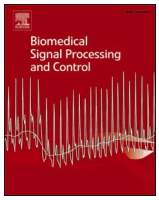
Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Biomedical Signal Processing and Control 77 (2022) 103836

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc



A novel method using Covid-19 dataset and machine learning algorithms FOR THE MOST ACCURATE DIAGNOSIS that can be obtained in medical diagnosis

Emre Avuçlu

Department of Software Engineering, Faculty of Engineering, Aksaray University, Aksaray, Turkey

ARTICLE INFO

Keywords:
The most accurate diagnosis
Biomedical images
COVID-19
Feature extraction
Machine learning algorithms

ABSTRACT

Pandemics and many other diseases threaten human life, health and quality of life by affecting many aspects. For this reason, the medical diagnosis to be applied for any disease is important in terms of the most accurate determination by the doctors and the appropriate treatment for the determined diagnosis. The COVID-19 pandemic that started in China in December 2019 spread all over the world in a short time. Researchers have begun to do different studies to make the most accurate diagnosis of COVID-19. Due to the rapid spread of COVID-19, doctors in the health sector of many countries were also caught off guard. Machine Learning Algorithms (MLAs) are of great importance in the development of computer-aided early and accurate diagnosis systems in today's medical field, as they greatly assist doctors in the medical diagnosis process. In this study, a method was proposed for the most accurate diagnosis of COVID-19 patients using the COVID-19 image data. Images were first standardized and features extracted using RGB values of 800x800 images, and these features were used in train and test processes for MLAs. 5 different MLAs were used in experimental studies using statistical measurements (k Nearest Neighbor (k-NN), Decision Tree (DT), Multinominal Logistic Regression (MLR), Naive Bayes (NB) and Support Vector Machine (SVM)). A method was proposed that automatically finds the highest classification success that these algorithms can achieve. In experimental studies, the following accuracy rates were obtained in train operations for MLAs, respectively; 1, 1, 1, 0.69565, 0.92753. Accuracy results in test operations were obtained as follows; 0.85714, 0.79591, 0.91836, 0.61224, 0.89795. After the application of the proposed method, the test success rate for MLR increased from 0.91 to 0.98. As a result of applying the proposed algorithm, more accurate results were obtained. The results obtained were given in the experimental studies section in detail. The results obtained proved to be very promising. According to the results, it was seen that the proposed method could be used effectively in future studies.

1. Introduction

Medical diagnosis worldwide is primarily based on the patient's medical history and physical examination by the doctor. The success degree of medical diagnosis depends on multiple factors, the most important of which are the scientific competence, experience and technical materials of the doctor. Regardless of the competence of physicians or the high quality of technical equipment used, it is inevitable that they will make medical mistakes, especially in the diagnosis of the disease [1]. Researchers have stated that the number of false or late diagnoses increases with each passing year and these cause death of people living in many parts of the world [2,3]. It has been identified as the third leading cause of deaths in the United States due to misdiagnosis, after medical errors, cardiovascular diseases and cancer [4]. The probability of successful treatment of a patient diagnosed early usually increases depending on the time of early diagnosis. In addition, rehabilitation costs in early diagnosis determined by the doctor are much lower than in late diagnosis. In other words, unnecessary inspection procedures do not occur. Medical errors can result in the constant emergence of new and complex diseases. This is due to the lack of tools that help doctors make the right decision [5]. In some of the complex diseases, making the correct diagnosis decision can be very difficult and sometimes impossible [6,7]. Today, decision support systems are among the powerful tools that help doctors diagnose diseases and make decisions. Many studies have been conducted on decision support systems to help doctors make diagnoses more easily. In this

E-mail address: emreavuclu@aksaray.edu.tr.

https://doi.org/10.1016/j.bspc.2022.103836
Received 23 March 2022; Received in revised form 5 May 2022; Accepted 27 May 2022
Available online 30 May 2022

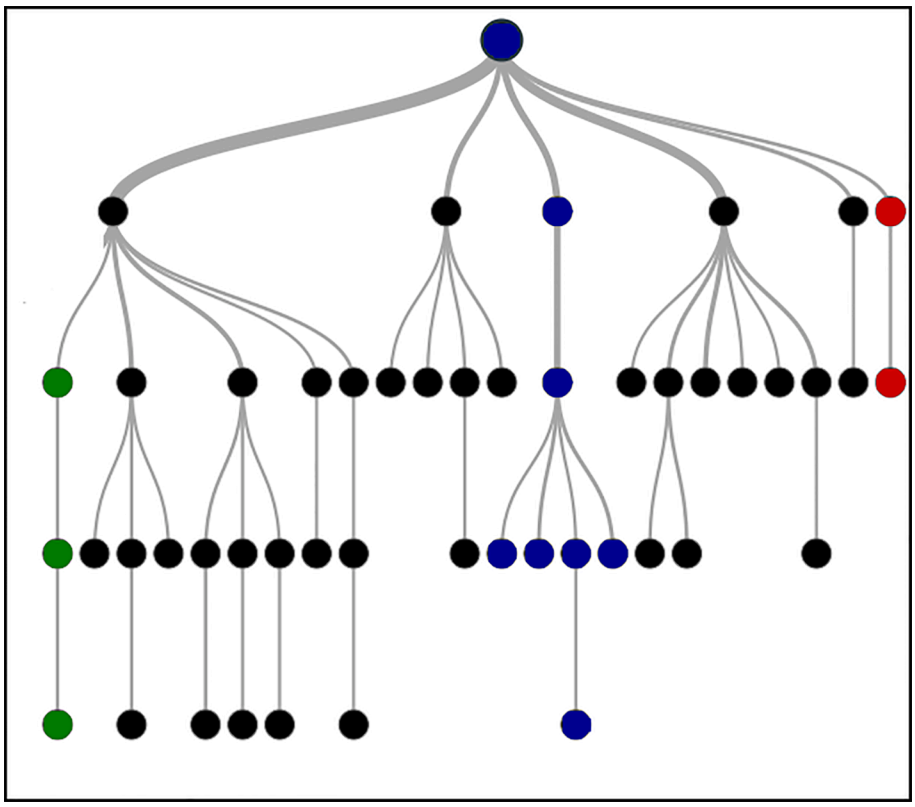


Fig. 1. DT algorithm.

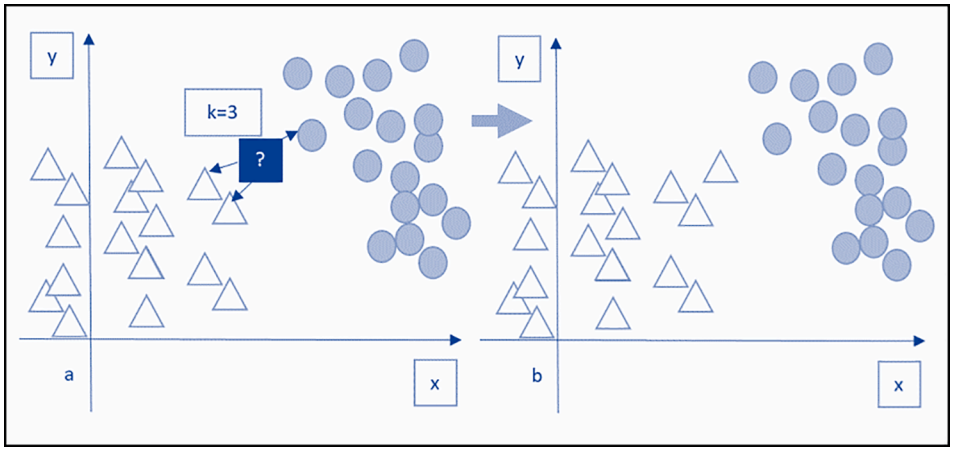


Fig. 2. Classification of a new data.

sense, expert systems and artificial intelligence techniques are successfully used to solve different problems in various medical branches [8]. Accurate diagnosis is of paramount importance for COVID-19, which spread rapidly all over the world and became a pandemic in a short time. Outbreaks of many infectious diseases such as plague, Spanish flu, severe acute respiratory failure syndrome (SARS), swine flu, Middle East Respiratory Syndrome (MERS) have occurred in the world history. As a result of the widespread trade network and travels, the spread of such epidemics has accelerated [9]. There are different interpretations from different sources about the number of cases when the COVID-19 disease was first seen [10,11]. It has been reported that these patients are Chinese citizens and that these people may be responsible for the spread of the epidemic [12–14]. In the press release of the World Health Organization (WHO) on March 13, 2020, it is informative about the way the disease will follow the next course of the disease, using the statement "Detect to break the COVID-19 spread chain and every case we treat will restrict the spread of the disease" [15]. It has been suggested that the virus, which is the causative agent of COVID-19 disease, is of a zoonotic nature, and that it is transmitted from animals called bats and pangolin, which are possible hosts, to humans [16]. According to some sources, it is stated that there is no reliable evidence to support this claim, although there is information that COVID-19 is artificially or intentionally produced in a laboratory setting [17,18]. Some results were obtained in a study prepared with the contribution of some organizations affiliated to the German Federal Government. It has been reported that the route of transmission is in the form of droplets, the symptoms of dry cough, fever, shortness of breath will be seen in infected people, lung findings will be found on radiography, and patients may have tremor, nausea and muscle pain. In the scenario that creates the risk analysis report, the distribution rate of the disease by age groups is 50% for 65 years and

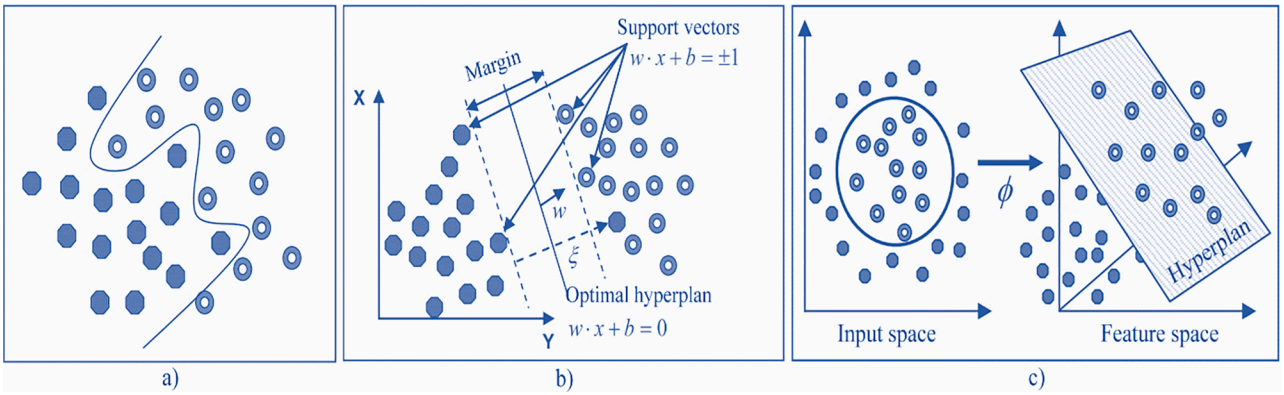


Fig. 3. Non-linear SVM; a) non-linearly separable data set, b) determination of the separation plane for non-linearly separable data sets, c) conversion of input space to property space.

Sensitivity or True Positive Rate	$TPR = \frac{TP}{TP+FN}$	Dice Similarity Coefficient	$DSC = \frac{2TP}{2TP+FP+FN}$
Specificity or True Negative Rate	$TNR = \frac{TN}{TN+FP}$	Accuracy	$ACC = \frac{TP+TN}{TP+TN+FP+FN}$
Precision or Positive Predictive Value	$PPV = \frac{TP}{TP+FP}$		
Negative Predictive Value	$NPV = \frac{TN}{TN+FN}$	F-Measurements	$FM = \frac{2}{\frac{1}{TPR}+\frac{1}{PPV}}$
False Positive Ratio	$FPR = \frac{FP}{TN+FP}$	Matthews Correlation Coefficient $MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$	
False Negative Ratio	$FNR = \frac{FN}{TP+FN}$		

Fig. 4. Statistical measurements.

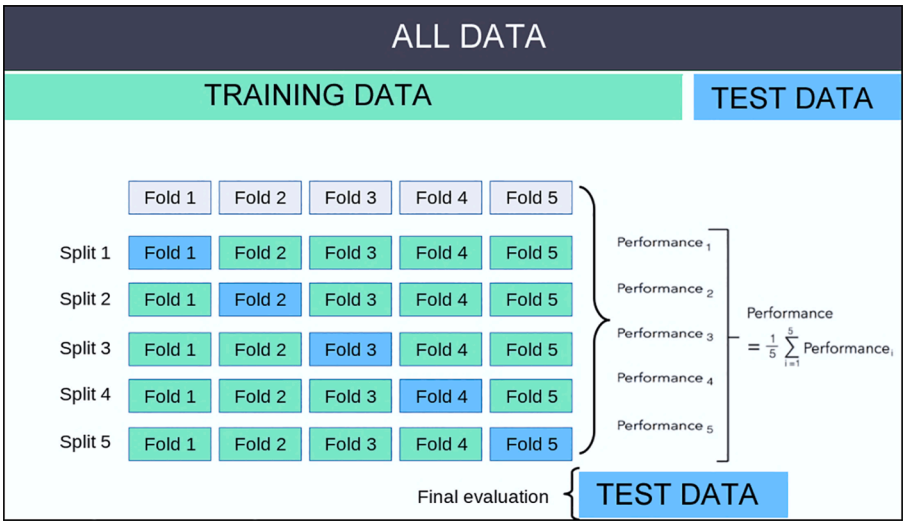


Fig. 5. 5-fold crossvalidation process.

over [19,20]. Due to the COVID-19 pandemic, the estimated mortality rate by WHO is 3.4% worldwide as of March 3, 2020, and when the analysis of these death cases; It has been found that 2/3 of males, more than 80% are over 60 years old, 1/3 are seen in females, and more than 75% have chronic diseases such as cardiovascular diseases, diabetes and cancer [21–23]. COVID-19, which has become a global pandemic, has caused some negative consequences, not only medical, but also social, professional, political, economic, ethical and moral. It has been observed that the epidemic spread has turned into a pandemic due to the countries not being ready for the COVID-19 epidemic, some delays and confusion in the measures to be taken, and the high transmission rate of the disease [24,25]. The illness of well-known people in the fields of politics, arts and sports in the world is another proof that the disease has spread and contagious in a pandemic [26,27]. Although COVID-19

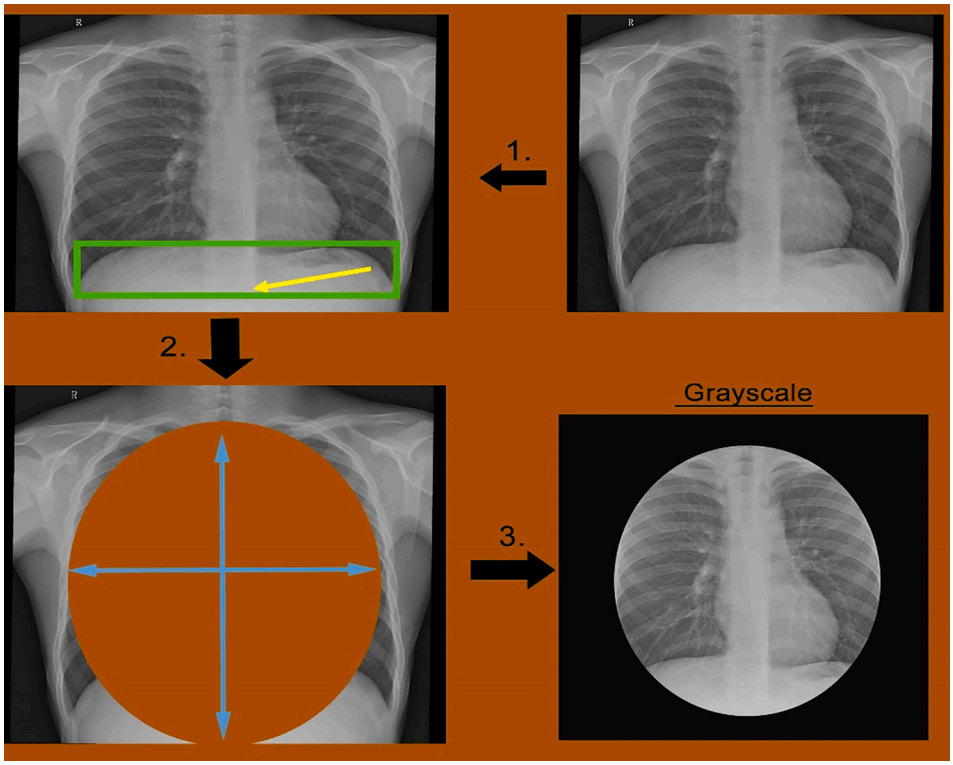


Fig. 6. Image preprocessing.

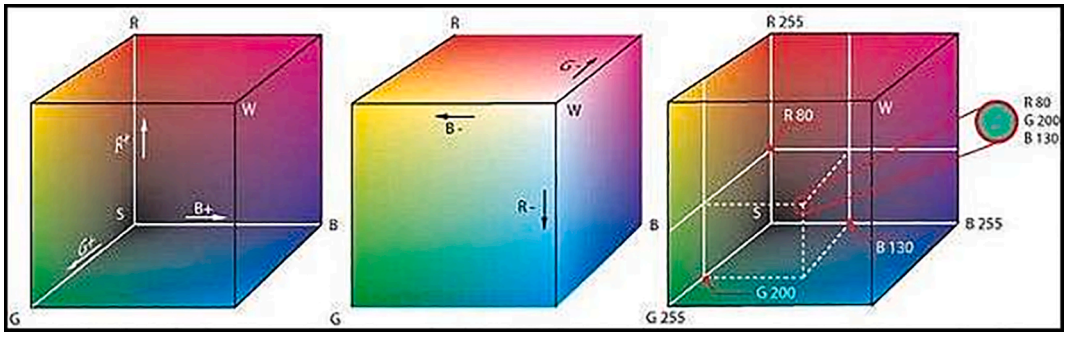


Fig. 7. Color space on the coordinate.

mainly causes lung pneumonia, it can cause disease in multiple organ systems [28]. While 14 of the patients in Italy had respiratory symptoms compatible with viral infection, a definite diagnosis was made in one patient with systemic sclerosis and the patient was lost [29]. They proposed a new feature extraction method for the automated COVID-19 classification process [30–33]. They developed a system to make a complete and accurate diagnosis of COVID-19. Using different deep learning methods, they achieved 96.45% precisions [34]. They processed the X-ray images with the convolutional attention network and obtained an accuracy of 98.02 ± 1.35% [35]. Wang et al., in a study conducted in a different field, said that they had the highest accuracy rate of 96.67% from MLR, as in this study [36].

Articles on COVID-19 and other diseases, which are mostly based on scientific data, are made to help experts in accurate diagnosis and treatment of diseases. The aim of this study is to propose an algorithm to make the most accurate diagnosis in MLAs training and testing processes to be used in any field. For this, the proposed algorithm was tested with two different datasets. The results proved to be promising. Thus, it is aimed to ensure that the studies to be carried out on behalf of humanity in the literature are carried out in the most accurate way possible. In the second part of the study, MLA, k-fold crossvalidation, Confusion Matrix (CM), dataset, feature extraction and the recommended method are presented. In the third part, experimental studies and statistical measurements obtained for two different databases are discussed and presented. The results obtained are given in the last section.

2. Materials and methods

This section provides information about the database, MLA, k-fold cross validation, CM, datasets, feature extraction, proposed method in the study.

2.1. Machine learning algorithms

Machine learning searches for some patterns in data with various algorithms and methods. The working structure of the 5 different machine learning algorithms used in this study are explained in the following titles respectively.

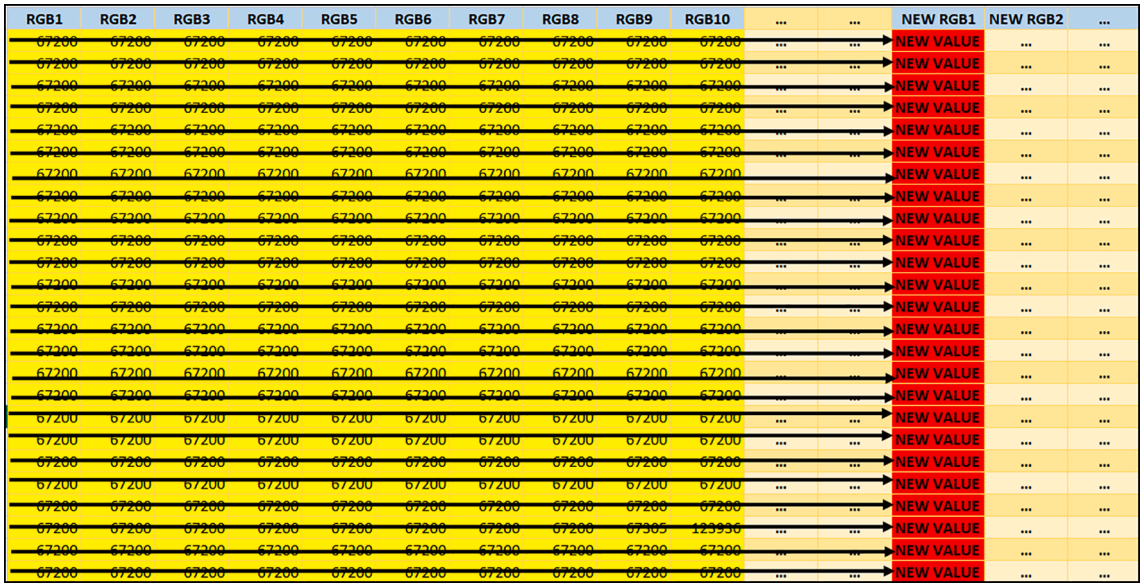


Fig. 8. Grouping RGB values.

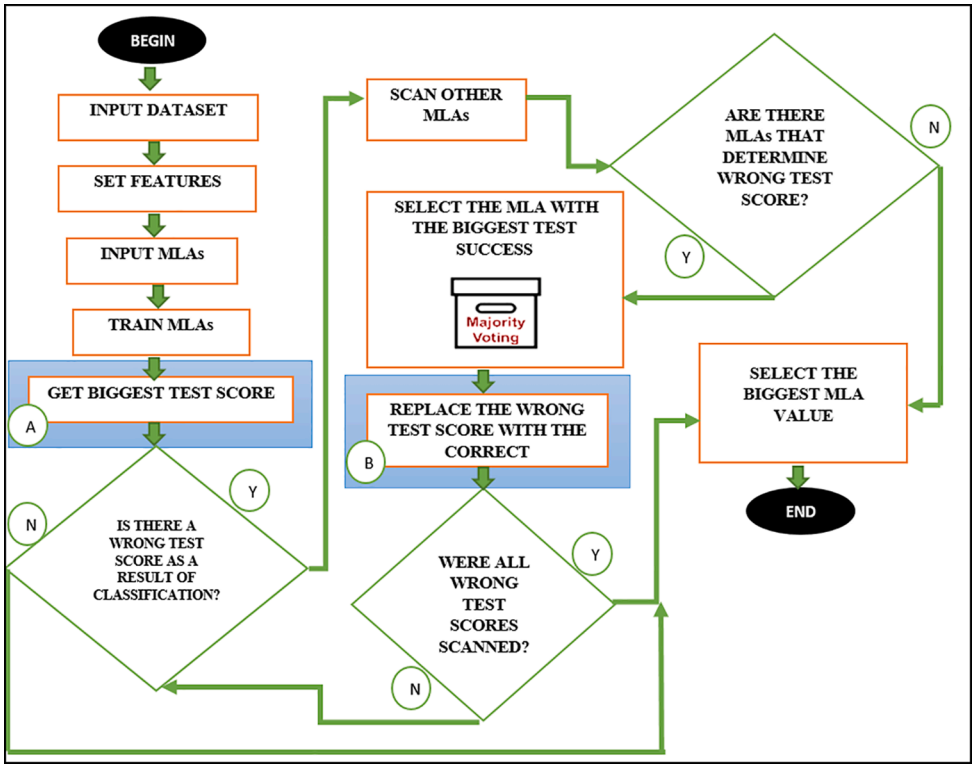


Fig. 9. Flowchart of the proposed method.

2.1.1. Decision tree classifier

Tree-based learning algorithms are among the most used supervised learning algorithms. Generally, they can be adapted to the solution of all the problems (classification and regression) dealt with. A decision tree is a structure used to divide a data set containing a large number of records into smaller sets by applying a set of decision rules. In other words, it is a structure that is used by applying simple decision-making steps, dividing large amounts of records into very small groups of records. Fig. 1 shows the general structure of the DT algorithm.

2.1.2. k-NN

It uses a variable k to determine the class closest to it. This determined variable k represents the number of k elements closest to the sample. There are different methods (Minkowski, Euclid, etc.) for calculating the distance of a new sample from the classified samples. The most common of these is the Euclidean distance calculation method (Eq. (1)).

$$d(i,j) = \sqrt{\sum_{p=1}^{n} \left(X_{ip} - X_{jp}\right)^2} \tag{1}$$

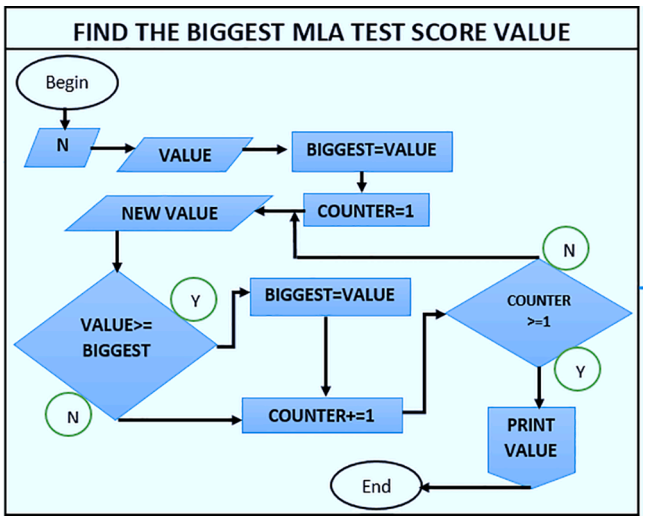


Fig. 10. Finding the biggest MLA score.

where n represents the dimension. i is a new sample (X_{jp}) to be classified, and the nearest k neighbors $X_{ip}(i = 1, 2, \cdots, k)$. Fig. 1 shows the process of classifying a new X_{jp} sample according to $k = 3$ in a two-dimensional ($n = 2$) space.

An example is shown in Fig. 2, since the new object is the closest and the largest number of triangles (Figure a), this object is included in the triangle class (Figure b).

2.1.3. Naïve Bayes

NB classification algorithm is a classification/categorization algorithm named after Mathematician Thomas Bayes. NB classification aims to determine the class, or category, of the data submitted to the system, with a series of calculations defined according to probability principles. Bayes' theorem defines the relationship between a random event that arises from a random process and conditional probabilities and marginal probabilities for another random event as in Eq. (2).

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \tag{2}$$

In Eq. (2), $P(X)$ represents the input probability of the problem $P(Y)$, represents the probability of a possible exit status, and $P(Y|X)$ represents the probability of a Y output versus input X [37].

2.1.4. Support Vector Machine

SVM is a MLA based on convex optimization that works according to the structural risk minimization principle [38]. Fig. 3 shows the SVM classification stages of a data set that cannot be separated linearly.

2.1.5. Multinomial Logistic Regression

Regression analysis determines the statistical relationship between two or more variables that have a cause-effect relationship and makes predictions about the subject using this relationship [39,40]. The nonlinear Logistic Regression (LR) model was designed for two dependent variables [41]. The MLR is used when the dependent variable contains at least three or more categories to explain cause-effect relationships between the dependent (Y) and independent variables (X) [42,43]. Since the goal here is to categorically estimate the value of the dependent variables, it is to estimate "membership" for two or more categories. As a result, we can say that one of the purposes is the classification process and the other is to examine the relationships between dependent and independent variables [44]. In LR, the ratio of probability of occurrence of a p event to the probability of occurrence of other events other than itself is called "odds" or "superiority" value (Eq. (3)). This ratio serves as a function that facilitates the transformation during linearization of the LR model.

$$Odds = \frac{p}{1-p} = \frac{\pi(x)}{1-\pi(x)} \tag{3}$$

LR model is a special form of general linear models obtained for dependent variables as binomial distribution and it is expressed as in Equation (4);

$$\pi(x) = \frac{\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p}{1 + e^{\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p}} \tag{4}$$

where π (x) represents the probability of occurrence of an event under investigation, α the dependent variable constant, $\beta_1, \beta_2, \cdots, \beta_p$ are the regression coefficients of the independent variables, $x_1, x_2, \cdots, x_p$ arguments, p the number of arguments, and e the error term. MLR model, as shown in Equation (5), is an expanded form of the two-state LR model.

$$\pi_j(x_i) = \frac{e^{\alpha_i+\beta_{1j}x_{1i}+\beta_{2j}x_{2i}+\ldots+\beta_{pj}x_p}}{1+\sum_{j=1}^{k-1} e^{\alpha_i+\beta_{1j}x_{1i}+\beta_{2j}x_{2i}+\ldots+\beta_{pj}x_p}} \tag{5}$$

where k represents k categories, $j_1, j_2, \cdots, j_k$, and the n levels of possible arguments $i_1, i_2, \cdots, i_n$.

In this section, information about the performance of machine learning algorithms is presented using CM. It is a matrix model that

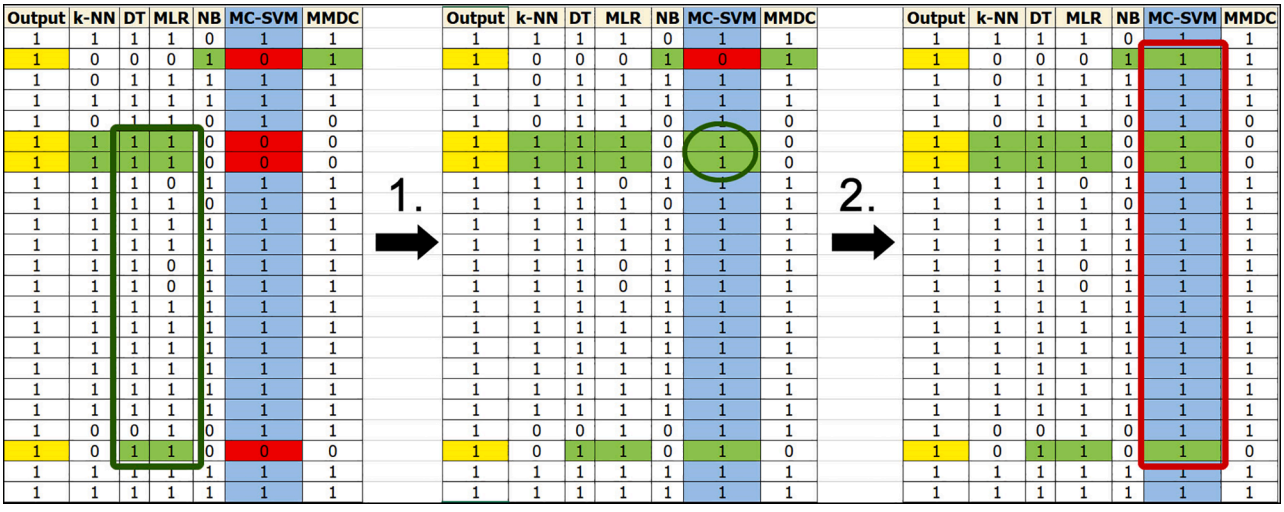


Fig. 11. Repetitive highest test score detection for MLAs.

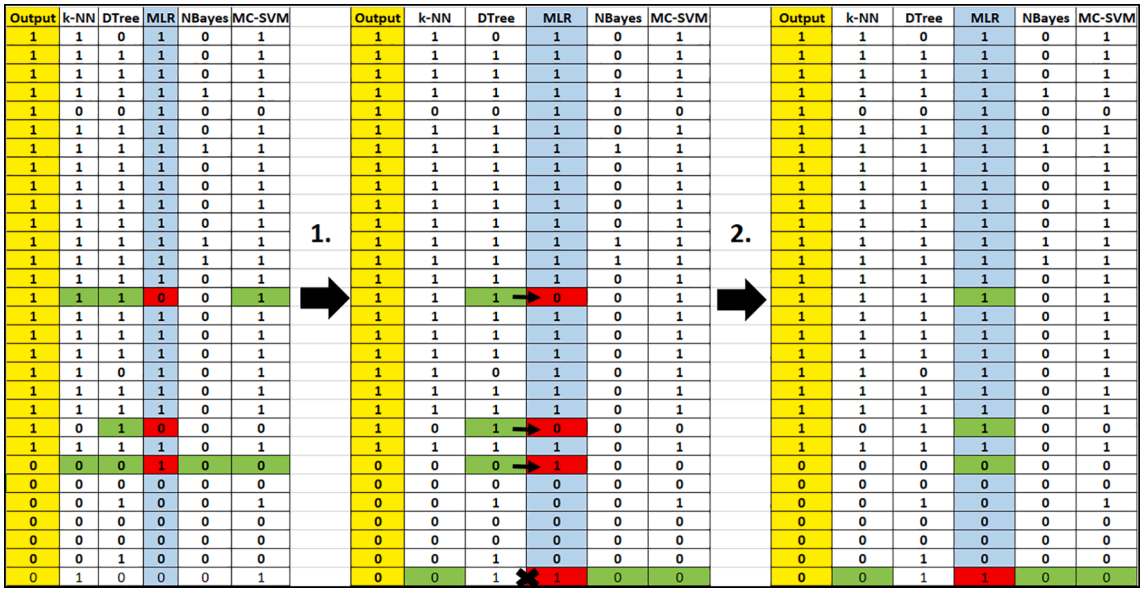


Fig. 12. Obtaining the biggest test score for the COVID-19 image dataset.

Table 1
Parameters used for MLA.

k-NN	Distance Method: Euclidean	k:3
DT	Learning method: C4.5	–
MLR	Estimate method: Gradient Descent	–
NB	Distribution: Gaussian	–
SVM	Kernel: Gaussian	Tolerance:0.001

provides a holistic approach to the classification performance of an intelligent system algorithm. The CM is structurally expressed as in Equation (6).

$$CM = \begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix} \tag{6}$$

In this study, 9 statistical measurement methods were applied. These measurements are shown in Fig. 4.

In this study, a 5-fold crossvalidation process was also performed. Fig. 5. The working structure of the 5-fold crossvalidation process is shown.

2.2. Image preprocessing and feature extraction

The COVID-19 database used in this study was taken from the Kaggle site [45]. The images in the database were first standardized. In order to make the diagnosis more accurate, all images were image preprocessed and cleared of noise. All images were taken with the third row L bone indicated by the yellow arrow in the green rectangle in Fig. 6 as the border. As a result, noises and unnecessary areas around the images were eliminated as shown in Fig. 6.

In addition, all images were cropped in 800 × 800 sizes so that the images in the cage space remain. Using RGB values on the cropped images, the features of the images were extracted. These features were used in Train and Test operations for MLAs.

Table 2
Statistical results for COVID-19 dataset.

	CL0	CL1	TP	FN	FP	TN	TPR	SPC	PPV	NPV	FPR	FNR	ACC	MCC	FM
k-NN															
CL0	21	5	21	5	2	21	0.81	0.91	0.91	0.81	0.09	0.19	0.86	0.72	0.86
CL1	2	21	21	2	5	21	0.91	0.81	0.81	0.91	0.19	0.09	0.86	0.72	0.86
DT															
CL0	19	7	19	7	3	20	0.73	0.87	0.86	0.74	0.13	0.26	0.8	0.6	0.79
CL1	3	20	20	3	7	19	0.87	0.73	0.74	0.86	0.27	0.14	0.8	0.6	0.8
MLR															
CL0	24	2	24	2	2	21	0.92	0.91	0.92	0.91	0.09	0.09	0.92	0.84	0.92
CL1	2	21	21	2	2	24	0.91	0.92	0.91	0.92	0.08	0.08	0.92	0.84	0.91
NB															
CL0	26	0	26	0	19	4	1	0.17	0.58	1	0.83	0	0.61	0.32	0.73
CL1	19	4	4	19	0	26	0.17	1	1	0.58	0	0.42	0.61	0.32	0.3
MC-SVM															
CL0	23	3	23	3	2	21	0.88	0.91	0.92	0.88	0.09	0.12	0.9	0.8	0.9
CL1	2	21	21	2	3	23	0.91	0.88	0.88	0.92	0.12	0.08	0.9	0.8	0.89

Table 3
ROC curve for k-NN.

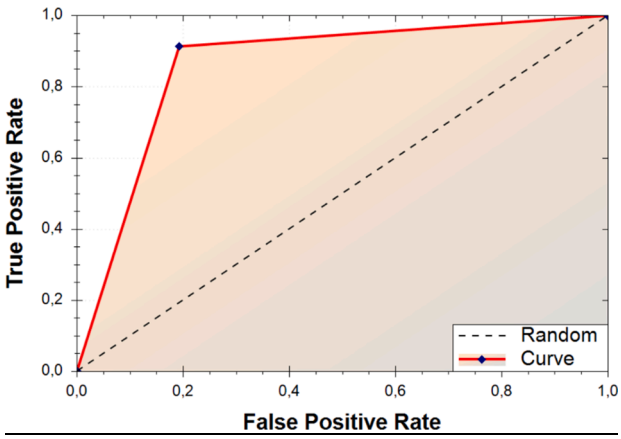


Table 4
ROC curve for DT.

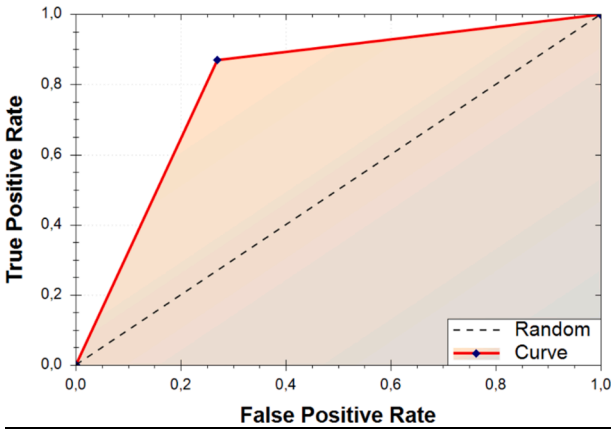
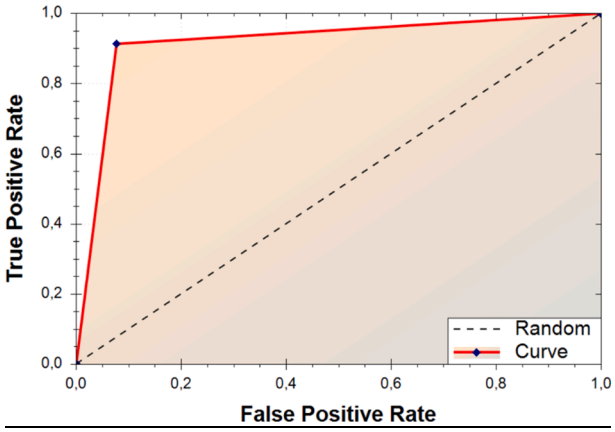


Table 5
ROC curve for MLR.



2.2.1. RGB color mode space

The RGB color system reproduces all colors from the three primary colors blue, green, and red. It is one of the most widely used color spaces in image processing. For images with 8 bits per channel, intensity values range from 0 (black) to 255 (white) for each RGB (red, green, blue) component in a color image. RGB color space can be represented as a three-dimensional Cartesian coordinate system as shown in Fig. 7.

Table 6
ROC curve for NB.

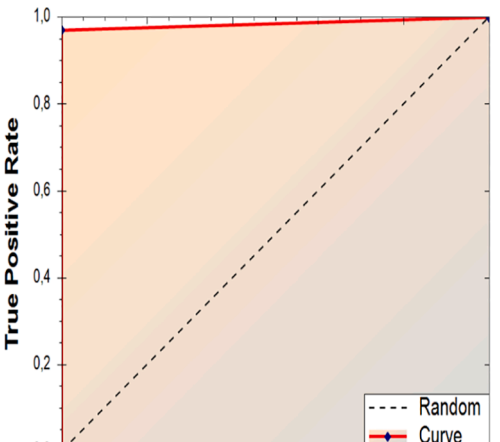
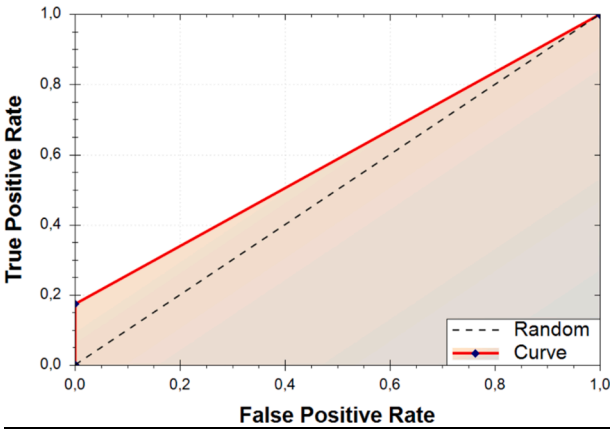


Table 7
ROC curve for SVM.



Three colors or channels are used in RGB images to render colors on the screen. In images of 8 bits per channel, three channels are converted into 24 bits (8 bits × 3 channels) color information for each pixel [46]. All of the images in this study have been resized in 800x800 dimensions.

2.3. Proposed method for the most accurate medical diagnosis

First, an algorithm was developed to obtain RGB values. In this study, the images with 800 × 800 pixels consisted of 640,000 pixels in total. The training process could not be carried out as it would be a big problem to present such many pixels as data for MLAs. For this reason, the pixels of each image were divided into 3200 groups of pixels without losing its features, and a total of 200 features were obtained. For a better understanding of the applied algorithm, the following 10 yellow-colored RGB properties are summed and written as new values in the first excel cell as seen in the red excel cell. Then the second ten RGB values are taken and written into NEW RGB two excel cells. This process continues until all pixels are gone. Then the same feature is applied for all rows as shown in Fig. 8.

In the proposed method, the normal classification process is done first. As a result of the classification process, the most accurate MLA result of the general classification is determined. It is determined by the MLA Majority Voiting method, which has the highest test accuracy. In

Table 8
5 Fold crossvalidation CM results.

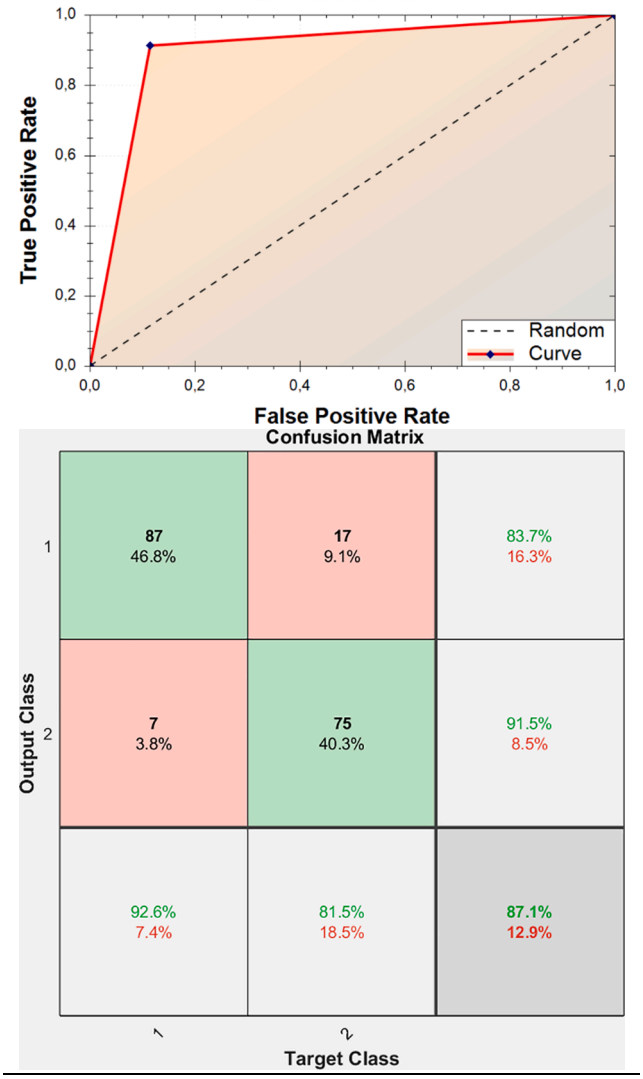


Table 9
Train and test score results.

	Train Score	Test Score
k-NN	100%	85%
DT	100%	79%
MLR	100%	91%
NB	69%	61%
MC-SVM	92%	89%

other words, the classification is determined by the multiplicity of success. The general flow diagram of the proposed method is as shown in Fig. 9.

To determine the largest test score, "A" shown in Fig. 9 operates according to the flow chart shown in Fig. 10. The two algorithms shown in Fig. 10 are run in the section with the blue box indicated by "B".

2.3.1. Majority Voting

In cases where more than one machine learning algorithm is used, the classification result is obtained as in equation (7).

$$f_{MV} = \begin{cases} 1 & \left(\sum_{i=0}^{n} ML_i\right) \geq 2 \\ 0 & otherwise \end{cases} \quad (7)$$

Where ML_i refers to the estimation results (1 or 0) produced by MLAs. The n shown in the formula represents the number of MLAs used in the related field.

The most accurate classification MLA was chosen. After that, the places shown in the red box in Fig. 11 with false test scores for this MLA are determined. Other MLAs are checked for red boxes. Finding correct MLAs are detected as in the green boxes. As the MLA with the highest classification rate from Step 1 to Step 2 is the SVM, the algorithm will select the correct feature classification from it. This process will then repeat the same for every next step. If there is no correct test score left in an MLA during the repetition process, the classification process is terminated. This process depends on the database. Fig. 11 shows an example visual that illustrates this situation.

In Fig. 11, the blue column is the MLA (SVM) that has the highest classification success at the beginning, and the red boxes are the values that SVM incorrectly predicted. These values allow other MLAs to be scanned for the attribute on that row, and the wrong values for SVM are replaced with the correct values. Thus, the diagnosis is made more accurate. This process continues until all of the wrong predictions for SVM are correct, but if other MLAs do not make correct predictions during the iteration, this process ends. As can be seen from the values drawn in the red rectangle after the second step, the actual output values in the yellow box are the same. Thus, the most accurate diagnosis is made.

Fig. 12 shows some of the test scores and corrections obtained for the COVID-19 image dataset in this study. In the first step of the proposed algorithm, the other MLA which is the most correct for the wrong predictions of MLR (with the highest test score) is determined. Since the value predicted incorrectly by the MLR MLA in Fig. 12 is correctly predicted by DT MLA, this will be selected by MLA. In order to determine the MLA to be selected, the MLA with the most accurate diagnosis and the highest test classification is selected. As a result of the algorithm applied, the MLR test success rate has been increased from 91% to 98%.

3. Experimental studies and discussions

The proposed method to make the most accurate medical diagnosis using MLAs in this study was applied for two different datasets (COVID-19 image dataset and Cardiotocography dataset). In this section, the data obtained from the statistical measurements made for these databases are given in order. In this study, 75% of the data was used for train

Table 10
Value of MLR MLA after the proposed method.

	Normal		After proposed method	
	Train Accuracy Score	Test Accuracy Score	Train Accuracy Score	Test Accuracy Score
k-NN	1	0. 85,714	–	–
DT	1	0. 79,591	–	–
MLR	1	0. 91,836	1	0.98
NB	0. 69,565	0. 61,224	–	–
MC-SVM	0. 92,753	0. 89,795	–	–

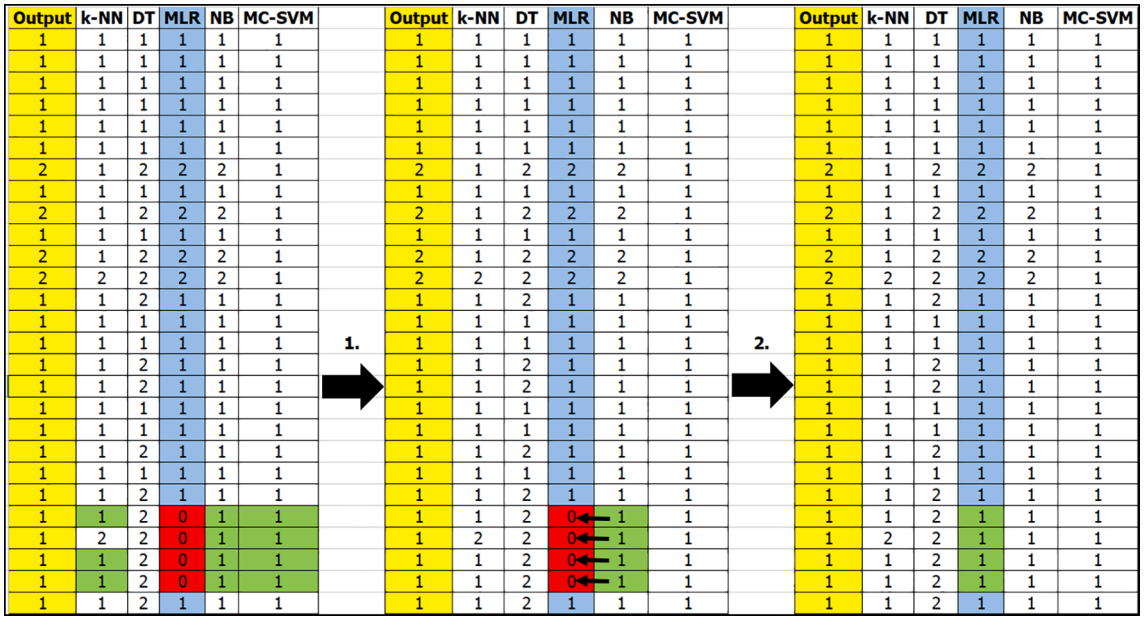


Fig. 13. Application of the proposed method for cardiotocography dataset.

Table 11
Statistical results obtained from Cardiotocography dataset.

k-NN																
	CL0	CL1	CL2	TP	FN	FP	TN	TPR	SPC	PPV	NPV	FPR	FNR	ACC	MCC	FM
CL0	40	45	2	40	47	9	440	0.46	0.98	0.82	0.9	0.02	0.1	0.9	0.56	0.59
CL1	8	391	39	391	47	51	47	0.89	0.48	0.88	0.5	0.52	0.5	0.82	0.38	0.89
CL2	1	6	4	4	7	41	484	0.36	0.92	0.09	0.99	0.08	0.01	0.91	0.15	0.14
DT																
CL0	87	0	0	87	0	0	449	1	1	1	1	0	0	1	1	1
CL1	0	283	155	283	155	1	97	0.65	0.99	1	0.38	0.01	0.62	0.71	0.49	0.78
CL2	0	1	10	10	1	155	370	0.91	0.7	0.06	1	0.3	0	0.71	0.19	0.11
MLR																
CL0	86	0	1	86	1	6	443	0.99	0.99	0.93	1	0.01	0	0.99	0.95	0.96
CL1	6	417	15	417	21	1	97	0.95	0.99	1	0.82	0.01	0.18	0.96	0.88	0.97
CL2	0	1	10	10	1	16	509	0.91	0.97	0.38	1	0.03	0	0.97	0.58	0.54
NB																
CL0	87	0	0	87	0	0	449	1	1	1	1	0	0	1	1	1
CL1	0	416	22	416	22	1	97	0.95	0.99	1	0.82	0.01	0.18	0.96	0.87	0.97
CL2	0	1	10	10	1	22	503	0.91	0.96	0.31	1	0.04	0	0.96	0.52	0.47
MC-SVM																
CL0	64	23	0	64	23	6	443	0.74	0.99	0.91	0.95	0.01	0.05	0.95	0.79	0.82
CL1	6	396	36	396	42	27	71	0.9	0.72	0.94	0.63	0.28	0.37	0.87	0.6	0.92
CL2	0	4	7	7	4	36	489	0.64	0.93	0.16	0.99	0.07	0.01	0.93	0.3	0.26

and 25% for test in experimental studies. The parameters used for MLA are shown in Table 1.

3.1. COVID-19 image dataset

The data obtained as a result of feature extraction from the COVID-19 image database are statistically measured for each MLA. The data obtained from CM are shown in Table 2 (CL: Class, CL0: Image with COVID-19, CL1: Image NOT COVID-19).

ROC curves obtained for MLAs from the COVID-19 image database are shown in Tables 3–7, respectively. Table 3 shows the ROC curve for k-NN.

Table 4 shows the ROC curve for DT.

Table 5 shows the ROC curve for MLR. The ROC curve obtained after the application of the proposed method is also shown.

Table 6 shows the ROC curve for NB.

Table 7 shows the ROC curve for SVM.

The CM results obtained as a result of 5 fold cross validation from MILAs are shown in Table 8.

The highest value was obtained from MLR MLA with 88.7% value. Table 9 shows the train and test scores obtained from MLAs for the COVID-19 dataset.

The highest value was obtained from MLR MLA with 91%. The worst results were obtained from NB MLA. Table 10 shows the train and test Accuracy values obtained from MLAs for COVID-19 dateset.

After the application of the proposed algorithm, the accuracy rate

Table 12
ROC curve for k-NN.

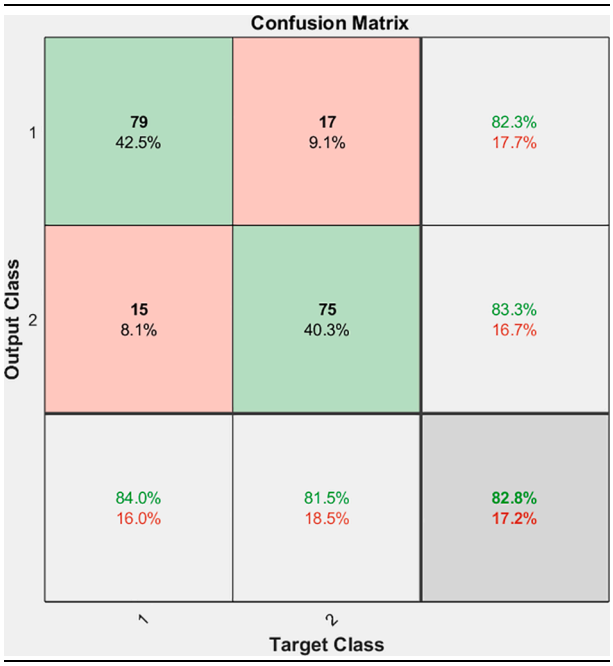
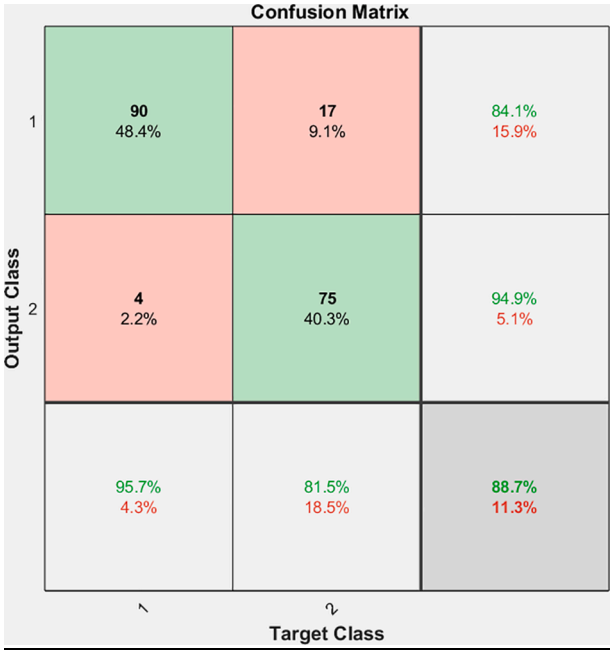


Table 13.R
OC curve for DT.



increased to 0.98 since 1 error could not be corrected.

3.2. Cardiotocography dataset

The proposed algorithm has been tested in a second database [47,48]. It was re-set by MLA, which predicts the wrong values of MLR with the highest test success most correctly. Fig. 13 illustrates part of this process.

In the first step, NB MLA was chosen because it was the NB that gave the most accurate diagnosis to wrong values of MLR. The success rate of the test did not increase to 100%, as the NB MLA gave incorrect answers

Table 14
ROC curve for MLR.

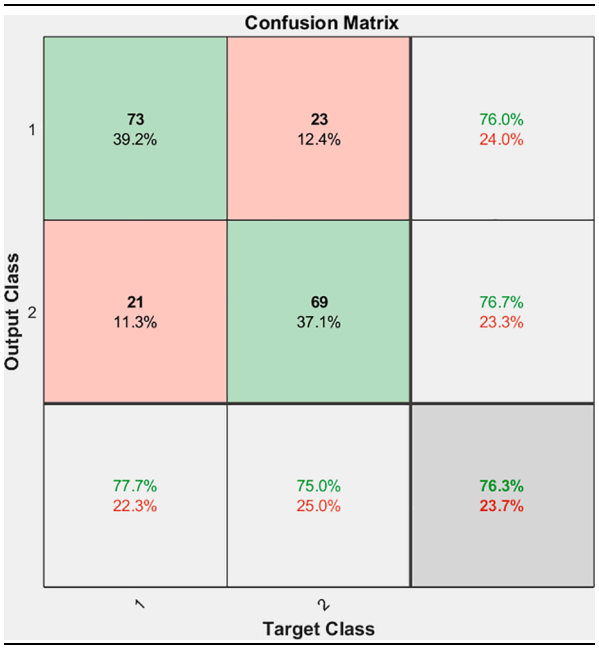
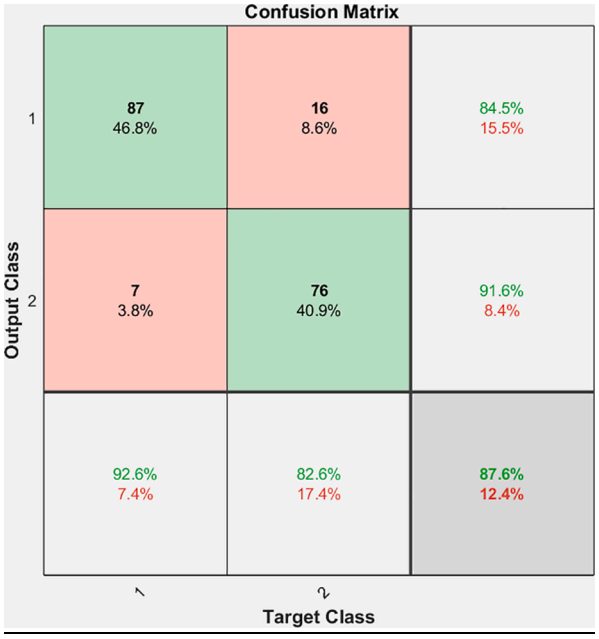


Table 15
ROC curve for NB.



to some diagnoses like MLR. As a result, the value of MLR 'with an accuracy of 0.95 was increased to 0.97. Thus, the most accurate estimation process that can be obtained in this database was carried out.

CM values obtained as a result of statistical measurements applied on cardiotocography dataset are shown in Table 11.

The highest accuracy value was obtained from MLR MLA with 0.99. The ROC curves obtained for MLAs from the cardiotocography database are shown in Tables 12–16. Table 12 shows the ROC curve for k-NN.

Table 13 shows the ROC curve for DT.

Table 14 shows the ROC curve for MLR. The AUC values obtained after the application of the proposed method are also shown.

Table 15 shows the ROC curve for NB.

Table 16
ROC curve for SVM.

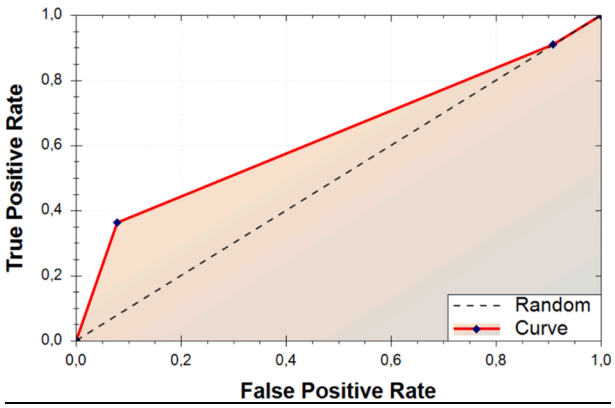


Table 17
5 Fold cross validation result COM results.

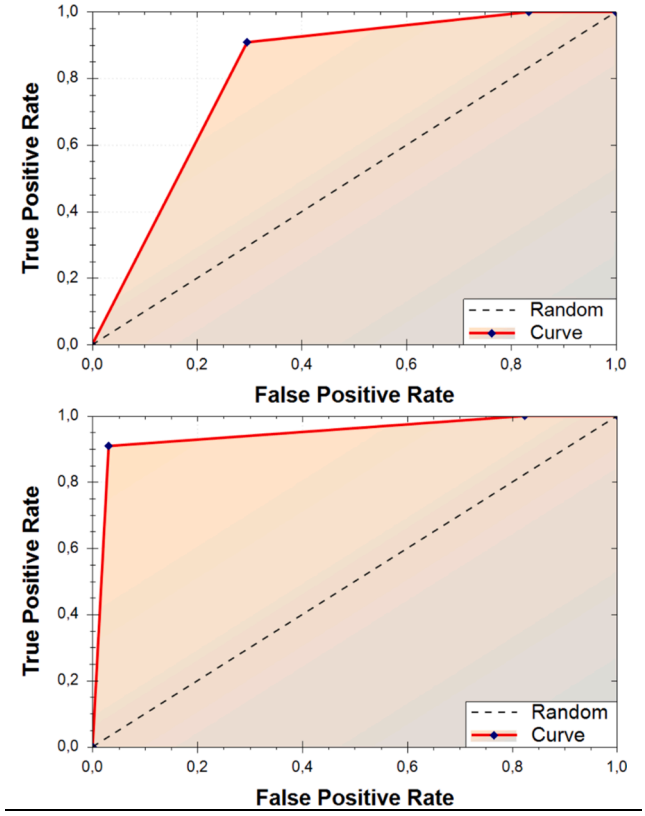


Table 18
Train and test scores for MLAs.

	Train Score	Test Score
k-NN	99%	81%
DT	99%	70%
MLR	99%	95%
NB	97%	95%
MC-SVM	93%	87%

Table 16 shows the ROC curve for SVM.

CM results obtained as a result of 5 fold crossvalidation process from MLAs for cardiotocography dataset are shown in Table 17.

The highest value was obtained from SVM MLA with 98.8% value. The worst results were obtained from the k-NN MLA. Table 18 shows the train and test scores obtained from MLAs for Cardiotocography dateset.

The highest value was obtained from MLR and NB MLAs with 95%. The worst test score was obtained from DT MLA. Table 19 shows the train and test Accuracy values obtained from MILAs for Cardiotocography dataset. Since 1 error could not be corrected, the accuracy rate increased to 0.97.

As a result, the test success rate for COVID-19 image dataset increased from 91% to 98% and for Cardiotocography dataset from 95% to 97%. This increase value is very important, as the tiniest accuracy value in medical diagnosis is vital to the diagnosis. It should be noted that these values may vary for each database.

There are a few points to be considered here, it is possible to list them as follows;

- The proposed method can be used if the number of classes in the database is different.
- After the application of this method, an MLA can make an accurate diagnosis for one, many or all exits. This did not happen in databases, but test success rates can be increased to 100%. This is all about the database.
- If more than one MLA has been diagnosed with high diagnosis, which one will be chosen is determined by first checking the test score accuracy and then the train score accuracy.
- The proposed method can turn very bad results into very good results for some databases, depending on the characteristics of the database.
- It may not be possible to obtain any efficiency using the proposed method. Because if a result that will change the data obtained after the first classification is not in the next iteration, there may not be any test score change.

As a result, the proposed method can be used in studies to be done for the most accurate diagnosis and definition in any field.

4. Conclusions and future works

Since health is the most important value in people's lives, the most accurate medical diagnosis for any disease is of great importance. Likewise, COVID-19, which turned into a pandemic in a short time, is an internationally alarming, sudden and rapidly emerging health problem. For this reason, the most accurate diagnosis of COVID-19 is vital for patients to receive the right treatment. Radiological examinations such as Computed Tomography can be seriously helpful in the diagnosis of the disease in COVID-19 patients. In this study, for more accurate detection of COVID-19, COVID-19 diagnosis was made using the image database. Experimental studies were conducted using k-NN, DT, MLR, NB, SVM MLA using statistical methods. In the results of the train, the following ratios were obtained for k-NN, DT, MLR, NB, SVM, respectively; 100%, 100%, 100%, 69%, 92%. Likewise, the following rates of test success were obtained for k-NN, DT, MLR, NB, SVM, respectively; 85%, 79%, 91%, 61%, 89%. After the application of the proposed method, the test score for MLR was obtained as 98%. It was seen as a result of the experimental results that this feature extraction method could be used in any study in the literature. Performance comparison can be made by applying to databases that have undergone different preprocessing in future image processing methods. The method proposed in this study can be integrated into real medical devices and evaluate images directly, and can provide significant convenience to doctors during the diagnosis of COVID-19.

CRediT authorship contribution statement

Emre Avuçlu: Conceptualization, Methodology, Software, Validation, Investigation, Resources, Writing – review & editing.

Table 19
Train and test accuracy values.

	Normal		After proposed method	
	Train Accuracy Score	Test Accuracy Score	Train Accuracy Score	Test Accuracy Score
k-NN	0.99937	0.8115	–	–
DT	0.99937	0.7089	–	–
MLR	0.99245	0.9570	1	0. 97
NB	0.97106	0.9570	–	–
MC-SVM	0.93584	0.8712	–	–

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Web site, Available at: www.who.int/patientsafety/topics/primary-care/technical_series/en/., Access Date:11.09.2016.
[2] Web site, Available at: Woolever, D. R; The impact of a patient safety program on medical error reporting.2005.
[3] Web site, Available at: www.sciencedaily.com/releases/2016/05/160504085309.htm./., Erişim Tarihi: 11.09.2016.
[4] Web site, Available at: www.usnews.com/news/articles/2016-05-03/medical-errors-are-third- leadingcause-of-death-in-the-us., Access Date:11.10.2016.
[5] Web site, Available at: https://en.wikipedia.org/wiki/Medical_error., Access Date: 12.10.2016.
[6] C. Castaneda, K. Nalley, C. Mannion, P. Bhattacharyya, P. Blake, A. Pecora, A. Goy, K.S. Suh, Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine, J. Clin. Bioinf. 5 (1) (2015).
[7] A. Pannu, Artificial intelligence and its application in different areas, Artif. Intell. 4 (10) (2015).
[8] C. Angeli, Diagnostic expert systems: From expert's knowledge to real-time systems, Adv. Knowl. Based Syst. Model Appl. Res. 1 (2010) 50–73.
[9] C. Çetin, A. Kara, Global surveillance, travel, and trade during a pandemic, Turk. J. Med. Sci. 50 (2020) 527–533.
[10] H. Lu, C.W. Stratton, Y.-W. Tang, Outbreak of pneumonia of unknown etiology in Wuhan, China: the mystery and the miracle, J. Med. Virol. 92 (4) (2020) 401–402.
[11] Zhao S. Musa SS. Lin Q. Ran J. Yang G. Wang W. Et al. Estimating the Unreported Number of Novel Coronavirus (COVID-19) Cases in China in the First Half of January 2020: A Data-Driven Modelling Analysis of the Early. Outbreak. J. Clin. Med. 2020; 9, 388.
[12] Hui DS. Azhar EI. Madani TA. Ntoumi F. Kock R. Dar O. et al. The continuing COVID-19 epidemic threat of novel coronavirusesto global health — The latest 2019 novel coronavirus outbreak inWuhan, China. International Journal of Infectious Diseases 91. 2020;264-266.
[13] Yeni Koronavirüs "COVID-19" Nedir? Available at: https://www.yeditepe.edu.tr/tr/duyuru/yeni-koronavirus-COVID-19-nedir (Access date: 05.04.2020).
[14] Tesini BL. Coronaviruses and Acute Respiratory Syndromes (COVID-19, MERS, and SARS), Available at: https://www.msdmanuals.com/professional/infectious-diseases/respiratory viruses/coronaviruses-and-acute-respiratory-syndromes-COVID-19,-mers,-and-sars (Access date: 15.04.2020).
[15] Koronavirüs neden hızlı yayıldı? Corona Virüsü. Available at: https:// www.e-psikiyatri.com/koronavirus-neden-hizli-yayildi-corona-virusu (Access date: 15.4.2020).
[16] C.X. Deng, The global battle against SARS-CoV-2 and COVID-19, Int. J. Biol. Sci. 16 (10) (2020) 1676–1677.
[17] V.D. Menachery, B.L. Yount, K. Debbink, S. Agnihothram, L.E. Gralinski, J. A. Plante, R.L. Graham, T. Scobey, X.-Y. Ge, E.F. Donaldson, S.H. Randell, A. Lanzavecchia, W.A. Marasco, Z.-L. Shi, R.S. Baric, A SARS-like cluster of circulating bat coronaviruses shows potential for human emergence, Nat. Med. 21 (12) (2015) 1508–1513.
[18] S.-L. Liu, L.J. Saif, S.R. Weiss, L. Su, No credible evidence supporting claims of the laboratory engineering of SARS-CoV-2, Emerg. Microbes Infect. 9 (1) (2020) 505–507.
[19] Germany's 2012 Covid scenario became real in 2020. Available at: https:// www.foreigner.fi/articulo/news/germany-s-2012-covidscenario-became-real/20200325014404004958.html (Access date: 05.15.2020).
[20] D. Bundestag, 17. Wahlperiode, Drucksache 17/12051. 03 (01) (2013) 2–88.
[21] 3.4% Mortality Rate estimate by the World Health Organization (WHO) as of March 3 2020. https://www.worldometers. info/coronavirus/coronavirus-death-rate/#correct (Access date: 14.05.2020).
[22] Coronavirus: case fatality rates by age. Available at: https://ourworldindata.org/uploads/2020/03/COVID-CFR-by-age-1536x1190.png (Access date: 06.06.2020).
[23] Covid-19 Coronavirus Pandemic (5.15.2020). Available at: https://www. worldometers.info/coronavirus/ (Access date: 15.04.2020).
[24] VanderWeele TJ. Chen Y. Long K. Kim ES. Trudel-Fitzgerald C. Kubzansky LD. Positive Epidemiology? Epidemiology. 2020; (31),2;189-192.
[25] D. Morens, G. Folkers, A. Fauci, What is a pandemic? J. Infect. Dis. 200 (7) (2009) 1018–1021.
[26] Trump keeps touting an unproven coronavirus treatment. It's now being tested on thousands in New York. https://www. washingtonpost.com/business/2020/03/26/trump-keepstouting-an-unproven-coronavirus-treatment-its-now-beingtested-thousands-new-york/ (Access date: 16.05.2020).
[27] Trump touted hydroxychloroquine as a cure for COVID-19. Don't believe the hype. https://www.theguardian.com/ world/2020/mar/28/coronavirus-cure-fact-check-hydroxychloroquine-trump (Access date: 19.04.2020).
[28] Hu Y, Sun J, Dai Z, et al. Prevalence and severity of corona virus disease 2019 (COVID-19): A systematic review and meta-analysis. Journal of clinical virology. 2020;127:104371.
[29] Favalli EG, Agape E, Caporali R. Incidence and Clinical Course of COVID-19 in Patients with Connective Tissue Diseases: A Descriptive Observational Analysis. The Journal of rheumatology. 2020; https://doi.org/10.3899/jrheum.200507.
[30] T. Tuncer, E. Aydemir, F. Özyurt, S. Dogan, S.B. Belhaouarı, E. Akbal, An automated Covid-19 respiratory sound classification method based on novel local symmetric Euclidean distance pattern and ReliefF iterative MRMR feature selector, Int. Adv. Res. Eng. J. 5 (3) (2021) 334–343, https://doi.org/10.35860/iarej.898830.
[31] T. Tuncer, F. Ozyurt, S. Dogan, A. Subasi, A novel Covid-19 and pneumonia classification method based on F-transform, Chemom. Intell. Labor. Syst. 210 (2021) 104256.
[32] T. Tuncer, E. Akbal, E. Aydemir, S.B. Belhaouarı, S. Dogan, A novel local feature generation technique based sound classification method for Covid-19 detection using lung breathing sound, Eur. J. Tech. (EJT) 11 (2) (2021) 165–174, https://doi.org/10.36222/ejt.986599.
[33] Narin Aslan, Gonca Ozmen Koca, Mehmet Ali Kobat, Sengul Dogan, Multi-classification deep CNN model for diagnosing COVID-19 using iterative neighborhood component analysis and iterative ReliefF feature selection techniques with X-ray images, Chemometrics and Intelligent Laboratory Systems, Volume 224,2022, 104539, ISSN 0169-7439, https://doi.org/10.1016/j.chemolab.2022.104539.
[34] S. Wang, M.A. Khan, V. Govindaraj, S.L. Fernandes, Z. Zhu, et al., Deep rank-based average pooling network for covid-19 recognition, Comput. Mater. Continua 70 (2) (2022) 2797–2813.
[35] Y.-D. Zhang, Z. Zhang, X. Zhang, S.-H. Wang, MIDCAN: a multiple input deep convolutional attention network for Covid-19 diagnosis based on chest CT and chest X-ray, Pattern Recogn. Lett. 150 (2021) 8–16.
[36] S.-H. Wang, Y.-D. Zhang, M. Yang, B. Liu, J. Ramirez, J.M. Gorriz, Unilateral sensorineural hearing loss identification based on double-density dual-tree complex wavelet transform and multinomial logistic regression, ICA 26 (4) (2019) 411–426.
[37] U. Orhan, K. Adem, "The Effects of Probability Factors in Naive Bayes Method", Elektrik-Elektronik ve Bilgisayar Mühendisliği Sempozyumu, Bursa 722–724 (2012).
[38] Soman, K. P., Loganathan, R. and Ajay, V., "Machine learning with SVM and other kernel methods", PHI Learning Pvt. Ltd., Delhi/India, 1-10 (2009).
[39] F.E. Harrell, Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression and Survival Analysis, Springer-Verlag, New York, 2001, pp. 215–267.
[40] Hosmer, D. W., Lemeshow, S. and Sturdivant, R. X., "Applied Logistic Regression 3rd Ed.", Wiley&Sons Publications, Canada, 8-35 (2013).
[41] Stock, J. H. and Watson, M. W., "Introduction to Econometrics 2nd Ed.", Addison-Wesley, Boston, 389-390 (2007).
[42] Leech, N. L., Barrett, K. C. and Morgan, G. A., "SPSS For Intermediate Statistics: Use and Interpretation 2nd Ed.", Lawrance Erlbaum Associates Publishers, New Jersey, 109-110 (2004).
[43] E. Arı, Z. Yıldız, Parallel lines assumption in ordinal logistic regression and analysis approaches, Int. Interdiscipl. J. Sci. Res. 1 (3) (2013) 8–23.
[44] Büyüköztürk, Ş., Çokluk Bökeoğlu, Ö. ve Şekercioğlu, G., "Sosyal Bilimler İçin Çok Değişkenli İstatistik SPSS ve LISREL Uygulamaları", Pegem Akademi Publishing, Ankara, 59-65 (2010).
[45] Web site, COVID-19 Xray Dataset (Train & Test Sets). Accessed date: (2020, 10, 19). Retrieved from Kaggle.com: https://www.kaggle.com/khoongweihao/covid19-xray-dataset-train-test-sets.

[46] Web site, Available at: https://tr.wikipedia-on-ipfs.org/wiki/RGB_renk_uzay%C4%B1.html, Accessed date [14.05.2020].
[47] A. de Campos, et al., SisPorto 2.0 A program for automated analysis of cardiotocograms, J Matern Fetal Med 5 (2000) 311–318.
[48] Web site, Available at: https://archive.ics.uci.edu/ml/datasets/cardiotocography, Accessed date [19.09.2020].