Research Article GU J Sci 37(1): 169-181 (2024) DOI: 10.35378/gujs.1150388

Gazi University

Journal of Science

http://dergipark.gov.tr/gujs

Detection of COVID-19 Disease with Machine Learning Algorithms from CT Images

Mahmut Nedim EKERSULAR*, Ahmet ALKAN

Kahramanmaras Sutcu Imam University, Faculty of Engineering and Architecture, Department of Electrical and Electronics Engineering, 46000, Kahramanmaras, Turkey

Highlights

- This paper focuses on the detection of COVID-19.
- LBP features were used to train machine learning models in the study.
- The high classification accuracy obtained with the test images shows that the model is reliable.

Article Info

Received: 28 July 2022
Accepted:14 Feb 2023

Keywords

Classification
Machine learning
Local binary pattern
Computerized tomography
COVID-19

Abstract

COVID-19, caused by the SARS-COV-2 virus, which has killed more than 6 million people, is one of the most contagious diseases in human history. It has seriously affected every area that people come into contact with, from business life to economy, from transportation to education, from social life to psychology. Although the developed vaccines provide a partial decrease in the number of deaths, the mutations that the virus constantly undergoes and the increase in the transmission rate accordingly reduce the effectiveness of the vaccines, and the number of deaths tends to increase as the number of infected people. It is undoubtedly important that the detection of this epidemic disease, which is the biggest crisis that humanity has experienced in the last century after World War II, is carried out accurately and quickly. In this study, a machine learning-based artificial intelligence method has been proposed for the detection of COVID-19 from computed tomography images. The features of images with two classes are extracted using the Local Binary Pattern. The images reserved for training in the dataset were used for training machine learning models. Trained models were tested with previously unused test images. While the Fine K-Nearest Neighbors model reached the highest accuracy with a value of 0.984 for the training images, the highest accuracy value was obtained by the Cubic Support Vector Machine with 0.93 for the test images. These results are higher than the deep learning-based study using the same data set.

1. INTRODUCTION

The COVID-19 disease, which turned into a global epidemic with the detection of the first official case in China in the last months of 2019, and spreading all over the world after a short time, continues to affect humanity. The family of viruses called coronaviruses is mostly found in animals. It has not yet been determined how SARS-COV-2, the seventh type of coronavirus that infects humans, was first transmitted to humans [1]. This virus, which usually manifests itself with symptoms of fever, cough, and shortness of breath, can result in pneumonia, severe respiratory failure, kidney failure, and death in severe cases [2]. The spread of the virus between people is transmitted by air, it is transmitted very quickly and infects many people, and deaths increase accordingly. As of July 2022, 574 million people were infected with this virus, and more than 6 million people died [3]. Although there has been a partial decrease in the number of cases and deaths with the introduction of various types of vaccines, the very rapid mutation of the virus and the increase in its contagiousness with each mutation have led to an increase in the number of cases and deaths again. These situations show that the disease will affect humanity for a while. Although vaccines are the most important tools in ending the epidemic, the inability to apply enough vaccines to reach herd immunity and the decrease in the protection of vaccines due to mutations reminds us that mask, social distance, and hand cleaning rules are still the most important protection methods.

*Corresponding author, e-mail: mahmutnedimekersular@ksu.edu.tr

The worldwide accepted test to detect whether the SARS-COV-2 virus exists in the human body is RT-PCR (Reverse Transcription-Polymerase Chain Reaction). The presence of the genetic material of the virus is sought in the swabs taken from the nose and mouth of the person, and if detected, the test will be positive [4]. However, due to the high false-negative rate of the test, the long duration of the results, and the inability to catch the genetic material of the virus in the early stages of the disease, the doctor evaluates the lung tomography / X-ray and the clinical condition of the patient together to make the diagnosis.

The fact that COVID-19 especially affects the lungs and causes pneumonia by destroying the lung parenchyma cells reveals the importance of radiological imaging to evaluate the condition of the lungs. Understanding the extent of the damage of the disease on the lungs is only possible with such imaging methods. The presence of ground glass opacities in images is characterized by COVID-19 [5].

Computed tomography refers to a computerized x-ray imaging procedure; where a narrow x-ray is aimed at the patient and quickly rotated around the body to generate a signal. These signals are processed in the computer, creating cross-sectional images of the body, or in other words slices. These slices are called tomographic images and contain more detailed information than conventional x-rays. After several slices are collected in succession by the system, they can be digitally stacked to create a three-dimensional image that allows for easier detection and localization of possible tumoral structures or abnormalities in addition to the patient's basic structures. Unlike conventional x-ray, which uses a stationary x-ray tube, the CT scanner uses a motorized (moving) X-ray source that rotates around the circular opening of the annular structure called the gantry. During the CT scan, the patient lays down on a slowly moving bed with the x-ray tube rotating around the patient and the body shoots narrow x-rays. Instead of film, CT scanners use special digital x-ray sensors placed directly opposite the x-ray source. As the x-rays leave the patient, they are collected by sensors and transmitted to the computer [6].

Evaluation of a large number of images obtained with the experience of the specialist creates both a time cost for the doctor and the possibility of affecting the accuracy of the result depending on the doctor's current attention. Since early initiation of treatment is critical in COVID-19 disease, the appropriate time window for treatment may be lost due to the misdiagnosis of the patient. For this reason, it would be a rational approach to use computer technology in the classification of images. Accordingly, machine learning-based approaches can be used when classifying computed tomography or x-ray images. In [7], CT lung images obtained through the Kaggle platform were classified as non-Covid-Covid using the CNN-based classifier. 80% of the images were used for training and the rest for testing, and a success rate of 94.21% was achieved [8]. Rasheed et al. machine learning (Logistic Regression) and deep learning classifiers (CNN) used X-Ray images to classify as Covid- not Covid. The number of images in the data set GAN (Generative Adversarial Network) and performance results were measured with and without PCA (Principal Component Analysis). The accuracy values measured without feature extraction were found to be 95.2% and 97.6% for Linear Regression and CNN models, respectively. When the feature extraction process is included, accuracy values of 100% for CNN and 95.2% for LR were achieved at 0.99 variance. In the study in which five different classifiers were included [9] Covid -Normal, Covid -Bacterial Pneumonia, Covid -Viral Classifier performances were measured according to pneumonia status. XGB-L was determined as the most successful classifier and Covid-Viral in this classifier. An accuracy of 97.56% was found for the pneumonia condition. In [10], where a semi-supervised learning method was used, it was tried to diagnose COVID-19 from CT images. Since a limited number of labeled data is used, the Mixmatch method has been developed to prevent the model from memorizing. The images obtained by mixing the few labeled images and many unlabeled images are given as input to the semi-supervised learning algorithm, while the labeled images are presented to the supervised learning model. The estimation of the classes was made by the ensemble method and an accuracy rate of 0.90 was found. In the study of Baghdadi et al. [11], two data sets consisting of computed tomography images, one with two classes and the other with three classes, were used. Before using the datasets, preprocessing steps such as resizing, scaling, and replication to equalize the number of images was passed. Sparrow for hyperparameter optimization of pre-trained networks algorithm was used and the most optimal network was tried to be determined by comparing various transfer learning networks with each other. As a result of the study, MobileNetV3Large was found as the most suitable network for the two-class dataset, while SeNet154 gave the best result for the three-class dataset. Accuracy rates were 99.74% for MobileNetV3Large and 98% for SeNet154.

The remaining part of the study is organized as Section 2, where explanations are made about the data set and the methods used, and Section 3, where the application results and evaluations are made.

2. MATERIAL METHOD

2.1. Dataset

An open access dataset was used in the study [12]. CT images in the dataset were collected from hospitals in Sao Paulo, Brazil. It consists of 1252 images of Sars-Cov-2 positive and 1229 images of Sars-Cov-2 negative, and the total number of images is 2481. Other properties of the images in the data set are given in Table 1.

Table 1. *Properties of images in the dataset*

	COVID	NORMAL	**TOTAL**
Number of Images	1252	1229	2481
Image Size	Variable		
Bit Depth	24-bit (8-bit x 3)		
Image File Format	PNG		

Sample images from both classes of the data set used in the study are shown in Figure 1.

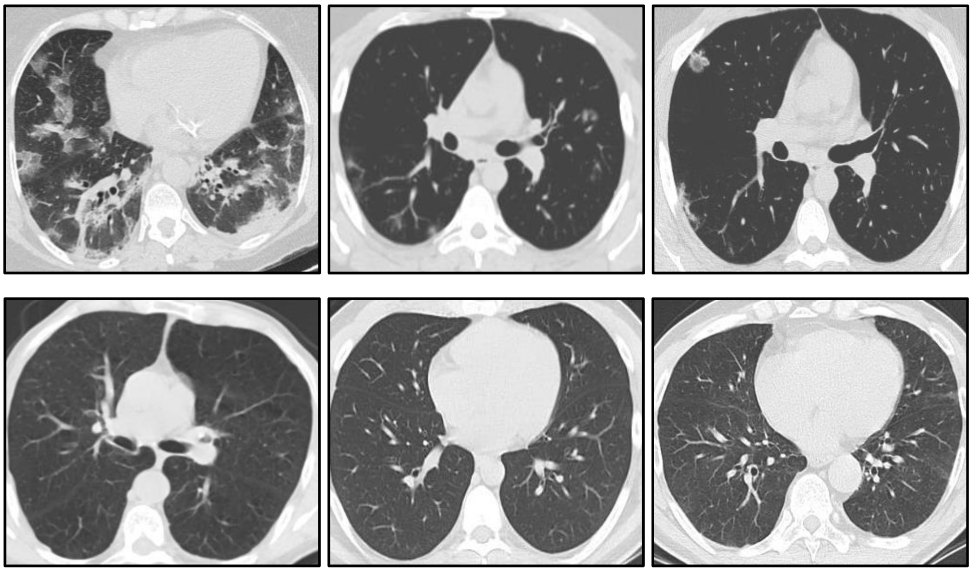


Figure 1. *Sample images from the dataset Top Row: Covid, Bottom Row: Normal*

The flow followed in the study is shown in Figure 2.

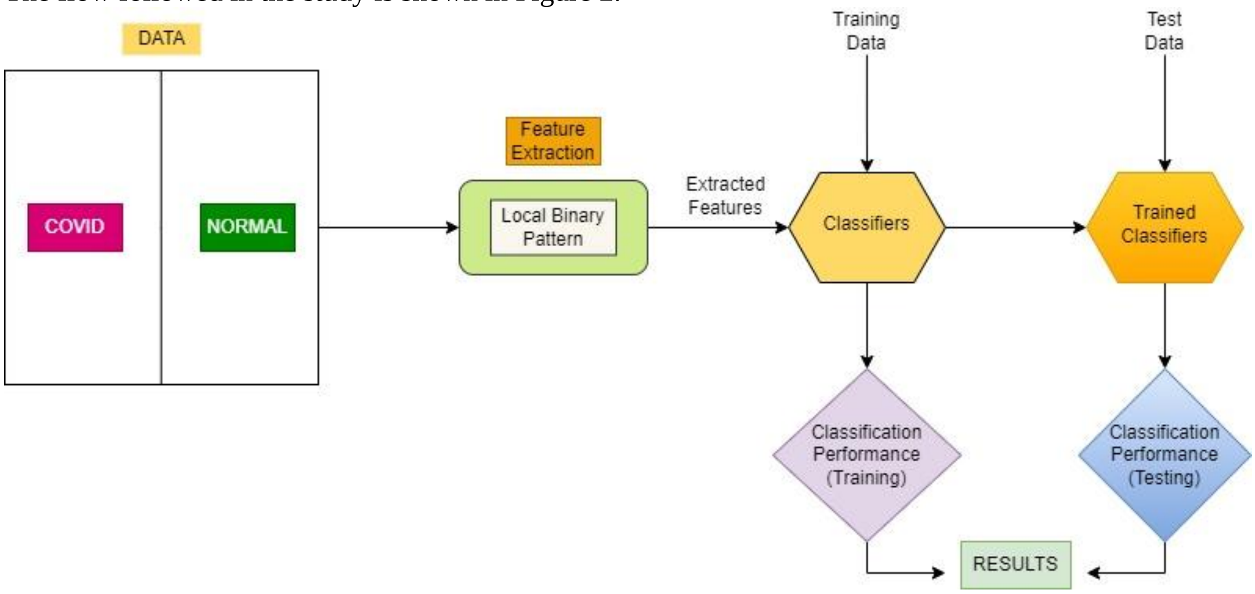


Figure 2. *Flow chart of the study*

2.2. Feature Extraction

2.2.1. Local Binary Pattern

Local Binary Pattern (LBP), which was first introduced to the literature by Ojala et al., is a successful texture scanning operator. In the original LBP operator, the pixel in the middle of the 3 x 3 window is selected as the threshold value. The remaining eight pixels are compared against this threshold, and then binary values are assigned. The resulting binary sequence of numbers is called the LBP code, which identifies distinctive features in the image such as edges, corners, light and dark regions [13]. Figure 3 shows how the LBP operator works for a 3*3 matrix with 45 in its center.

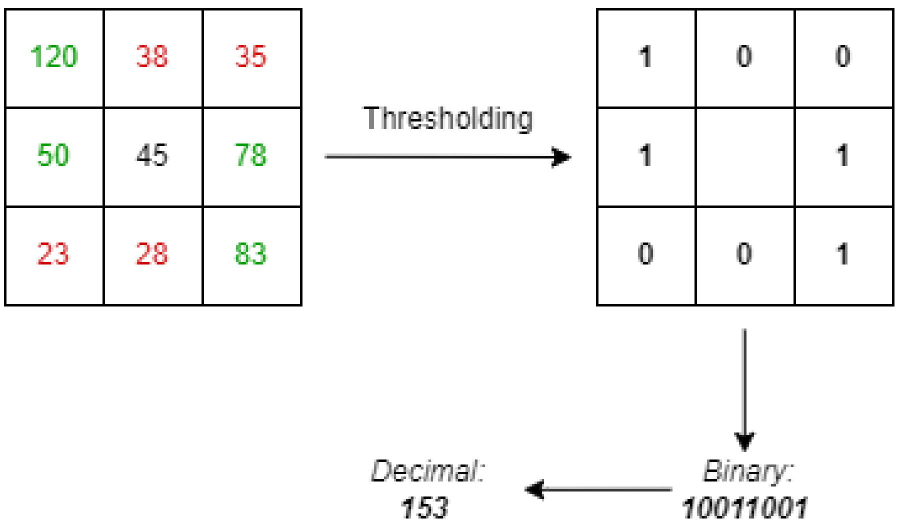
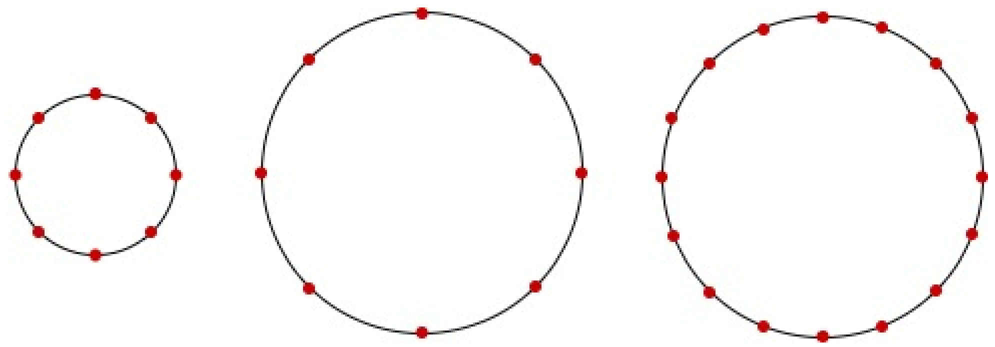


Figure 3. *Original LBP operator*

Because the original LBP uses a 3x3 quadratic neighborhood, it cannot extract features independent of the rotation of the image. Therefore, the method was developed, and circular topologies with different radii and several neighbors were used. By using the circular topology, it provided a better analysis of textures of different sizes in the image. Different neighborhoods for the circular topology are shown in Figure 4.



a b c

Figure 4. *Circular topology. (a) pattern with (8,1) neighborhood, (b) pattern with (8,2) neighborhood, (c) pattern with (16,2) neighborhood*

As in the original LBP, the LBP codes of the central pixels with circular neighborhoods are generated by the formula given below (Equation (1)):

$$LBP_{P,R} = \sum_{P=0}^{P-1} s(g_p - g_c)2^P , \qquad s(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x \leq 0 \end{cases}. \tag{1}$$

Here g_c and g_p are the center pixel and intensity value of pixel p, respectively. After generating the LBP codes for each pixel, the histograms of the codes obtained are used as the feature vector.

2.3. Classifiers

2.3.1. Linear Discriminant

Linear discriminant analysis (LDA) is based on the assumption that all classes can be separated linearly. From this point of view, multiple LDA functions representing hyperplanes in the feature space are created to separate the classes from each other [14].

Although generally used to classify features between two different classes, linear discriminant analysis can also be used to separate datasets with more than two classes.

For a data set with two classes, the data for each class is separated by a hyperplane, maximizing the separation of these two classes. This hyperplane is created according to the criteria of maximizing the distance between the means of two classes and minimizing the variation between each category.

LDA is a size reduction technique as well as being used as a classifier.

2.3.2. Ensemble

Ensemble learning is a method that aims to improve result accuracy by combining several machine learning models. This approach provides better prediction performance compared to a single model [15].

2.3.3. Decision tree

The decision tree is a machine learning algorithm that can be used in classification and regression problems. Decision trees, which are an iterative method, proceed in nodes, as the name suggests, similar to the branching of trees. The nodes start with the root, branch out, and reach the leaves that produce results.

According to the response from each node, the algorithm proceeds by branching differently [16]. Figure 5 shows a representation of decision trees.

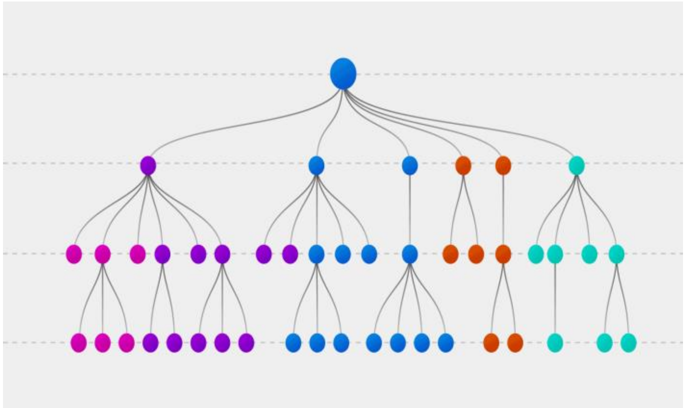


Figure 5. *Decision tree*

2.3.4. Support Vector Machine

In the support vector machine, the goal is to classify the input set, which is considered to be the x-dimensional feature vector, by trying to find the x-1 dimensional hyperplane that maximizes the minimum distance between any data point. [17]. The hyperplane with the largest distance between the two classes is best for SVM [18]. Figure 6 shows the optimal hyperplane separating the two classes.

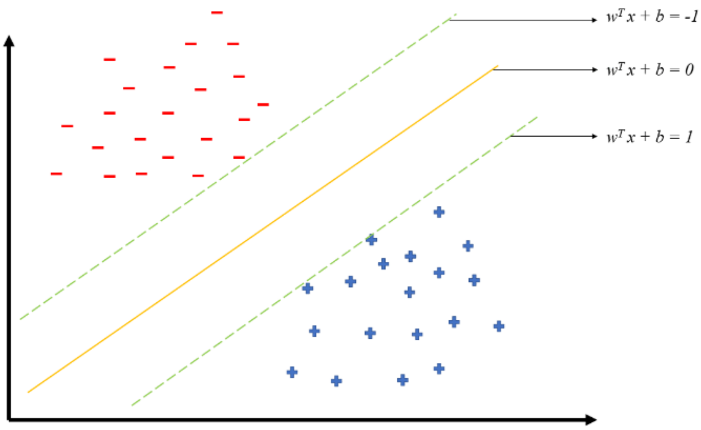


Figure 6. *Support vector machine*

In SVM, the goal is to find the variables *w* and *b* for the line that separates the two classes.

2.3.5. K-Nearest Neighbors

K-Nearest Neighbor, a non-parametric classification method first developed in 1951, is used for regression problems as well as classification. The main goal of the KNN algorithm is to search for the closest points to the new point. The K value indicates the nearest neighbors to the unknown point. Euclidean is often used as a distance function when finding nearest neighbors [19]. The distance between two points is found with the Euclidean function as (Equation (2)):

$$d = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}\,. \tag{2}$$

Here $\boldsymbol{x_i}$ and $\boldsymbol{y_i}$ represents the points, while k is the nearest neighbor number.

2.3.6. Neural networks

A neural network is an artificial representation of the human brain that stimulates the learning process [20]. Neural network-based classification uses the learning strategies of the human brain for decision-making. Artificial neural networks consist of many cells and layers. The weights of the intercellular connections are given randomly at the beginning, and the network reviews and adjusts these weights every cycle [21]. Figure 7 shows that there are many layers between the input and output.

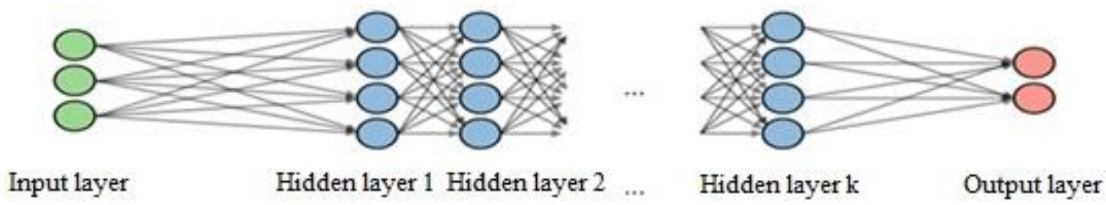


Figure 7. *Deep neural network*

3. EXPERIMENTAL RESULTS

In this part of the study, classification performances for the data set used are given separately for both training and test images. To evaluate the classification performance, confusion matrices, Receiver Operating Characteristic (ROC) curves, and related performance metrics were used.

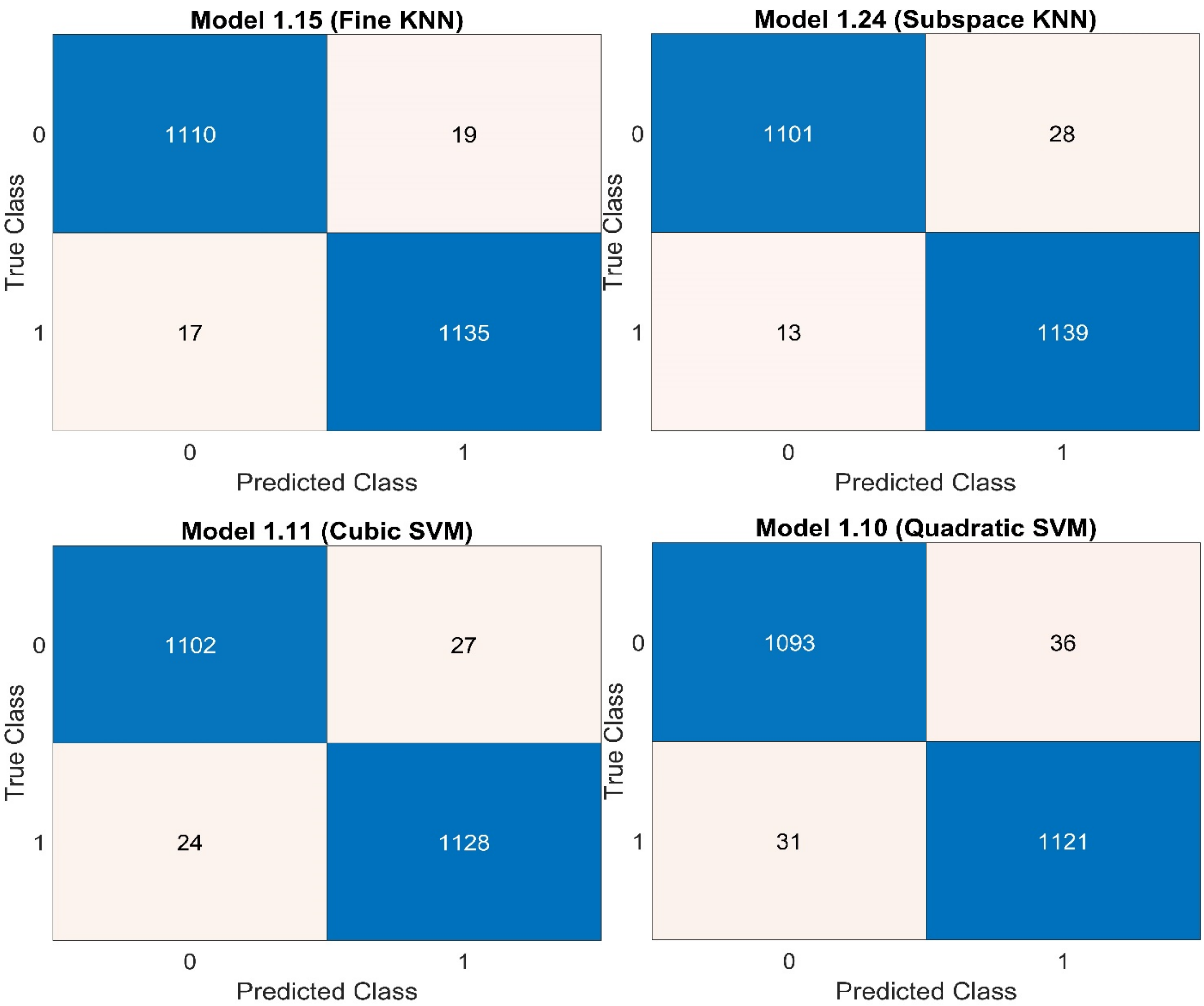


Figure 8. *The four most successful classifiers for training images (from left to right, in descending order)*

A portion of a total of 2481 images in the data set was reserved as the test set. This separation was such that for every 100 images in each class (not Covid-Covid), 10 were reserved for testing. Thus, 200 images were obtained for two classes in the test dataset (100 Covid-100 not Covid).

Figure 8 shows the confusion matrices of the four most successful classifiers for training images. The most successful classifiers for the training dataset were obtained as Fine KNN, Ensemble (Subspace KNN), Cubic SVM, and Quadratic SVM, respectively. 10-fold cross-validation was performed for all classifiers.

Fine KNN, the most successful classifier for training images, predicted 1135 of 1152 Covid lung images as having Covid and misclassified 17 as non-Covid. In addition, it labeled 1110 of 1129 normal lung images as normal and made a wrong prediction, considering 19 of them as Covid.

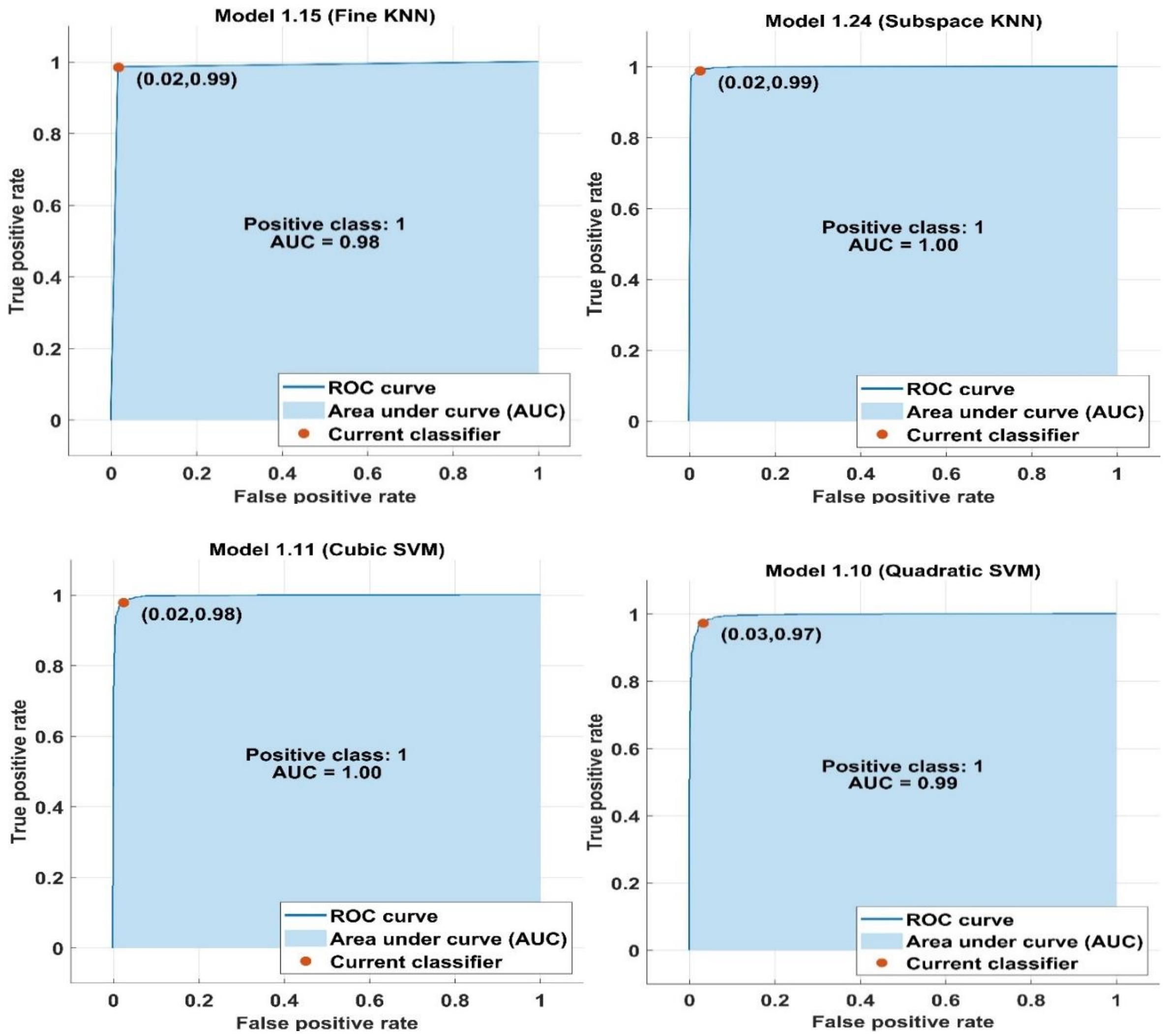


Figure 9. *ROC curves of the top four most successful classifiers for training images*

The curve that shows how the true positive rate (TPR) changes with the false positive rate (FPR) is called the ROC curve. The region under the curve is known as AUC (Area Under Curve), and the convergence of this area to 1 indicates that the classification performance increases [22]. Figure 9 shows the ROC graphs of the four models that made the most successful classification for the training images. The AUC for the Fine KNN, Ensemble (Subspace KNN), Cubic SVM, and Quadratic SVM classifiers were respectively: 0.98, 1.00, 1.00, 0.99.

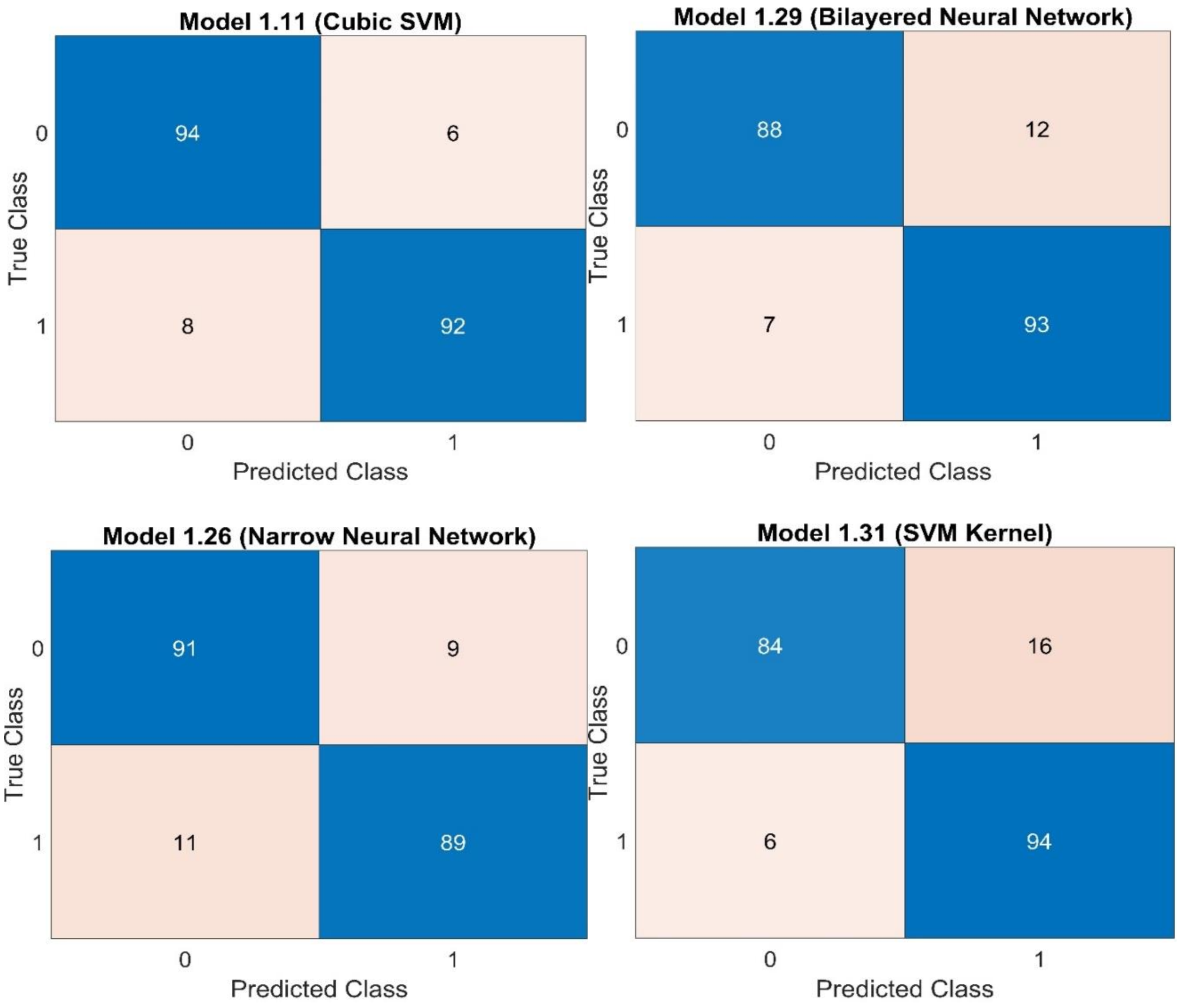


Figure 10. *The four most successful classifiers for test images (from left to right, in descending order)*

The confusion matrices of the four classifiers that made the most successful classification prediction in the images used for the test are given in Figure 10. Cubic SVM classifier, which reached 93% accuracy, labeled 94 of 100 normal lung images as normal and mislabeled 6 of them as Covid. Again, the same classifier estimated 8 of 100 Covid lung images as not Covid, but 92 as Covid. The ROC graphs of the classifiers that achieved the highest accuracy for the test images are given in Figure 11.

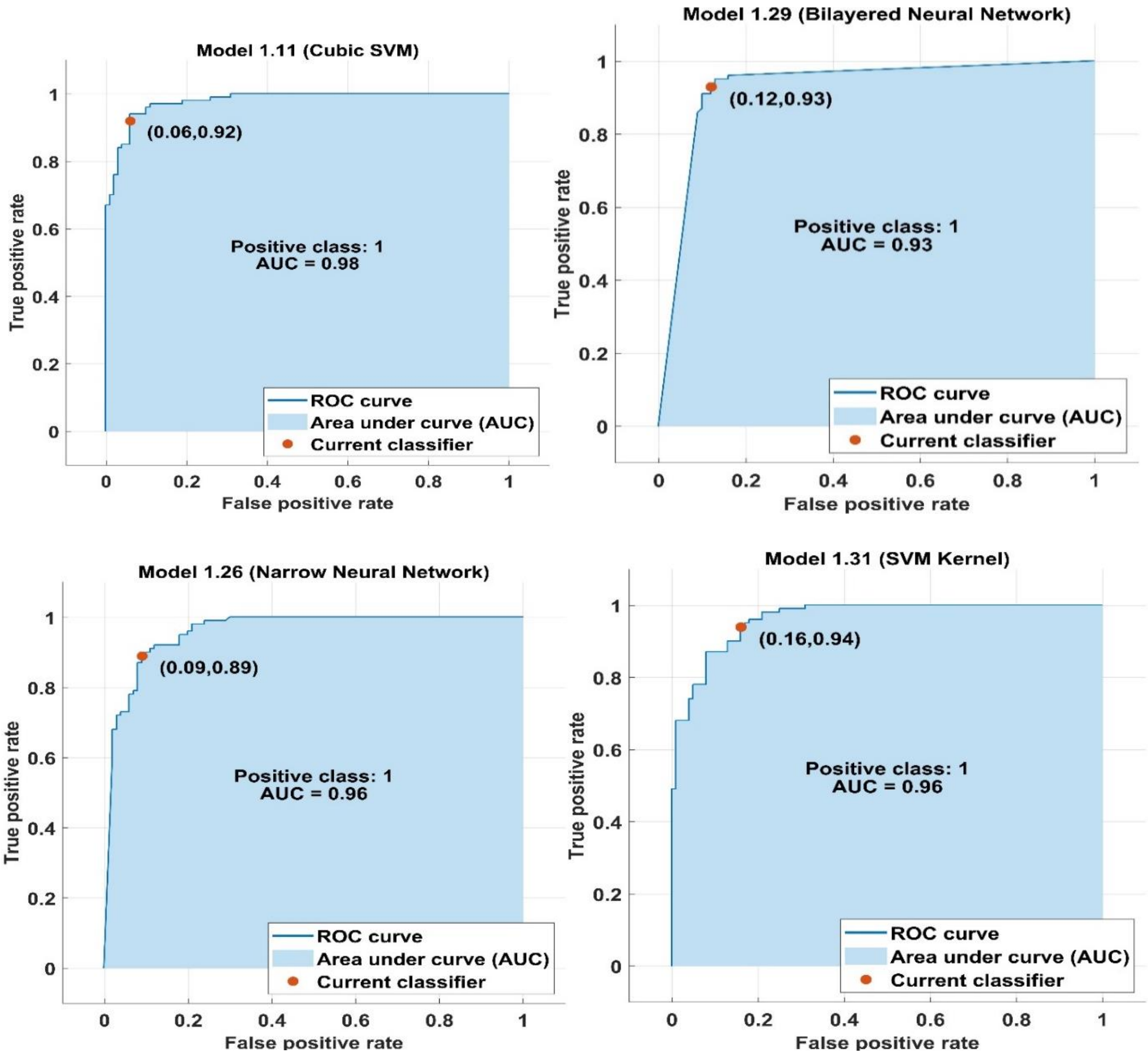


Figure 11. *ROC curves of the top four classifiers for test images*

Table 2 shows the most successful classifiers and their performance metrics for the training and test set. Fine KNN classifier reached the highest accuracy with 98.4% for the training set. For the test set, the highest accuracy was obtained with the Cubic SVM classifier with 93%. Although the Fine KNN classifier reached the highest accuracy for the training set, the same classifier was able to achieve 87% accuracy on the test set. The Cubic SVM classifier, on the other hand, had an accuracy of 97.7% for the training set, while it reached the highest accuracy of 93% for the test set. For this reason, Cubic SVM has been evaluated as a more reliable classification model.

Table2. *Performance values of the classifiers used in the study*

		Fine KNN	Ensemble (Subspace KNN)	Cubic SVM	Quadratic SVM	Bilayered Neural Network	Narrow Neural Network	SVM Kernel
Accuracy (%)	Training	**98,4**	98,2	97,7	97	95,4	96,4	96,7
	Test	87	83,5	**93**	88,5	90,5	90	89
Error (%)	Training	1,5	1,8	2,2	2,9	4,6	3,5	3,2
	Test	13	16,5	7	11,5	9,5	10	11
Sensitivity (%)	Training	98.3	97,5	97,6	96,8	94,8	96,2	96,2
	Test	82	75	94	91	88	91	84
Specificity (%)	Training	98,5	98,8	98	97,3	96	96,7	97,1
	Test	92	92	92	86	93	89	94
Precision (%)	Training	98,5	98,8	97,8	97,2	95,8	96,6	97
	Test	91,1	90,3	92,1	86,6	92,6	89,2	93,3
False Positive Rate (%)	Training	1,4	1,1	2	2,6	4	3,3	2,8
	Test	8	8	8	14	7	11	6
F-Score (%)	Training	98,4	98,1	97,7	97	95,3	96,4	96,6
	Test	86,3	81,9	93	88,7	90,2	90,1	88,4
MCC	Training	0.9684	0.9641	0.9553	0.9413	0.9080	0.9299	0.9343
	Test	0.7437	0.6799	0.8602	0.7710	0.8110	0.8002	0.7839
Cohen's Kappa	Training	0.9684	0.9640	0.9553	0.9412	0.9079	0.9298	0.9342
	Test	0,74	0,67	0.86	0,77	0.81	0.80	0,78

4. DISCUSSION

The methodology and performance values of the study are given above in detail. In this section, a comparison is made over the classification performance of this study and another study using the same data set [7].

The compared study classified the same dataset using a CNN network. The data set was not divided into test and training, the network was fed only with training images, and as a result, 94% classification accuracy was achieved. The reliability of the model is questionable as it has not been tested with test images. Since no generalization error is obtained, it is not known whether the model has passed into an over-fitting state.

In our study, the data set was divided into two as training and test images, and the trained model was tested with test images. The classification performance for training images was 98.4%, even though a deep learning model was not used, and this value is higher than the result obtained in the compared study. The 93% classification accuracy obtained for the test images is very close to the 94% accuracy value obtained by the other study in education. Therefore, it can be said that the training accuracy obtained in the compared study was achieved with test images without using deep learning in our study. The reasonable difference between the accuracy rates obtained with the training and test images indicates that our model is not in an over-fitting state.

5. CONCLUSION

In this study, we classified the images as not Covid-Covid, using the dataset consisting of a total of 2481 images. To train the machine learning algorithms, we extracted the features of the images using the Local Binary Pattern method. Fine KNN, Subspace KNN, Cubic SVM, and Quadratic SVM were the most successful classifiers in the training phase for images trained with LBP features. In the classification process, we made with the part of the data set reserved for testing, the classifiers with the highest accuracy were respectively: Cubic SVM, Bilayered Neural Network, Narrow Neural Network, and SVM Kernel. The acceptable difference between training and test accuracy showed us that the models did not overlearn. The separation of the data set into training and test groups, the testing of the trained models with the test set, and the performance values obtained as a result show that the system is reliable.

CONFLICTS OF INTEREST

No conflict of interest was declared by the authors.

REFERENCES

[1] Sünnetci, K.M., Alkan, A., Tar, E., "Göğüs X-Ray görüntülerinin AlexNet tabanlı sınıflandırılması", Journal of Computer Science, IDAP-2021(Special): 375-384, (2021).

[2] https://covid19.saglik.gov.tr/TR-66300/covid-19-nedir-.html. Access Date: 23.07.2022

[3] https://www.worldometers.info/coronavirus. Access Date: 23.07.2022

[4] https://www.iaea.org/newscenter/news/how-is-the-covid-19-virus-detected-using-real-time-rt-pcr. Access Date: 23.07.2022

[5] https://www.cdc.gov/coronavirus/2019-ncov/science/science-and-research.html. Access Date: 23.07.2022

[6] https://en.wikipedia.org/wiki/CT_scan Access Date: 23.07.2022

[7] Kaya, B., Önal, M., "A CNN Based Method for Detecting Covid-19 from CT Images", Journal of Computer Science, IDAP-2021(Special): 1-10, (2021).

[8] Rasheed, J., Hameed, A. A., Djeddi, C., Jamil, A., Al-Turjman, F., "A machine learning-based framework for diagnosis of COVID-19 from chest X-ray images", Interdisciplinary Sciences: Computational Life Sciences, 13(1): 103–117, (2021).

[9] Hussain, L., Nguyen, T., Li, H., Abbasi, A.A., Lone, K.J., Zhao, Z., Zaib, M., Chen, A., and Duong, T.Q., "Machine-learning classification of texture features of portable chest X-ray accurately classifies COVID-19 lung infection", Biomedical Engineering Online, 19(88), (2020).

[10] Zhang, Y., Su, L., Liu, Z., Tan, W., Jiang, Y., Cheng, C., "A semi-supervised learning approach for COVID-19 detection from chest CT scans", Neurocomputing, 503: 314-324, (2022).

[11] Baghdadi, N. A., Malki, A., Abdelaliem, S. F., Magdy Balaha, H., Badawy, M., Elhosseini, M., "An automated diagnosis and classification of COVID-19 from chest CT images using a transfer learning-based Convolutional Neural Network", Computers in Biology and Medicine, 144: 105383, (2022).

[12] https://www.kaggle.com/plameneduardo/sarscov2-ctscan-dataset. Access Date: 23.07.2022

[13] Kızrak, A., "Local Binary Pattern Yöntemi ile Yüz İfadelerinin Tanınması", Şekil Tanıma Proje Raporu, (2014).

[14] Vaibhaw, Sarraf, J., Pattnaik, P. K., "Brain–computer interfaces and their applications", An Industrial IoT Approach for Pharmaceutical Industry Growth, 2: 31–54, (2020).

[15] https://www.geeksforgeeks.org/ensemble-classifier-data-mining. Access Date: 23.07.2022

[16] https://medium.com/@ekrem.hatipoglu/machine-learning-prediction-algorithms-decision-tree-random-forest-part-5-2970905c021e. Access Date: 23.07.2022

[17] Alkan, A., “Analysis of knee osteoarthritis by using fuzzy c-means clustering and SVM classification”, Scientific Research and Essays, 6(20): 4213-4219, (2011).

[18] Özbeyaz, A., “EEG-Based classification of branded and unbranded stimuli associating with smartphone products: comparison of several machine learning algorithms”, Neural Computing and Applications, 33: 4579–4593, (2021).

[19] Hasoon, J. N., Fadel, A. H., Hameed, R. S., Mostafa, S. A., Khalaf, B. A., Mohammed, M. A., and Nedoma, J., “COVID-19 anomaly detection and classification method based on supervised machine learning of chest X-ray images”, Results in Physics, 31: 105045, (2021).

[20] Alkan, A., “EEG işaretlerinin klasik ve modern yöntemlerle önişlenmesi ve sınıflandırılması”, Ph.D. Thesis, Sakarya University Instıtute of Scıence and Technology, Sakarya, (2005).

[21] Ekersular, M.N., “Bilgisayarlı tomografi (BT) görüntülerinde safra kesesi taşının morfometrik özelliklerinin belirlenmesi ve karaciğer içi safra kanallarının tespiti”, MSc. Thesis, Kahramanmaraş Sütçü İmam University, Institute of Science and Technology, Kahramanmaraş, 47-49, (2019).

[22] Bradley, A., “The use of the area under the ROC curve in the evaluation of machine learning algorithms”, Pattern Recognition, 30(7): 1145-1159, (1997).