genes calculated from the best performing Stacking method can be used as biomarkers for COVID-19. Genes determined with these results may be determinant in early diagnosis and treatment of the disease.

Conflict of interests
The authors declare that they have no competing interests.

Financial Disclosure
This software was supported by Inonu University Scientific Research Projects Coordination Unit within the scope of TOA-2020-2204 numbered research project. As a project team, we would like to thank Inonu University Scientific Research Projects Coordination Unit for this support.

References

1. Lu R, Zhao X, Li J, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet. 2020;395:565-74.
2. Rothan HA, Byrareddy SN. The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak. J Autoimmun. 2020:102433:1-4.
3. Vetrugno L, Baciarello M, Bignami E, et al. The "pandemic" increase in lung ultrasound use in response to Covid-19: can we complement computed tomography findings? A narrative review. J Ultrasound. 2020;12:1-11.
4. Timurkan MÖ, Aydın H. Transmission and replication dynamics of SARS CoV-2. Eurasian J Vet Sci.2020;17-22.
5. Chung HM, Gray P. Data mining. Manag Inf Syst. 1999;16:11-6.
6. Akman M, Genç Y, Ankarali H. Random forests yöntemi ve sağlık alanında bir uygulama. Turkiye Klinikleri J Biostat. 2011;3:36-48.
7. Zhang C, Ma Y. Ensemble machine learning: methods and applications: Springer; 2012.
8. Hsieh S-L, Hsieh S-H, Cheng P-H, et al. Design ensemble machine learning model for breast cancer diagnosis. J Med Syst. 2012;36:2841-7.
9. https://www.ebi.ac.uk/arrayexpress/search.html?query=COVID-19+.
10. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007;23:2507-17.
11. Van Der Maaten L, Postma E, & Van den Herik J. Dimensionality reduction: a comparative. J Mach Learn Res. 2009;10:13.
12. Fonti V. Research Paper in Business Analytics: Feature Selection with LASSO. Amsterdam: VU Amsterdam. 2017.
13. Sperandei S. Understanding logistic regression analysis. Biochem Med. 2014;24:12-8.
14. Haykin S. Neural Networks, a comprehensive foundation, Prentice-Hall Inc. Upper Saddle River, New Jersey. 1999;7458:161-75.
15. Johnson KW, Torres Soto J, Glicksberg BS, at al. Artificial intelligence in cardiology. J Am Coll Cardiol. 2018;71:2668-79.
16. Kaya MO. Performance Evaluation of Multilayer Perceptron Artificial Neural Network Model in the Classification of Heart Failure. The J Cog Syst. 2021;6:35-8.
17. Tunç Z, Çolak C, Özdemir R. Classification of Hydrocephalus Disease and Determination of Related Factors by Machine Learning Method. Journal of Inonu University Health Sciences. 2018:14-20.
18. Kecman V. Learning and soft computing: support vector machines, neural networks, and fuzzy logic models: MIT press; 2001.
19. Breiman L. Random forests. Machine learning. 2001;45:5-32.
20. Pal M. Random forest classifier for remote sensing classification. Int J Remote Sens. 2005;26:217-22.
21. Lior R. Data mining with decision trees: theory and applications: World Scientific; 2014.
22. Rokach L. Pattern classification using ensemble methods: World Scientific; 2010.
23. Ferreira AJ, Figueiredo MA. Boosting algorithms: A review of methods, theory, and applications. Ensemble machine learning: Springer; 2012;35-85.
24. Le T, Le Son H, Vo MT, et al. A cluster-based boosting algorithm for bankruptcy prediction in a highly imbalanced dataset. Symmetry, 2018;10:250.
25. Divina F, Gilson A, Goméz-Vela F, et al. Stacking ensemble learning for short-term electricity consumption forecasting. Energies. 2018;11:949.
26. Mierswa I, Klinkenberg R. RapidMiner Studio (9.2)[Data science, machine learning, predictive analytics]. 2018.
27. Martin DB, Nelson PS. From genomics to proteomics: techniques and applications in cancer research. Trends Cell Biol. 2001;11:60-5.
28. Del Boccio P, Urbani A. Homo sapiens proteomics: clinical perspectives. Ann Ist Super Sanita. 2005;41:479-82.
29. Witten IH, Frank E. Data mining: practical machine learning tools and techniques with Java implementations. Acm Sigmod Record. 2002;31:76-7.

30. Johri S, Jain D, Gupta I. Integrated analysis of bulk multi omic and single-cell sequencing data confirms the molecular origin of hemodynamic changes in Covid-19 infection explaining coagulopathy and higher geriatric mortality. medRxiv. 2020.

31. Klein RS. A moving target: the multiple roles of CCR5 in infectious diseases. The University of Chicago Press; 2008.

32. Panda AK, Padhi A, Prusty BAK. CCR5 Δ32 minorallele is associated with susceptibility to SARS-CoV-2 infection and death: An epidemiological investigation. Clin Chim Acta; 2020.

33. Gómez J, Cuesta-Llavona E, Albaiceta GM, et al. The CCR5-delta32 variant might explain part of the association between COVID-19 and the chemokine-receptor gene cluster. medRxiv. 2020.

34. Vaninov N. In the eye of the COVID-19 cytokine storm. Nat Rev Immunol. 2020;20:277.

35. Corley MJ, Sugai C, Schotsaert M, et al. Comparative in vitro transcriptomic analyses of COVID-19 candidate therapy hydroxychloroquine suggest limited immunomodulatory evidence of SARS-CoV-2 host response genes. bioRxiv. 2020.