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.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
COVID-19 detection using X-ray images and statistical measurements

Emre AVUÇLU

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

ARTICLE INFO

Keywords:
Biomedical images
COVID-19
Feature extraction
Machine learning algorithms

ABSTRACT

The COVID-19 pandemic spread all over the world, starting in China in late 2019, and significantly affected life in all aspects. As seen in SARS, MERS, COVID-19 outbreaks, coronaviruses pose a great threat to world health. The COVID-19 epidemic, which caused pandemics all over the world, continues to seriously threaten people’s lives. Due to the rapid spread of COVID-19, many countries’ healthcare sectors were caught off guard. This situation put a burden on doctors and healthcare professionals that they could not handle. All of the studies on COVID-19 in the literature have been done to help experts to recognize COVID-19 more accurately, to use more accurate diagnosis and appropriate treatment methods. The alleviation of this workload will be possible by developing computer aided early and accurate diagnosis systems with machine learning. Diagnosis and evaluation of pneumonia on computed tomography images provide significant benefits in investigating possible complications and in case follow-up. Pneumonia and lesions occurring in the lungs should be carefully examined as it helps in the diagnostic process during the pandemic period. For this reason, the first diagnosis and medications are very important to prevent the disease from progressing. In this study, a dataset consisting of Pneumonia and Normal images was used by proposing a new image preprocessing process. These preprocessed images were reduced to 15x15 unit size and their features were extracted according to their RGB values. Experimental studies were carried out by performing both normal values and feature reduction among these features. RGB values of the images were used in train and test processes for MLAs. In experimental studies, 5 different Machine Learning Algorithms (MLAs) (Multi Class Support Vector Machine (MC-SVM), k Nearest Neighbor (k-NN), Decision Tree (DT), Multinominal Logistic Regression (MLR), Naive Bayes (NB)). The following accuracy rates were obtained in train operations for MLAs, respectively; 1, 1, 1, 0.746377, 0.963768. Accuracy results in test operations were obtained as follows; 0.87755, 0.857143, 0.857143, 0.877551, 0.938776.

1. Introduction

Outbreaks of many infectious diseases such as smallpox, plague, cholera, Spanish flu, HIV/AIDS, severe acute respiratory failure syndrome (SARS), swine flu, Middle East Respiratory Syndrome (MERS) have been observed in the world history. Thanks to the widespread trade network and travels, new environments have been created for human and animal interactions and as a result, the spread of such epidemics has accelerated [1]. Despite the persistent diseases and epidemics throughout history, improvements in healthcare are an important tool in reducing the effects of these diseases. The new coronavirus disease 2019, the epidemic caused by the disease defined as COVID-19, was declared as a global pandemic by the World Health Organization (WHO) on March 12, 2020 [2]. As a result of genetic analysis of the epidemic on animals, it is thought that this virus, which was previously found in bats, mutated and settled as a host to animals such as the anteater species called pangolin, and gained the ability to infect humans from animal markets in China and Southeast Asia, which are shown as the source of the epidemic [3]. In addition, analysis of publicly available genome sequences of SARS-CoV-2 and related viruses found no evidence that the virus was produced in a laboratory or otherwise designed [4]. If someone tried to create a new coronavirus that causes disease, they would do so using the genetic sequence of a virus known to cause illness in humans. However, the SARS-CoV-2 genetic sequence was found to be significantly different from the backbone of known existing human coronaviruses and more similar to the associated viruses found in bats and anteaters [5]. An important question regarding the spread of the epidemic is why SARS-CoV2 is spreading more than other coronaviruses. In SARS, it has been shown that the viral load, which peaks 6–11 days after onset, is high since the onset of symptoms in COVID-19, decreases after 5–6 days, and in severe COVID-19 cases, viral load is detected longer than moderate cases. This makes isolation and quarantine of symptomatic individuals infected with SARS-CoV-2 much more difficult [6–11]. In addition, the SARS-CoV-2 spike protein binds to receptors on
human cells (acetylcholinesterase-2, ACE2) with a much higher power than other coronaviruses [12]. Another reason is that due to the higher mortality rate in SARS compared to COVID-19, it was observed that the epidemic remained at the epidemic level and did not cause a pandemic [13]. There are different interpretations from different sources about the number of cases when the disease was first seen. According to one of these, the number of cases is 42, according to another source, the number of cases is 41 [14,15]. Among these patients, all of them were determined to be Chinese citizens, except those who traveled from Wuhan to Thailand [16–18]. Detect, isolate, test and treat to break the COVID-19 spread chain, the World Health Organization said in a press release on March 13, 2020. However, he said that every case we find and treat will limit the spread of the disease. Using this statement is informative about the course of the disease next [19]. Many hypotheses have been put forward about the virus that is the cause of the new COVID-19 disease and the formation of this virus. It has been said that the virus has a zoonotic nature and is transmitted from host animals called bats and

![Fig. 1. DT algorithm.](image1)

![Fig. 2. Classification of a new data.](image2)

![Fig. 3. a) non-linearly b) determination of the separation plane c) conversion of input space to property space.](image3)

![Fig. 4. CM general structure.](image4)
Table 1

| Statistical Measurements | $TPR = \frac{TP}{TP + FN}$ | False Negative Ratio | $FNR = \frac{FN}{TP + FN}$ |
|---------------------------|-----------------------------|----------------------|-----------------------------|

| Sensitivity or True Positive Rate |
|----------------------------------|
| Specificity (SPO) or True Negative Rate |
| $TNR = \frac{TN}{TN + FP}$ |
| $TNR$ for Accuracy: $ACC = \frac{TP + TN}{TP + TN + FP + FN}$ |

| Negative Predictive Value |
|---------------------------|
| $NPV = \frac{TN}{TN + FN}$ |
| $NPV$ for F-Measurements: $FM = \frac{TP}{TP + FP}$ |

| False Positive Ratio |
|----------------------|
| $FPR = \frac{FP}{TN + FP}$ |
| Matthews Correlation Coefficient (MCC): $MCC = \sqrt{(\frac{TP + FN}{TP + FN + TN + FP}(TN + FP)(TN + FN))}$ |

Table 2

| Number of images in the database. |
|-----------------------------------|
| Normal Images | Pneumonia Images |
|----------------|------------------|
| Train | 74 | 74 |
| Test | 20 | 20 |

Fig. 5. The general formulation of 5-fold.

According to some sources, it is stated that there is no reliable evidence to support the claim that COVID-19 is caused by a coronavirus designed in this way, although the information that COVID-19 is artificially or intentionally produced in the laboratory environment is on the agenda [21,22]. It has been stated that the route of transmission for COVID-19 in Germany 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 the radiography, and patients may have tremors, nausea and muscle pain. The distribution rate of the disease by age groups is 50 % for 65 years and over. It has been reported that the patient needs hospital care and up to 60 days [23,24]. Official statements by the Chinese government to the WHO report that the first confirmed case was diagnosed on December 8, 2019. Government officials did not openly acknowledge human-to-human transmission until January 21, 2020 [25]. Worldwide, as of March 3, 2020, the estimated mortality rate by WHO is 3.4 %, when the analysis of these death cases is; It has been determined that it is seen mainly in 2/3 of men, 1/3 in women, more than 80 % of them are over 60 years old, and more than 75 % have chronic diseases such as cardiovascular diseases, diabetes and cancer. At the time of writing, it was reported that 2 004 819 cases were infected worldwide and 126 830 people died due to the disease [26–28]. COVID-19, which has become a global pandemic, has caused some negative consequences not only in the field of medicine, but also in social, professional, political and economic fields. 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, the delay and confusion in the measures to be taken, and the high transmission rate of the disease [29,30]. In the report that published the statement of the WHO, it was stated that 1760 people working in health institutions in Wuhan had positive tests and 6 of them died [31]. The illness of well-known people in the fields of politics, arts and sports in the world is another indicator of the spread and contagiousness of the disease in a pandemic. The transmission of these news through visual and printed media further increased the panic situation in the society [32,33]. Although COVID-19 mainly causes lung pneumonia, it can cause disease in many organs [34]. Many personal illness factors have been found that negatively affect the course of COVID-19. Among these, the most commonly identified are the male sex, diabetes mellitus, hypertension, and coronary artery disease [35]. It has been reported that it increases the risk of infection, especially in the respiratory system, due to the defense-suppressing drugs used in the treatment of inflammatory rheumatic diseases, which are observed in 2–3.5 % of the population [36,37]. However, the course of COVID-19 in individuals with rheumatological diseases has not been clearly determined due to the fact that COVID-19 is a new and recently defined disease and that inflammatory type rheumatic diseases are less common in the society compared to other diseases. Rare reports have also been made about the course of previous corona virus infections (MERS and SARS-1) in patients with suppressed defense systems [38]. It has been reported that fever and cough are the most common complaints. In a cross-sectional study conducted in Italy, the prevalence of COVID-19 was found to be 0.22% in a telephone survey of 458 patients followed in the rheumatology department of a tertiary hospital. This rate was observed to be at a level similar to the prevalence of the disease in the region where the patients are located. While patients had symptoms consistent with COVID-19, only one patient was diagnosed with a definite diagnosis and reported to have
symptoms severe enough to be hospitalized [39]. Again, in the tele-
phone or face-to-face evaluation of the cohort of 123 connective tissue
patients in Italy, 14 of the patients had respiratory symptoms compatible
with viral infection, while a definite diagnosis was made in one patient
with systemic sclerosis and the patient was lost [40].

The aim of this study is to both guide the experts who use MLAs in
diagnosing COVID-19 and to inform how the result may change in case
of unnecessary data while diagnosing COVID-19. A new method was
proposed for this. The feature extraction processes and obtained results
related to the proposed method were explained. In the second part of the
study, MLAs, k-fold crossvalidation, Confusion Matrix (CM), dataset and
their properties are presented. In this study, some experimental studies

Fig. 6. Image preprocessing.

Fig. 7. General view of a part of the dataset.
were conducted with different MLAs using the COVID-19 image dataset. Experimental studies were compared with statistical measurements. The results obtained from the experimental results were compared among themselves and with other studies in the literature.

2. Materials and methods

In this section, information was given about the MLAs used in this study and how they work. The statistical methods used to compare the results of the study were explained with their formulas. Information about the database containing COVID-19 images used in the study and k-fold cross validation was given. Finally, the proposed method and the image pre-processing process were explained in detail with the visuals.

2.1. Machine learning algorithms

Machine Learning is a concept that allows the machine to learn from examples and experience without being explicitly programmed. The MLAs used in this study were described sequentially.

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

Table 3
Parameters used for MLAs.

| k-NN | Distance Method: Euclidean | k:4 |
|-----|---------------------------|-----|
| DT  | Learning method: C4.5     |     |
| MLR | Estimate method: Lower-bound Newton–raphson |     |
| NB  | Distribution: Gaussian    |     |
| SVM | Kernel: Gaussian          | Tolerance:0.001 |

Table 4
CM results for normal data.

| CL0 | CLI | TP  | FN  | FP  | TN  | TPR | SPC | PPV | NPV  | FPR | FNR | ACC  | MCC  | FM  |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|------|-----|-----|------|------|-----|
| k-NN|     |     |     |     |     |     |     |     |      |     |     |      |      |     |
| CL0 | 23  | 1   | 23  | 1   | 5   | 20  | 0.96| 0.8 | 0.82 | 0.95| 0.3 | 0.05 | 0.88 | 0.77 | 0.88 |
| CL1 | 5   | 20  | 20  | 5   | 1   | 23  | 0.8 | 0.96| 0.95 | 0.82| 0.04| 0.18 | 0.88 | 0.77 | 0.87 |
| DT  |     |     |     |     |     |     |     |     |      |     |     |      |      |     |
| CL0 | 17  | 7   | 17  | 7   | 23  | 23  | 0.71| 0.92| 0.89 | 0.77| 0.08| 0.23 | 0.82 | 0.64 | 0.79 |
| CL1 | 2   | 23  | 23  | 2   | 7   | 17  | 0.92| 0.71| 0.77 | 0.89| 0.29| 0.11 | 0.82 | 0.64 | 0.84 |
| MLR |     |     |     |     |     |     |     |     |      |     |     |      |      |     |
| CL0 | 19  | 5   | 19  | 5   | 4   | 21  | 0.79| 0.84| 0.83 | 0.81| 0.16| 0.19 | 0.82 | 0.63 | 0.81 |
| CL1 | 4   | 21  | 21  | 4   | 5   | 19  | 0.84| 0.79| 0.81 | 0.83| 0.21| 0.17 | 0.82 | 0.63 | 0.82 |
| NB  |     |     |     |     |     |     |     |     |      |     |     |      |      |     |
| CL0 | 24  | 0   | 24  | 0   | 7   | 18  | 0.72| 0.77| 1    | 0   | 0.28| 0    | 0.86 | 0.75 | 0.87 |
| CL1 | 7   | 18  | 18  | 7   | 0   | 24  | 0.72| 1   | 1    | 0.77| 0   | 0.23 | 0.86 | 0.75 | 0.84 |
| MC- |     |     |     |     |     |     |     |     |      |     |     |      |      |     |
| CL0 | 24  | 0   | 24  | 0   | 4   | 21  | 1   | 0.84| 0.86 | 1   | 0.16| 0    | 0.92 | 0.85 | 0.92 |
| SVM |     |     |     |     |     |     |     |     |      |     |     |      |      |     |
| CL1 | 4   | 21  | 21  | 4   | 0   | 24  | 0.84| 1   | 1    | 0.86| 0   | 0.14 | 0.92 | 0.85 | 0.91 |

Table 5
CM results for reduced data.

| CL0 | CLI | TP  | FN  | FP  | TN  | TPR | SPC | PPV | NPV  | FPR | FNR | ACC  | MCC  | FM  |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|------|-----|-----|------|------|-----|
| k-NN|     |     |     |     |     |     |     |     |      |     |     |      |      |     |
| CL0 | 24  | 2   | 24  | 2   | 4   | 19  | 0.92| 0.83| 0.86 | 0.9 | 0.17| 0.1  | 0.88 | 0.76 | 0.89 |
| CL1 | 4   | 19  | 19  | 4   | 2   | 24  | 0.83| 0.92| 0.9  | 0.86| 0.08| 0.14 | 0.88 | 0.76 | 0.86 |
| DT  |     |     |     |     |     |     |     |     |      |     |     |      |      |     |
| CL0 | 19  | 7   | 19  | 7   | 0   | 23  | 0.73| 1   | 1    | 0.77| 0   | 0.23 | 0.86 | 0.75 | 0.84 |
| CL1 | 0   | 23  | 23  | 0   | 7   | 19  | 0.73| 0.77| 1    | 0.27| 0   | 0.86 | 0.75 | 0.87 |
| MLR |     |     |     |     |     |     |     |     |      |     |     |      |      |     |
| CL0 | 20  | 6   | 20  | 6   | 1   | 22  | 0.77| 0.96| 0.95 | 0.79| 0.04| 0.21 | 0.86 | 0.73 | 0.85 |
| CL1 | 1   | 22  | 22  | 1   | 6   | 20  | 0.96| 0.77| 0.79 | 0.95| 0.23| 0.05 | 0.86 | 0.73 | 0.86 |
| NB  |     |     |     |     |     |     |     |     |      |     |     |      |      |     |
| CL0 | 26  | 0   | 26  | 0   | 6   | 17  | 1   | 0.74| 0.81 | 1   | 0.26| 0    | 0.88 | 0.77 | 0.9  |
| CL1 | 6   | 17  | 17  | 6   | 0   | 26  | 0.74| 1   | 1    | 0.81| 0   | 0.19 | 0.88 | 0.77 | 0.85 |
| MC- |     |     |     |     |     |     |     |     |      |     |     |      |      |     |
| CL0 | 26  | 0   | 26  | 0   | 3   | 20  | 1   | 0.87| 0.9  | 1   | 0.13| 0    | 0.94 | 0.88 | 0.95 |
| SVM |     |     |     |     |     |     |     |     |      |     |     |      |      |     |
| CL1 | 3   | 20  | 20  | 3   | 0   | 26  | 0.87| 1   | 1    | 0.9 | 0   | 0.1  | 0.94 | 0.88 | 0.93 |
Fig. 9. ROC curves for k-NN.

Fig. 10. ROC curves for DT.

Fig. 11. ROC curves for MLR.
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 [41].

Looking at Fig. 1, it is seen that the tree has 3 types of nodes and the meanings of these nodes are as follows:

1. A root node that does not have a branch before it and from which zero or more branches can arise.
2. Internal nodes, which are just one branch coming towards it before it and two or more branches coming out of it.
3. They are leaf or pole (terminal) nodes, which are only a branch coming towards them before them and no branches emerge from them.

2.1.2. k-NN

k-NN is one of the algorithms used for classification and regression in Supervised Learning. It is considered the simplest machine learning algorithm. There are different methods (Minkowski, Euclid, etc.) for calculating the distance of a new sample. Euclidean, which is the most widely used measurement metric, was also preferred in this study (Equation (1)).

\[ d(i,j) = \sqrt{\sum_{p=1}^{n} (X_{ip} - X_{jp})^2} \] (1)

Where \( n \) represents the dimension. \( i \) is a new sample \( X_{ip} \) to be classified, and the nearest \( k \) neighbors \( X_{ip}(i = 1, 2, \ldots, k) \). Fig. 2 shows the process of classifying a new \( X_{ip} \) sample according to \( k = 3 \) in a two-dimensional \( (n = 2) \) space [41].

An example is shown in Fig. 2, since the new object (?) is the closest and the largest number of triangles (Fig. 2- a), this object is included in the triangle class (Fig. 2- b). New object with question mark, check 3 objects closest to it. Whichever object is most closest to the question mark, the question mark is included in that group.

---

**Fig. 12.** ROC curves for NB.

**Fig. 13.** ROC curves for SVM.
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 (Equation (2)).

\[
P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}
\]  

In Equation (2), \(P(X|Y)\) 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. \(P(X|Y)\): It is the probability of occurrence of X event when Y event occurs. \(P(Y|X)\): It is the probability of occurrence of event Y when event A occurs. \(P(X)\), \(P(Y)\): They are prior probabilities of X and Y events.

2.1.4. Support Vector Machine

SVM was first proposed by Vapnik and Chervonenkis in 1963. SVM is a MLA that works on the principle of structural risk minimization and based on convex optimization [43]. It is mainly designed to solve binary classification problems. SVM is divided into two groups, Linear SVM and Nonlinear SVM, according to the state of the data. SVM is an effective learning algorithm in complex data sets and identifying patterns that are difficult to analyze. The vast majority of classification problems in the real world consist of more than two classes. A multi-class SVM classifier is needed to solve such problems. Multiple classification can be achieved by combining binary classifiers [44]. Fig. 3 shows the general structures of SVM [41].

2.1.5. Multinomial Logistic regression

Logistic Regression (LR) is a nonlinear regression model designed for two dependent variables [45–47]. MLR is used to explain the cause-effect relationships between the dependent variable (Y) and the independent variables (X) when the dependent variable contains at least three or more categories and its values are obtained by classifying scale [48,49]. Since the purpose of this analysis is to estimate the value of dependent variables categorically, what we are trying to do here is actually the “membership” estimation for two or more categories. Accordingly, it can be stated that one of the purposes is classification and the other is to investigate the relationships between dependent and independent variables [50]. In other words, it aims to establish a model that will assign the observations to the classes they belong to in the most accurate way and determine the structures and risk factors related to the observations [51]. 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 (Equation (3)) and this ratio serves as a function that facilitates the transformation during linearization of the LR model.

\[
\text{Odds} = \frac{p}{1-p} = \frac{\pi(x)}{1-\pi(x)}
\]  

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);

\[
\pi(x) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p 
\]  

Where, \( \pi(\alpha) \) represents the probability of occurrence of an event under investigation, \( \alpha \) the dependent variable constant, \( \beta_1, \beta_2, \ldots, \beta_p \) are the regression coefficients of the independent variables, \( x_1, x_2, \ldots, x_p \) 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.

\[
\pi_j(x_i) = \frac{e^{\alpha_i + \beta_{1j} x_{i1} + \beta_{2j} x_{i2} + \ldots + \beta_{pj} x_{ip}}}{1 + \sum_{j=1}^{k} e^{\alpha_i + \beta_{1j} x_{i1} + \beta_{2j} x_{i2} + \ldots + \beta_{pj} x_{ip}}}
\]  

Where, \( k \) represents \( k \) categories, \( j_1, j_2, \ldots, j_k \), and the \( n \) levels of possible arguments \( i_1, i_2, \ldots, i_n \). It is observed that the use of logistic regression analysis has increased especially in the last 20 years in military matters, meteorology, internal migration movements and education. One of the most important reasons for this increase is the widespread use of statistical package programs. However, it is seen that one of the areas where it is used more widely is medicine [50].

It is frequently used to compare the estimates and actual values of the target attribute to evaluate the performance of the classification models used in machine learning. Regardless, classification estimates will have one of four assessments (Fig. 4):

- Saying True to True (True Positive) : TRUE
- Saying False to False (True Negative) : TRUE
- Saying True to False (False Positive) : FALSE
- Saying False to True (False Negative) : FALSE

![Fig. 16. 5 fold crossvalidation CM results for MLR.](image1)

![Fig. 17. 5 fold crossvalidation CM results for RF.](image2)
In this study, nine statistical measurement methods were applied. These measurements are shown in Table 1. 5-fold cross validation process was performed in this study. The general formulation of cross validation is as in Fig. 5. Cross-validation is known as a statistical resampling method used to evaluate the performance of MLAs on data that they have not seen, as objectively and accurately as possible. If we examine the working logic over the K-Fold Cross-validation method, which is one of the most basic cross-validation methods:

The data set is shuffled randomly. The data set is divided into k groups. The following steps are applied in order for each group:

- The selected group is used as the validation set.
- All other groups (k-1 groups) are treated as train sets.
- The model is built using the Train set and evaluated with the validation set, and as a result, the evaluation score of the model is stored in a list.

2.2. Image preprocessing and feature extraction

The COVID-19 database used in this study was taken from the Kaggle site [52,53]. The number of data in the dataset is as in the Table 2. There are pneumonia images in the database. Pneumonia is a disease that can result in death in individuals in a short time from the flow of fluid into the lungs. There are a total of 188 images in the database, consisting of normal and pneumonia images. Images of this dataset include X-ray images of patients affected by the first variant of COVID-19. Table 2. shows the number of images used in the database.

The images in the database were first standardized. For all images, the 3rd row L bone border indicated by the yellow arrow in the green rectangle in Fig. 6 was chosen. In addition, all images were cut so that images inside the lattice space remain.

As can be seen in Fig. 6, noises and unnecessary areas around the images were removed as a result. The images were all converted to the same size. Then it was cropped in 15X15 dimensions to extract features. Image cropping was done because the normal image size is large. Converted images to a reduced size of 15X15 units. If the image size was large, it would cause very large data to be generated after obtaining the RGB values for each pixel of the image. Thus, processing the data was easy. The general view of the database obtained in 15X15 dimensions is shown in Fig. 7.

RGB color mode space: The RGB color system generates 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. When the values of all components are 255, the result is pure white and when the values are 0, the result is pure black. Each axis consists of blue, green, and red colors. Three colors or channels are used in RGB images to create colors on the screen. In images with 8 bits per channel, three channels are converted into 24 bits (8 bits × 3 channels) color information for each pixel [54]. The color gray with hexadecimal color code #808080 is a shade of gray. #808080 in RGB color modeling; It consists of 50.2% red, 50.2% green and 50.2% blue. All of the images in this study were resized in 15 × 15 dimensions and the RGB value of each pixel obtained in 225 pieces was used as a feature. Feature reduction has been applied in case all values in the same column are the same. These features were not very distinctive in the classification process. 16 attributes were reduced as a result of feature reduction. Experimental studies were

---

**Table 6**

| Normal | Reduced |
|--------|---------|
| Train Accuracy Score | Train Cohen Kappa | Test Accuracy Score | Test Cohen Kappa | Train Accuracy Score | Train Cohen Kappa | Test Accuracy Score | Test Cohen Kappa |
| k-NN  | 1       | 0.755814 | 0.877551 | 1                | 1.000000 | 0.877551 | 0.752941 |
| DT    | 1       | 0.630962 | 0.816327 | 1                | 1.000000 | 0.857143 | 0.718159 |
| MLE   | 1       | 0.632193 | 0.816327 | 1                | 1.000000 | 0.857143 | 0.716763 |
| NB    | 0.763263 | 0.715824 | 0.887143 | 0.504436         | 0.746377 | 0.495509 | 0.877551 |
| MC    | 0.971014 | 0.873709 | 0.918367 | 0.941968         | 0.963768 | 0.927567 | 0.938776 |
| SVM   |         |         |         |                  |          |          |          |

---

Fig. 18. 5 fold crossvalidation CM results for SVM.
were discussed in detail in the following sections. Observed that more accurate results were obtained in the test processes. MLAs in feature reduction processes. In experimental studies, it was each image. Values in these columns were not presented as input to feature reduction, ROC curves and statistical measurement processes in 2.60 ghz, 16 gb RAM and 64 bit processor was used. Feature extraction, containing the RGB values in each pixel of the 15

Table 3 shows the parameters used more in the literature and selected because they gave better results in this study. The data obtained as a result of feature extraction from the COVID-19 image database were statistically measured for each MLA. The values obtained from the CM as a result of the normal classification process are shown in Table 4 (CL 1. Method Database Accuracy (%) -

| Method              | Database    | Accuracy |
|---------------------|-------------|----------|
| xDNN [55]           | X-ray images| 88.60    |
| SVM, k-NN, Naive Bayes [55] | X-ray images| 80.5, 83.9, 70.5 |
| COVID-Net [56]      | X-ray images| 90.6    |
| ResNet-50 [56]      | X-ray images| 92.3    |
| VGG-19 [56]         | X-ray images| 83.0    |
| Neuro-heuristic approach [57] | X-ray images| 79.06 |
| Different Pre-trained CNN [58] | X-ray images| 93.3 |
| CNN [59]            | X-ray images| 85.61   |
| Different pre-trained model [60] | X-ray images| 92.31 |
| ResNet50 [61]       | X-ray images| 93.06   |
| Fusion of Deep Learning and Machine Learning Technique (k-NN, SVM, DT, RF, SVM) [62] | X-ray images| 86 |
| Multi-class classification [63] | X-ray images| 79.52 |
| COVID-Net [64]      | X-ray images| 89.17 ± 0.015 |
| ResNet18 [64]       | X-ray images| 91.26 ± 0.014 |
| ResNet [64]         | X-ray images| 90.37 ± 0.015 |
| MobileNet-v2 [64]   | X-ray images| 86.83 ± 0.017 |
| Proposed Method- MC-SVM | X-ray images| 93.87 |

Comparison with other studies.

Table 7 shows the train and test scores obtained from the feature extraction method applied on the image and after feature reduction. The highest test score was obtained from SVM MLA. The Cohen Kappa test is a statistical method that measures the reliability of agreement between two or more observers. The reliability of conducted using both different properties. Fig. 8 shows the reduced version of the data containing the numerical values obtained after feature extraction in the data set. These data are numeric data containing the RGB values in each pixel of the 15 × 15 images. In this study, both all the data and the reduced data were analyzed and evaluated separately, and the results were discussed in detail in the following sections.

3. Experimental studies and discussions

In the experimental studies, a computer with an intel core i7 6700hq 2.60 ghz, 16 gb RAM and 64 bit processor was used. Feature extraction, feature reduction, ROC curves and statistical measurement processes in this study were performed in C# programming language. 5-fold cross-validation CM processes were performed with Matlab.

In this study, 75 % of the data was used for train and 25 % for test in experimental studies. The parameters used for MLA are shown in Table 3.

Table 3 shows the parameters used more in the literature and selected because they gave better results in this study. The data obtained as a result of feature extraction from the COVID-19 image database were statistically measured for each MLA. The values obtained from the CM as a result of the normal classification process are shown in Table 4 (CL = Class, CL0: Image with COVID-19, CL1: Image NOT COVID-19).

The values obtained from the CM as a result of the reduced data classification process are shown in Table 5.

Looking at the Table 4, and Table 5, it can be seen that as a result of feature reduction, the k-NN AUC value remained constant and all other MLAs AUC values increased. When we look at Table 4 and Table 5, it is seen that after the data reduction process, the number of correct predictions increased by 1 for k-NN MLA, 2 for DT MLA, 1 for MLR MLA, 2 for NB MLA, and 2 for MC-SVM MLA. The similarity or closeness of the results is related to the statistical calculations of the MLAs used. This may differ in different databases.

ROC curves were also drawn for both normal and reduced data for each MLA. ROC curves for k-NN are shown in Fig. 9.

Looking at Fig. 9, it can be seen that there was no improvement after data reduction for k-NN. This is a special case for k-NN MLA and means that it has no effect for the statistical methods used in k-NN. This situation may cause different results in other data. Here it only concerns the values of numeric data. ROC curves for DT are shown in Fig. 10.

When the ROC curves for DT are examined in Fig. 10, it is seen that the AUC value increased from 82 % to 86 % after the data reduction process. ROC curves for MLR are shown in Fig. 11.

When the ROC curves for MLR are examined in Fig. 11, it is seen that the AUC value increased from 82 % to 86 % after the data reduction process. ROC curves for SVM are shown in Fig. 12.

When the ROC curves for NB are examined in Fig. 12, it is seen that the AUC value increased from 86 % to 88 % after the data reduction process. ROC curves for SV are shown in Fig. 13.

When the ROC curves for SVM are examined in Fig. 13, it is seen that the AUC value increased from 92 % to 94 % after the data reduction process.

It was observed that the AUC value did not change for some MLAs as a result of feature reduction. This may indicate that the proposed method may not be effective in some statistical methods. It should be noted here that the first of these, the proposed method did not reduce the AUC value anyway, and the other one, perhaps the proposed method for different databases may be more effective for each MLA.

The highest AUC value was obtained from SVM MLA after feature reduction. 5 Fold cross validation process was applied for MLAs. Fig. 14. 5-fold cross-validation CM results are shown for both normal and feature reduced data for k-NN MLA.

Looking at Fig. 14, it is seen that for the CL0 (with COVID-19) class, the correct prediction value of 91 increased to 92 after data reduction. Likewise, it is seen that the number of 1 correct predictions increases after data reduction in the CL1 class. Fig. 15. 5-fold cross-validation CM results are shown for both normal and feature reduced data for DT MLA.

Looking at Fig. 15, it is seen that for the CL0 (with COVID-19) class, the correct prediction value of 78 increased to 80 after data reduction. Likewise, it is seen that the number of 3 correct predictions increases after data reduction in the CL1 class. Fig. 16. 5-fold cross-validation CM results are shown for both normal and feature reduced data for MLR MLA.

Looking at Fig. 16, it is seen that for the CL0 (with COVID-19) class, the correct prediction value of 78 increased to 80 after data reduction. Likewise, it is seen that the number of 4 correct predictions increases after data reduction in the CL1 class. Fig. 17. 5-fold cross-validation CM results are shown for both normal and feature reduced data for SVM MLA.

Looking at Fig. 17, it is seen that for the CL0 (with COVID-19) class, the correct prediction value of 92 increased to 93 after data reduction. Likewise, it is seen that the number of 1 correct predictions increases after data reduction in the CL1 class. Fig. 18. 5-fold cross-validation CM results are shown for both normal and feature reduced data for SVM MLA.
the comparative agreement was measured with Cohen’s kappa coefficient. The agreement between the raters was checked. It was observed that there was generally a significant and almost perfect agreement between the comparison values. It is also seen that Cohen Kappa values increase in test processes. Looking at the table according to Cohen Kappa values, it can be seen that there is good agreement between the models used in this study for each of the results. In addition, it can be said that the answers given by the compared models are compatible with each other.

It was observed that SVM MLA algorithm gave the best result. In Table 7, other studies on COVID-19 using the same and different databases in the literature are given.

Accuracy values obtained in the studies in Table 7 may vary according to the model used in the study. In a study using MLA algorithms, the following results were obtained for SVM, k-NN, NB MLA, respectively: 80.5 %, 83.9 %, 70.5 % [55]. Different data augmentation methods and different deep learning models were compared [58]. Different results were obtained by applying different models with deep learning. In the literature, there are studies using deep learning models, MLA and different optimization algorithms [60–64]. In another study using MLA algorithms, the highest accuracy rate was 86%. Compared to other studies in the literature, the results obtained from this study were found to be better. Among the different MLAs used in this study, the highest 93.87 % was obtained from the MC-SVM MLA. When the studies using different algorithms are examined in Table 7, it is seen that the accuracy rate of the proposed method is higher than the studies in the literature.

4. Conclusions and future works

COVID-19 is an urgent public health problem of international concern. COVID-19 has caused infections that have a huge impact on people’s lives and have resulted in global epidemics. Although radiological examinations such as Computed Tomography in COVID-19 patients are not fully diagnostic tests, they can be seriously helpful in diagnosis and differential diagnosis. Since December 2019, the fight against COVID-19 infection continues without slowing down. The infection affects more and more people every day. Articles about COVID-19, which have been brought to the literature and based mostly on scientific data, help experts in the correct diagnosis and treatment of the disease. In this study, COVID-19 was diagnosed using the COVID-19 image database. Experimental studies were performed using k-NN, DT, MLR, NB, SVM MLA using statistical methods. In the train results, the following ratios were obtained for k-NN, DT, MLR, NB, SVM, respectively: 100 %, 100 %, 100 %, 74 %, 96 %. Likewise, the following rates of test success were obtained for k-NN, DT, MLR, NB, SVM, respectively: 87 %, 85 %, 85 %, 87 %, 93 %. The data obtained as a result of the feature reduction process used together with the proposed feature extraction method show that this method can be applied in any image dataset. The feature extraction method applied in this study will be applied to images of different sizes (for example, 20X20, 30X30, 50X50 ... ) and their accuracy will be examined. In addition, the performances of different feature extraction methods and feature extraction methods applied in this study will be compared. Finally, an algorithm will be developed to automatically select the highest accuracy rate obtained from MLA algorithms as a result of different feature extraction methods.

CRediT authorship contribution statement

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

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.

Data availability

No data was used for the research described in the article.

References

[1] C. Gein, A. Kara, Global surveillance, travel, and trade during a pandemic, Turk J. Med. Sci. 50 (2020) 527–533.
[2] World Health Organization. WHO announces COVID-19 outbreak a pandemic. 2020. Available from: http://www.euro.who.int/en/health-topics/health -emergencies/coronavirus-covid-19/news/news/2020/3/wh o-announcescovid-19-outbreak-a-pandemic Accessed: 6 May 2020.
[3] T. Zhang, Q. Wu, Z. Zhang, Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak, Curr. Biol. 30 (7) (2020) 1346–1351.
[4] L. Li, G. Li, J. Chen, X. Liang, Y. Li, Comparative genomic analysis revealed specific mutational pattern between human coronavirus SARS-CoV-2 and Bat-SARS-CoV RaTG13, bioReiv. (2020).
[5] Genomic epidemiology of novel coronavirus - Global subampling. 2019-2020 Available from: https://nextstrain.org/ ncov/global Accessed: 6 May 2020.
[6] E. Petersen, D. Gökengin, SARS-CoV-2 epidemiology and control, different scenarios for Turkey, Turk. J. Med. Sci. 50 (2020) 509–514.
[7] F.S. Wang, C. Zhang, What to do next to control the 2019-nCoV epidemic? The Lancet. 395 (10222) (2020) 391–393.
[8] K.K.W. To, O.T.Y. Tsang, W.S. Leung, et al., Temporal profiles of viral load in posterior oropharyngeal saliva samples and serum antibodies during infection by SARS-CoV-2: an observational cohort study, Lancet Infect. Dis. 20 (5) (2020) 566–574.
[9] P.K.C. Cheng, D.A. Wong, L.K.L. Tong, et al., Viral shedding patterns of coronavirus in patients with probable severe acute respiratory syndrome, The Lancet. 363 (9422) (2004) 1699–1700.
[10] Y. Pan, D. Zhang, P. Yang, et al., Viral load of SARS-CoV-2 in clinical samples, Lancet Infect Dis. 20 (4) (2020) 411–412.
[11] Y. Liu, L.M. Yan, L. Wan, et al., Viral dynamics in mild and severe cases of COVID-19, Lancet Infect Dis. (2020).
[12] A.C. Wallis, Y.J. Park, M.A. Tortorici, A. Wall, A.T. McGuire, D. Veesler, Structure, function, and antigenicity of the SARS-CoV-2 Spike glycoprotein, Cell 181 (2) (2020) 281–292.
[13] World Health Organization Department of Communicable Disease Surveillance and Response. Consensus document on the epidemiology of severe acute respiratory syndrome (SARS) [online]. 2003 Available from: https://www.who.int/ cts/sars/ WHOconsensus.pdf Accessed: 6 May 2020. 14. Akin, L. Gozel MG. Understanding dynamics of pandemics. Turk J Med Sci. 2020;50:515-9.
[14] H. Lu, C.W. Stratton, Y.W. Yang, Outbreak of pneumonia of unknown etiology in Wuhan, China: the mystery and the miracle, J Med Virol. 92 (2020) 401–402.
[15] Zhao S. Musa SS. Lin Q. Ran J. Yang G. Wang W. Et al. Estimating the Unreported -rate/#correct (Access date: 06.04.2020).
[16] H. Lu, C.W. Stratton, Y.W. Yang, Outbreak of pneumonia of unknown etiology in Wuhan, China: the mystery and the miracle, J Med Virol. 92 (2020) 401–402.
[17] Zhao S. Musa SS. Lin Q. Ran J. Yang G. Wang W. Et al. Estimating the Unreported -rate/#correct (Access date: 06.04.2020).
[18] Tesini BL. Coronavirus and Acute Respiratory Syndromes (COVID-19, MERS, and SARS), https://www.msdmanuals. com/professional/infectious-diseases/respiratory viruses/coronaviruses-and-acute-respiratory-syndromes-COVID-19, mers, and-sars (Access date: 30.03.2020).
[19] Koronavirüs neden hızlı yayıldı? Corona Virüsü. https://www.e-psiykisiyiari.com/ko ronavirüs-neden-hız-yayidi-corona-virüsü (Access date: 04.04.2020).
[20] C.X. Deng, The global battle against SARS-CoV-2 and COVID-19, Int. J. Biol. Sci. 16 (10) (2020) 1676–1677.
[21] Menachery DV, Yount Jr BL, Debinkh K. Agnihotham S. Gralinski LE. Plante JA. et. al. A SARS-like cluster of circulating but coronaviruses shows potential for human emergence. Nature Medicine. 2015; Number 12, December;1508-1514.
[22] S.L. Liu, L.J. Stif, S.R. Weiss, L. Su, No credible evidence supporting claims of the laboratory engineering of SARS-CoV-2, Emerging Microbes Infect. 9 (2020) 505–507.
[23] Germany’s 2012 Covid scenario became real in 2020. https://www.foreigner.fr/artic les/news/germany-s-2012-covidscenario-became-real/2020032501440040 4958.html (Access date: 04.04.2020).
[24] Deutscher Bundestag. 17. Wahlperiode. Drucksache 17/12051. 03. 01. 2013:2-88.
[25] First COVID-19 case happened in November, China government records show – report. https://www.theguardian.com/ world/2020/mar/13/first-COVID-19-ca se-happened-innovember-china-government-records-show-report (Access date: 04.04.2020).
[26] 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: 06.04.2020).
Hosmer, D. W., Lemeshow, S. and Sturdivant, R. X., EULAR COVID-19 Database. 18.05.2020 ed. https://www.eular.org/eular_covid19_database.cfm.

F. Atzeni, K. Bendtzen, F. Bobbio-Pallavicini, et al., Infections and treatment of patients with connective tissue diseases: a descriptive observational analysis, J. Rheumatol (2020), https://doi.org/10.3899/jrheum.200507.

E.G. Favalli, E. Agape, R. Caporali, Incidence and clinical course of COVID-19 in patients with connective tissue disease: a descriptive observational analysis, J. Rheumatol (2020), https://doi.org/10.1061/10.1061/jrheum.2010.102575.

F. Zhou, T. Yu, R. Du, et al., Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study, Lancet 395 (10229) (2020) 1054–1062.

O. Sangha, Epidemiology of rheumatic diseases, Rheumatology 39 (S2) (2000) 3–12.

F. Atzori, K. Bendtzen, F. Bobbio-Pallavicini, et al., Infections and treatment of patients with rheumatic diseases, Clin. experimental rheumatol. 26 (1 Suppl 48) (2008) 67–73.

EULAR COVID-19 Database. 18.05.2020 ed. https://www.eular.org/eular_covid19_database.cfm.

G. Emmi, A. Bettiol, I. Mattioli, et al., SARS-CoV-2 infection among patients with systemic autoimmune diseases, Autoimmun. Rev. (2020), 102575, https://doi.org/10.1016/j.autrev.2020.102575.

E.G. Favalli, E. Agape, R. Caporali, Incidence and clinical course of COVID-19 in patients with connective tissue disease: a descriptive observational analysis, J. Rheumatol (2020), https://doi.org/10.1061/10.1061/jrheum.2010.102575.

Emre Arıcan, A new data augmentation method to use in machine learning algorithms using statistical measurements, Measurement, Volume 180, 2021,109577, ISSN 0263-2241.

Orhan, U. and Adem, K., “The Effects of Probability Factors in Naïve Bayes Method”, Elektrik-Elektronik ve Bilgisayar Mühendisi Sempozyumu, Bursa, 722–724 (2012).

K.P. Somani, R. Loganathan, V. Ajay, Machine learning with SVM and other kernel methods, PHI Learning Pvt. Ltd., Delhi/India, 2009, pp. 1–10.

Z.G. Jiang, H.G. Fu, L.J. Li, Support Vector Machine for Mechanical Faults Classification, J. Zhejiang University Sci. 6 (5) (2005) 433–439.

Harrell, F. E., ‘Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression and Survival Analysis’, Springer-Verlag, New York, 215-267 (2001).

Homser, D. W., Lemeshow, S. and Sturdivant, R. X., ‘Applied Logistic Regression 3rd Ed.’, Wiley&Sons Publications, Canada, 8-35 (2013).