RESEARCH ARTICLE

Ipek Balikci Cicek[1]
Mehmet Onur Kaya[2]
Cemil Colak[1]

[1]Inonu University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Malatya, Turkey
[2] Firat University, Faculty of Medicine, Department of Basic Medical Sciences Head of Department of Biostatistics 23119 Elazig,, Turkey

Corresponding Author:
Ipek Balikci Cicek
mail: ipek.balikci@inonu.edu.tr

Received: 14.07.2021
Acceptance: 04.11.2021
DOI: 10.18521/ktd.958555

Konuralp Medical Journal
e-ISSN1309–3878
konuralptipdergi@duzce.edu.tr
konuralptipdergisi@gmail.com
www.konuralptipdergi.duzce.edu.tr

Assessment of COVID-19-Related Genes Through Associative Classification Techniques

ABSTRACT

Objective: This study aims to classify COVID-19 by applying the associative classification method on the gene data set consisting of open access COVID-19 negative and positive patients and revealing the disease relationship with these genes by identifying the genes that cause COVID-19.

Methods: In the study, an associative classification model was applied to the gene data set of patients with and without open access COVID-19. In this open-access data set used, 15979 genes are belonging to 234 individuals. Out of 234 people, 141 (60.3%) were COVID-19 negative and 93 (39.7%) were COVID-19 positives. In this study, LASSO, one of the feature selection methods, was performed to choose the relevant predictors. The models' performance was evaluated with accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1-score.

Results: According to the study findings, the performance metrics from the associative classification model were accuracy of 92.70%, balanced accuracy of 91.80%, the sensitivity of 87.10%, the specificity of 96.50%, the positive predictive value of 94.20%, the negative predictive value of 91.90%, and F1-score of 90.50%.

Conclusions: The proposed associative classification model achieved very high performances in classifying COVID-19. The extracted association rules related to the genes can help diagnose and treat the disease.

Keywords: Association Rules, Associative Classification, COVID-19, Genes, Classification.

COVID-19 ile İlgili Genlerin İlişkisel Sınıflandırma Teknikleriyle Değerlendirilmesi

ÖZET

Amaç: Bu çalışma, açık erişimli COVID-19 negatif ve pozitif hastalardan oluşan gen veri seti üzerinde ilişkisel sınıflandırma yöntemini uygulayarak COVID-19'u sınıflandırmayı ve COVID-19'a neden olan genleri tanımlayarak bu genlerle hastalık ilişkisini ortaya çıkarmayı amaçlamaktadır.

Gereç ve Yöntem: Bu çalışmada açık erişimli COVID-19 olan ve olmayan hastaların gen veri setine ilişkisel sınıflandırma yöntemi uygulandı. Kullanılan açık erişimli veri setinde 234 kişiye ait 15979 gen bulunmaktadır. 234 kişiden 141'i (%60.3) COVID-19 negatif ve 93'ü (%39.7) COVID-19 pozitifti. Bu çalışmada, ilgili tahmin edici değişkenleri seçmek için değişken seçim yöntemlerinden LASSO gerçekleştirilmiştir. Modelin performansı doğruluk, dengelenmiş doğruluk, duyarlılık, seçicilik, pozitif tahmin değeri, negatif tahmin değeri ve F1 skoru ile değerlendirildi.

Bulgular: Çalışmanın bulgularına göre, ilişkisel sınıflandırma yönteminden performans ölçütleri doğruluk %92.70, dengelenmiş doğruluk %91.80, duyarlılık %87.10, seçicilik %96.50, pozitif tahmin değeri %94.20, negatif tahmin değeri %91.90 ve F1 puanı %90.50 olarak elde edilmiştir.

Sonuç: Önerilen ilişkisel sınıflandırma yöntemi, COVID-19'u sınıflandırmada çok yüksek performans elde etmiştir. Genlerle ilgili çıkarılan birliktelik kuralları, hastalığın teşhis ve tedavisine yardımcı olabilir.

Anahtar Kelimeler: Birliktelik Kuralları, İlişkisel Sınıflandırma, COVID-19, Gen, Sınıflandırma.

INTRODUCTION

The SARS-CoV-2 virus, which emerged in Wuhan, China's Hubei province on December 31, 2019, quickly spreads to six continents and hundreds of countries, making history the first pandemic caused by coronaviruses (1). This virus has been called SARS-CoV-2 because of its similarity to the Coronavirus (SARS CoV) related to the severe acute respiratory syndrome. The name of the disease it caused has been accepted as COVID-19 worldwide (2). The COVID-19, which has a high contagion property, emanates to the whole world, especially to Europe, in a short time (3). By the World Health Organization (WHO), the COVID-19 outbreak has been declared an International Health Emergency (4).

The COVID-19 pandemic affected the whole world in March, and as of December 2020, 69 million people were reported to be sick in the world. It caused a total of 1,516,516 deaths on six continents around the world (5). The COVID-19 is a highly contagious disease that causes physical, psychological, and widespread systemic to function disorders in patients, especially respiratory disorders (6). The incubation period for COVID-19 is considered within 14 days after exposure, and most cases occur approximately four to five days after exposure (7). The WHO's situation report on February 19 confirmed that the average incubation period is 4-5 days, but it is extended up to 14 days (8). COVID-19 symptoms are not specific. There is no specific clinical feature that can reliably distinguish COVID-19 from other respiratory viral infections. WHO defined common symptoms as fever, fatigue, and dry cough. Other symptoms were reported as shortness of breath, myalgia, sore throat, and very few people diarrhea (9).

According to the data available so far, advanced age (60 years and over), adults with chronic diseases (cardiovascular diseases, hypertension, diabetes, chronic obstructive pulmonary disease, asthma, hypertension, and cancer), obesity, and tobacco use constitute a group at risk for the disease (10).

The related researches have not yet determined the transmission route, diagnosis, clinical features, treatment, and prevention methods of COVID-19. Therefore, it is important to examine genomic sequences for COVID-19 in different clinical studies. Additionally, genomic characterization will help us to describe the origin and evolution of the virus accurately. Demonstrating the mechanism of SARS-CoV-2 replication in various cell-based models can help us understand the pathogenesis and identify specific targets to develop effective antiviral drugs (11).

Data mining can be defined simply as the discovery of useful information hidden in data (12). Data mining enables researchers to make effective and informed decisions with techniques offered by different disciplines such as artificial intelligence, machine learning, statistics, and optimization. It also enables revealing hidden, implicit, beneficial relationships, patterns, relations, or trends that are difficult to reveal with classical methods (13).

Models used in data mining are examined under four headings. These models are; classification, clustering, predictive models, and association rules analysis (14). There is an association rules model under the associative analysis, which is one of the data mining models. Association rules are widely used in data mining due to their easy understanding and usefulness. Methods of data mining that analyze the co-occurrence of events are called the rules of the association. While doing this analysis, association rules express the occurrence of events together with certain probabilities. The association rules' purpose is to give relationships and associations as rules (15, 16).

Associative classification is a classification approach and uses the logic of combining the classification and association rule model, which are among the data mining methods while creating the model. In associative classification, classification models are created with the set of rules obtained by association rule analysis. In the associative classification approach, the response/target variable being on the right side of the obtained rule made it easier to understand and interpret (17).

This study aims to classify COVID-19 by applying the associative classification method on the gene data set consisting of open access COVID-19 negative and positive patients and revealing the disease relationship with these genes by identifying the genes that cause COVID-19.

MATERIAL AND METHODS

Dataset: In the study, an associative classification model was applied to the gene data set of patients with and without open access COVID-19. In this open-access data set used, 15979 genes are belonging to 234 individuals. Out of 234 people, 141 (60.3%) were COVID-19 negative and 93 (39.7%) were COVID-19 positives. Testing for COVID-19 was carried out in the UCSF Clinical Microbiology Laboratory using polymerase chain reaction (PCR) of nasopharyngeal (NP) swab or pooled NP + Oropharyngeal (OP) swab. In all our analyses, we defined patients with COVID-19 as those with a positive SARS-CoV-2 result by PCR. Detailed protocol and sample information are available at the relevant web address (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE156063) (18).

Feature Selection: Working with high dimensional data to be analyzed and modeled increases computation time, making interpretation difficult, and leading to computational inefficiency. For this reason, reducing the data size in high dimensional data increases the ease of analysis. The

main purpose of the size-reduction methods is to minimize the loss of data contained by the maximum reduction of data size. Choosing the most suitable data for the purpose is important in increasing the analysis's quality and obtaining meaningful results (19).

The gene data set in this study is a high dimensional data set. For this reason, feature selection has been made in this data set to reduce the calculation time, eliminate calculation inefficiency, and increase the quality of the analysis. LASSO feature selection method was used while performing the feature selection process.

LASSO feature selection was first proposed by Tibshirani (1996) to increase Least Squares (LS) prediction accuracy, and its usage areas have been expanded over time. Today, it is used in many different areas, especially in the field of health. When the areas of use are examined, it can be seen that the studies focus on data sets where the number of variables is higher than the number of observations, especially in very large data sets (20). The LASSO (Least Absolute Shrinkage and Selection Operator) estimator aims to increase the prediction accuracy by getting a certain penalty to the LS estimation equation. In the LASSO estimator, one or more parameters are narrowed to zero in the prediction equation due to penalty constraints. Thus, it allows obtaining models that are easy to interpret, especially in large data sets. LASSO is a widely preferred estimator because it provides ease of operation by enabling feature selection and parameter estimation simultaneously (21).

Association Rules and Associative Classification

Association Rules: Data mining is a technique that tries to identify previously unknown hidden relationships between data in databases. Data mining uses statistical, mathematical, and machine learning techniques while exploring and revealing these relationships (22). Association rules, which is one of the data mining methods, explain the occurrence of some events in the database with probabilistic expressions (23).

Association rules used to discover hidden relationships in large data sets are unsupervised data mining methods. Association rules define potential relationships of data. The aim is to reveal the rules of events that are likely to occur together. With this method, a series of operations are applied to the records in the databases in bulk, and the rules explaining the relationship between the records are derived (24). Form of association rules, "IF <if certain conditions are met>" is in the form "THEN <estimate the values of some attributes>" and are rules that are measures of support and confidence (25). Confidence and support value are units of measure that show power in the association rule. Rules with high trust and support values are called strong rules. The researcher determined the minimum support threshold value (min_support) and the minimum confidence threshold value (min_confidence) in the association rules. Association rules with higher values than the specified threshold values are taken into consideration (26).

Associative Classification: Associative classification is a new, effective supervised learning approach that aims to predict unseen situations. Particularly, an associative classification is an approach that uses rules obtained with association rules to create classification models. Associative classification effectively integrates classification with association rule mining and can produce more accurate results than other traditional data mining classification algorithms. On the right side of association rules in associative classification, it consists only of the class/response / dependent variable categories. The rules of the association are derived using precursor-successor clauses called if-then. Thus, it becomes easier for the user to understand and interpret the output. This situation ensures that associative classification is more advantageous than classical classification approaches (17).

There are many algorithms used and developed in associative classification. Classification based on association rules (CBA) algorithm was used in this study.

Classification Based on Association Rules (CBA): Classification based on association rules is an algorithm consisting of two parts that combine the classification and association rules. The first episode, called CBA-RG, is an adaptive version of Apriori used to find CARs (complete set of class association rules). The second part of CBA, called CBA-CB, is an algorithm that builds the classifier based on CARs found using CBA-RG. The classifier is built in three steps. First, the discovered CARs rank in order of priority; if the reliability of r1 is greater than r2 or if the support of r1 is greater than r2 while having the same confidence, the r1 rule is considered to precede r2. In step 2, all rules that correctly classify at least one case, and the majority class of undiscovered data are selected. Finally, in step 3, rules that do not improve the classifier's accuracy are discarded (27).

Performance Evaluation Criteria: The classification matrix for the calculation of performance metrics is given in Table 1.

Table 1. The classification 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)
Balanced accuracy = [[TP/(TP+FN))]+[TN/(FP+TN)]]/2
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)

Data Analysis: Quantitative data are summarized by median (minimum-maximum). Normal distribution was evaluated with the Kolmogorov-Smirnov test. In terms of input variables, the existence of a statistically significant difference and the relationship between the categories of the output variable, "positive " and "negative" groups, were examined using the Mann-Whitney U test. $p<0.05$ values were considered statistically significant. IBM SPSS Statistics 26.0 for the Windows package program was used in the analysis. A web-based application developed by İnönü University Faculty of Medicine Biostatistics and Medical Informatics Department was used (28).

RESULTS

In this study, 31 genes remained in the data set after the LASSO feature selection method was applied to the data set consisting of 15979 genes. Thirty-one genes obtained by the LASSO trait selection method are given in Table 2.

Descriptive statistics for the variables examined in this data set are given in Table 3.

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

Genes					
LMO3	RASL11A	ITGB1BP2	METRNL	GLTPD2	TBCE
PCSK5	RTN2	FAM83A	SIX5	DCUN1D3	
VSIG1	LGR6	AZGP1	CD163L1	TPSB2	
BACH2	TPT1	SCGB3A1	PCDHB9	ERVMER34-1	
PDGFRB	TNS3	IFI27	LDLRAD3	MTRNR2L12	
CR2	DUSP6	STK32A	ALOX15B	AC005832.4	

Table 3. Descriptive statistics for quantitative independent variables

Genes	**Groups**		**p-value***
	Negative	**Positive**	
	Median(min-max)	**Median(min-max)**	
LMO3	8 (0-95)	7 (0-168)	0.654
PCSK5	311 (0-21818)	891 (14-63448)	**<0.001**
VSIG1	12 (0-155)	6 (0-93)	**0.023**
BACH2	29 (0-708)	42 (0-2224)	0.186
PDGFRB	11 (0-592)	2 (0-218)	**<0.001**
CR2	7 (0-130)	8 (0-143)	0.579
RASL11A	16 (0-214)	10 (0-119)	0.046
RTN2	32 (0-585)	15 (0-158)	**0.002**
LGR6	18 (0-307)	50 (0-386)	**<0.001**
TPT1	3438 (322-38186)	2801 (69-19508)	**0.009**
TNS3	109 (0-3155)	94 (1-1677)	0.085
DUSP6	175 (0-2178)	102 (1-931)	**<0.001**
ITGB1BP2	6 (0-63)	4 (0-49)	0.085
FAM83A	656 (24-11025)	1700 (4-28484)	**<0.001**
AZGP1	31 (0-621)	26 (0-379)	**0.016**
SCGB3A1	30 (0-7220)	13 (0-1008)	**0.012**
IFI27	254 (3-6763)	1014 (22-4814)	**<0.001**
STK32A	3 (0-297)	4 (0-42)	0.757
METRNL	148 (6-2373)	80 (2-518)	**<0.001**
SIX5	35 (0-438)	28 (0-271)	0.227
CD163L1	2 (0-184)	1 (0-84)	0.168
PCDHB9	9 (0-148)	3 (0-51)	**<0.001**
LDLRAD3	5 (0-229)	5 (0-130)	0.155
ALOX15B	9 (0-200)	5 (0-85)	**0.017**
GLTPD2	21 (0-207)	12 (0-92)	**0.004**
DCUN1D3	314 (4-3904)	117 (3-657)	**<0.001**
TPSB2	4 (0-339)	7 (0-2935)	0.383
ERVMER34-1	20 (0-157)	14 (0-197)	0.063
MTRNR2L12	18 (0-1865)	15 (0-321)	0.282
AC005832.4	1 (0-100)	3 (0-50)	0.954
TBCE	371 (2-9996)	280 (0-51630)	0.548

***: Mann-Whitney U test.**

According to the findings obtained; There is a statistically significant difference between the dependent/target variable groups in terms of BPCSK5, VSIG1, PDGFRB, RTN2, LGR6, TPT1, DUSP6, FAM83A, AZGP1, SCGB3A1, IFI27, METRNL, PCDHB9, ALOX15B, GLTPD2, DCUN1D3 variables ($p<0.05$).

The distribution table for the dependent/target variable in the data set obtained with 31 genes is given in Table 4.

Table 4. Distribution table of the dependent/target variable

Negative		**Positive**	
Count	Percentage	Count	Percentage
141	60.3	93	39.7

The associative classification model was used to classify the dataset in this study. The classification matrix of this model is given below in Table 5.

Table 5. Classification matrix for the associative classification model

Prediction	**Reference**		
	Positive	**Negative**	**Total**
Positive	81	5	86
Negative	12	136	148
Total	93	141	234

The values for the classification performance metrics for the associative classification model are shown in Table 6. From the associative classification model, the obtained accuracy was 92.70%, balanced accuracy 91.80%, sensitivity 87.10%, Specificity 96.50%, positive predictive value 94.20%, negative predictive value 91.90%, and F1-score 90.50%.

Table 6. Values for the classification performance metrics of the associative classification model

Metric	**Value** (%)
Accuracy	92.70
Balanced accuracy	91.80
Sensitivity	87.10
Specificity	96.50
Positive predictive value	94.20
Negative predictive value	91.90
F1-score	90.50

Table 7 shows the association rules used by the classification algorithm. As expressed in Table 7, when TPT1=[69,8.14e+03), IFI27=[622,6.76e+03), METRNL=[2,155) and MTRNR2L12=[0,188) are considered, the probability of COVID-19 positive is 100%. Similarly, TPT1=[69,8.14e+03), ITGB1BP2=[0,18.5), IFI27=[622,6.76e+03) and DCUN1D3=[3,254) are taken into account, the probability of COVID-19 positive is 100%, and when ITGB1BP2=[0,18.5), IFI27=[622,6.76e+03), SIX5=[0,112) and DCUN1D3=[3,254) are regarded, the probability of COVID-19 positive is 100%. If FAM83A=[4,1.75e+03), IFI27=[3,622) and PCDHB9=[6.5,148) are considered, the probability of COVID-19 negative is 100%. Similarly, ITGB1BP2=[0,18.5), IFI27=[622,6.76e+03) and DCUN1D3=[3,254) are considered, the probability of COVID-19 positive is 98.1%.The other rules generated from the classification based on association rules model can be interpreted as the rules described earlier (Table 7).

DISCUSSION

In December 2019, the new COVID-19 outbreak in Wuhan, China's Hubei province, started as an epidemic and turned into a pandemic in a short time. The most important feature that distinguishes COVID-19 from other pandemics is that it is concentrated in underdeveloped countries and developing countries (29). This disease has become the most important health problem of the 21st century due to its high contagious feature, unfavorable clinical prognosis, and lethal effect in almost every age group, especially those aged 65 and above (30).

COVID-19 poses a serious threat to global public health today, and the very high human-to-human transmission capacity raises concerns about the control of the epidemic. It is not known how the pandemic will follow in the next period; It is thought that studies and investments on preventive healthcare services should be increased within health systems (31). Therefore, it is important to develop effective treatments to clarify the virus's source to combat the rapidly advancing COVID-19 pandemic. Therefore, it is necessary to reveal the genome structure of the virus. In the COVID-19 outbreak, host genomic factors cause the disease to manifest with quite different clinical symptoms. As disease-causing host genomic factors are discovered, new strategies that support rapid clinical practice can be put forward to achieve recovery in SARS-CoV-2 infected patients (32).

Due to the large size of the data in medical databases, effective data mining methods are needed. Information obtained from medical data processed using different data mining techniques is valuable for decision making, diagnosis, and predictions. The main challenge in data mining is to create a sensitive and efficient classifier (33).

In recent years, a new approach called associative classification that combines attribution and classification has been proposed. Association rule mining and classification are two important data mining methods, and associative classification combines these two methods. There is evidence that combining classification and association rule mining will give more efficient and more accurate classification performance than traditional classification techniques. Since the result of the rule

Table 7. Association rules used by the classification algorithm

Left-hand side rules	Right-hand side rules	Support	Confidence	Frequency
{TPT1=[69, 8.14e+03), IFI27=[622, 6.76e+03), METRNL=[2, 155), MTRNR2l12=[0, 188)}	{grup=positive}	0.222	1	52
{TPT1=[69, 8.14e+03), ITGB1BP2=[0, 18.5), IFI27=[622, 6.76e+03), DCUN1D3=[3, 254)}	{grup=positive}	0.218	1	51
{ITGB1BP2=[0, 18.5), IFI27=[622, 6.76e+03), SIX5=[0, 112), DCUN1D3=[3, 254)}	{grup=positive}	0.218	1	51
{FAM83A=[4, 1.75e+03), IFI27=[3, 622), PCDHB9=[6.5, 148)}	{grup=negative}	0.205	1	48
{LMO3=[0, 56), PDGFRB=[15.5, 592), DCUN1D3=[254, 3.9e+03)}	{grup=negative}	0.239	0.982	56
{ITGB1BP2=[0, 18.5), IFI27=[622, 6.76e+03), DCUN1D3=[3, 254)}	{grup=positive}	0.226	0.981	53
{LMO3=[0, 56), PCDHB9=[6.5, 148), DCUN1D3=[254, 3.9e+03)}	{grup=negative}	0.226	0.981	53
{TPT1=[69, 8.14e+03), IFI27=[622, 6.76e+03), METRNL=[2, 155)}	{grup=positive}	0.222	0.981	52
{VSIG1=[0, 30.5), TPT1=[69, 8.14e+03), IFI27=[622, 6.76e+03), DCUN1D3=[3, 254)}	{grup=positive}	0.218	0.981	51
{VSIG1=[0, 30.5), TNS3=[0, 560), IFI27=[622, 6.76e+03), STK32A=[0.5, 34.5)}	{grup=positive}	0.209	0.98	49
{LMO3=[0,56), METRNL=[155,2.37e+03), DCUN1D3=[254,3.9e+03),TBCE=[97.5,2.42e+03)}	{grup=negative}	0.205	0.98	48
{LMO3=[0, 56), LGR6=[0, 35.5), DUSP6=[172, 2.18e+03), TBCE=[97.5, 2.42e+03)}	{grup=negative}	0.205	0.98	48
{RASL11A=[0, 22.5), TNS3=[0, 560), IFI27=[622, 6.76e+03), MTRNR2L12=[0, 188)}	{grup=positive}	0.201	0.979	47
{LMO3=[0, 56), LGR6=[0, 35.5), DCUN1D3=[254, 3.9e+03), TBCE=[97.5, 2.42e+03)}	{grup=negative}	0.239	0.966	56
{LMO3=[0, 56), METRNL=[155, 2.37e+03), DCUN1D3=[254, 3.9e+03)}	{grup=negative}	0.226	0.964	53
{TPT1=[69, 8.14e+03), TNS3=[0, 560), IFI27=[622, 6.76e+03), ALOX15B=[0, 24)}	{grup=positive}	0.226	0.964	53
{TNS3=[0, 560), IFI27=[622, 6.76e+03), SIX5=[0, 112), ALOX15B=[0, 24)}	{grup=positive}	0.226	0.964	53
{LMO3=[0, 56), PDGFRB=[15.5, 592), CR2=[0, 102), LGR6=[0, 35.5)}	{grup=negative}	0.218	0.962	51
{LMO3=[0, 56), LGR6=[0, 35.5), METRNL=[155, 2.37e+03)}	{grup=negative}	0.214	0.962	50
{CR2=[0, 102), SCGB3A1=[41.5, 7.22e+03), IFI27=[3, 622)}	{grup=negative}	0.214	0.962	50
{RTN2=[0, 58), ITGB1BP2=[0, 18.5), IFI27=[622, 6.76e+03), ALOX15B=[0, 24)}	{grup=positive}	0.209	0.961	49
{IFI27=[3, 622), PCDHB9=[6.5, 148)}	{grup=negative}	0.244	0.95	57
{CR2=[0, 102), LGR6=[0, 35.5), FFAM83A=[4, 1.75e+03), DCUN1D3=[254, 3.9e+03)}	{grup=negative}	0.235	0.948	55
{SCGB3A1=[41.5, 7.22E+03), IFI27=[3, 622)}	{grup=negative}	0.214	0.943	50
{PCSK5=[0, 347), IFI27=[3, 622), TBCE=[97.5, 2.42e+03)}	{grup=negative}	0.244	0.934	57

obtained in the associative classification method is with the response variable, it is possible to create a more accurate classifier (34, 35).

Associative classification stands out as a new approach that provides easier interpretation for users when applied to medical data sets (34). In this study, an associative classification model was applied to an open-access gene data set. In this context, different factors (explanatory variables) that may be associated with COVID-19 (the dependent variable) positive-negative are estimated with the associative classification model, and rules have been obtained. According to the results of the findings, from the performance metrics obtained

from the associative classification model, the accuracy was 92.70%, balanced accuracy 91.80%, sensitivity 87.10%, specificity 96.50%, positive predictive value 94.20%, negative predictive value 91.90%, and F1-score 90.50%.

The same data set was used in an article where the most significant genes upregulated by SARS-CoV-2 were interferon-inducible, including IFI6, IFI44L, IFI27, and OAS2. Also, IFI27 was induced by SARS-CoV-2 significantly more than by other viruses, even at low viral load(18). In this study, the TPT1, IFI27, METRNL, MTRNR2l12, ITGB1BP2, DCUN1D3, LMO3, SIX5, VSIG1, STK32A, RASL11A, TNS3, ALOX15B, RTN2, ITGB1BP2 genes determine the status of being COVID-19 positives. In this study, the proposed associative classification method achieved very high performances in determining disease-related genes. Besides, this research presents an associative classification model to help researchers diagnose early for COVID-19 prediction.

In this study, the study was completed using an open source data set. If the study could be studied with real data, real experimental results could be reached and the results could be made more general and valuable. In addition, with the relational classification method preferred in the study, the results were obtained by considering the interdependent conditions of the genes associated with the disease, and a set of rules for the conditions causing the disease were obtained.

CONCLUSION

As a result, genes associated with these rules can help in the early diagnosis and treatment of the disease. The disease can be successfully managed if further research is encouraged to develop this area's prediction system.

REFERENCES

1. Dikmen AU, Kına MH, Özkan S, İlhan MN. COVID-19 epidemiyolojisi: Pandemiden ne öğrendik. Journal of biotechnology and strategic health research. 2020;4:29-36.
2. Yücel E, Tamay ZÜ. Astım ve COVID-19. Çocuk Dergisi.20(2):76-9.
3. Keskin M, Derya Ö. COVID-19 sürecinde öğrencilerin web tabanlı uzaktan eğitime yönelik geri bildirimlerinin değerlendirilmesi. İzmir Katip Çelebi Üniversitesi Sağlık Bilimleri Fakültesi Dergisi. 2020;5(2):59-67.
4. COVID WA. Outbreak a Pandemic; 2020. Back to cited text. (1).
5. Organization WH. Coronavirus disease (COVID-19) pandemic [cited 2020 14 December]. Available from: https://covid19.who.int/.
6. Aytür YK, Köseoğlu B, Taşkıran ÖÖ, Gökkaya NKO, Delialioğlu SÜ, Tur BS, et al. SARS-CoV-2 (COVID-19) sonrası pulmoner rehabilitasyon prensipleri: Akut ve subakut sürecin yönetimi için rehber. Fiziksel Tıp ve Rehabilitasyon Bilimleri Dergisi. 2020;23(2):111-23.
7. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. New England Journal of Medicine. 2020.
8. Mirzaei H, McFarland W, Karamouzian M, Sharifi H. COVID-19 among people living with HIV: a systematic review. AIDS and Behavior. 2020:1-8.
9. Sheikhi K, Shirzadfar H, Sheikhi M. A review on novel coronavirus (Covid-19): symptoms, transmission and diagnosis tests. Research in Infectious Diseases and Tropical Medicine. 2020;2(1):1-8.
10. Pala K. COVID-19 Pandemisi ve Türkiye'de Halk Sağlığı Yönetimi. Sağlık ve Toplum. 2020;30(Özel Sayı):39-50.
11. Tanrıverdi ES. COVID-19 Etkeninin Özellikleri.
12. Silahtaroğlu G. Kavram ve Algoritmalarıyla Temel Veri Madenciliği Papatya Yayıncılık Eğitim AŞ. İstanbul, Türkiye. 2008.
13. Witten IH, Frank E. Data mining: practical machine learning tools and techniques with Java implementations. Acm Sigmod Record. 2002;31(1):76-7.
14. Moss LT, Atre S. Business intelligence roadmap: the complete project lifecycle for decision-support applications: Addison-Wesley Professional; 2003.
15. Chen Y-L, Chen J-M, Tung C-W. A data mining approach for retail knowledge discovery with consideration of the effect of shelf-space adjacency on sales. Decision support systems. 2006;42(3):1503-20.
16. Vinodh S, Prakash NH, Selvan KE. Evaluation of leanness using fuzzy association rules mining. The International Journal of Advanced Manufacturing Technology. 2011;57(1-4):343-52.
17. Thabtah FA. A review of associative classification mining. Knowledge Engineering Review. 2007;22(1):37-65.
18. Mick E, Kamm J, Pisco AO, Ratnasiri K, Babik JM, Calfee CS, et al. Upper airway gene expression differentiates COVID-19 from other acute respiratory illnesses and reveals suppression of innate immune responses by SARS-CoV-2. medRxiv. 2020.
19. Çalışan M, Talu MF. Boyut İndirgeme Yöntemlerinin Karşılaştırmalı Analizi. Türk Doğa ve Fen Dergisi.9(1):107-13.
20. Zou H. The adaptive lasso and its oracle properties. Journal of the American statistical association. 2006;101(476):1418-29.

21. Zhang HH, Lu W. Adaptive Lasso for Cox's proportional hazards model. Biometrika. 2007;94(3):691-703.
22. Kamber M, Pei J. Data mining: Concepts and techniques: Morgan Kaufmann Publishers San Francisco; 2001.
23. Kumar AS, Wahidabanu R, editors. A frequent item graph approach for discovering frequent itemsets. 2008 International Conference on Advanced Computer Theory and Engineering; 2008: IEEE.
24. Fayyad UM, Piatetsky-Shapiro G, Smyth P, Uthurusamy R, editors. Advances in knowledge discovery and data mining1996: American Association for Artificial Intelligence.
25. Larose DT, Larose CD. Discovering knowledge in data: an introduction to data mining: John Wiley & Sons; 2014.
26. Han J, Pei J, Kamber M. Data mining: concepts and techniques: Elsevier; 2011.
27. Azmi M, Runger GC, Berrado A. Interpretable regularized class association rules algorithm for classification in a categorical data space. Information Sciences. 2019;483:313-31.
28. Arslan AK, Küçükakçalı Z, Balıkçı Çiçek İ, Çolak C. A Novel Interpretable Web-Based Tool On The Associative Classification Methods: An Application On Breast Cancer Dataset. The Journal of Cognitive Systems.5(1):33-40.
29. Bingül BA, Türk A, Ak R. Covid-19 Bağlamında Tarihteki Büyük Salgınlar ve Ekonomik Sonuçları. Electronic Turkish Studies. 2020;15(4).
30. Üstün Ç, Özçiftçi S. COVID-19 pandemisinin sosyal yaşam ve etik düzlem üzerine etkileri: Bir değerlendirme çalışması. Anadolu Kliniği Tıp Bilimleri Dergisi. 2020;25(Special Issue on COVID 19):142-53.
31. Çiftçi E, Çoksüer F. Yeni Koronavirüs İnfeksiyonu: COVID-19. Flora İnfeksiyon Hastalıkları ve Klinik Mikrobiyoloji Dergisi. 2020;25(1):9-18.
32. Murray MF, Kenny EE, Ritchie MD, Rader DJ, Bale AE, Giovanni MA, et al. COVID-19 outcomes and the human genome. Genetics in Medicine. 2020:1-3.
33. Liu B, Hsu W, Ma Y, editors. Integrating classification and association rule mining. KDD; 1998.
34. Jabbar MA, Deekshatulu BL, Chandra P, editors. Heart disease prediction using lazy associative classification. 2013 International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s); 2013: IEEE.
35. Haafeeza K, Mohanraj R. Classification of Multi Disease Diagnosing and Treatement Analysis Based on Hybrid Mining Technique. International Journal of Advanced Technology and Innovative Research. 2014;6(3):108-16.