Neural Computing and Applications (2022) 34:14729–14743
https://doi.org/10.1007/s00521-022-07325-y

ORIGINAL ARTICLE

Statistical analysis of blood characteristics of COVID-19 patients and their survival or death prediction using machine learning algorithms

Rahil Mazloumi[1] · Seyed Reza Abazari[1] · Farnaz Nafarieh[1] · Amir Aghsami[1,2] · Fariborz Jolai[1]

Received: 15 March 2021 / Accepted: 18 April 2022 / Published online: 11 May 2022

Abstract
This study's main purpose is to provide helpful information using blood samples from COVID-19 patients as a non-medical approach for helping healthcare systems during the pandemic. Also, this paper aims to evaluate machine learning algorithms for predicting the survival or death of COVID-19 patients. We use a blood sample dataset of 306 infected patients in Wuhan, China, compiled by Tangji Hospital. The dataset consists of blood's clinical indicators and information about whether patients are recovering or not. The used methods include K-nearest neighbor (KNN), decision tree (DT), logistic regression (LR), support vector machine (SVM), random forest (RF), stochastic gradient descent (SGD), bagging classifier (BC), and adaptive boosting (AdaBoost). We compare the performance of machine learning algorithms using statistical hypothesis testing. The results show that the most critical feature is age, and there is a high correlation between LD and CRP, and leukocytes and CRP. Furthermore, RF, SVM, DT, AdaBoost, DT, and KNN outperform other machine learning algorithms in predicting the survival or death of COVID-19 patients.

Keywords COVID-19 · Blood sample · Healthcare system · Machine learning · Statistical analysis

1 Introduction

Animal-origin COVID-19 first appeared in Wuhan, China, in December 2019, and on December 31, 2019, the pandemic virus was reported to the World Health Organization (WHO) as a new threat to communities. The disease outbreak rate has increased dramatically. According to WHO, in 72 countries, 1,05,586 positive cases have been reported by March 8, 2020 (WHO 2020) [1]. As of August 31, 2020, more than 180,000 people died in the USA [2]. Due to the symptoms' inconsistency in patients with COVID-19 and the diagnostic test mistakes, researchers face many challenges in this area [3].

The high cost, scarcity of diagnostic kits, and specialized laboratories in countries have led to more specialized tests being performed only on critically ill patients. In this situation, finding a way to reduce the number of tests that can provide a definitive answer to the medical staff can be very effective [4]. Each of the lactate dehydrogenase (LDH) level tests and the complete blood count (CBC) test, and others alone are no specific tests to measure a patient's deterioration, but together they can provide good performance. These tests can also be used in conjunction with reverse transcription-polymerase chain reaction (rRT-PCR), which is the most common test to detect COVID-19 for greater accuracy [5].

Machine learning methods have solved problems in many scientific fields over the past decade. These

✉ Amir Aghsami
a.aghsami@ut.ac.ir

Rahil Mazloumi
rahil.mazloumi@ut.ac.ir

Seyed Reza Abazari
reza.abazari@ut.ac.ir

Farnaz Nafarieh
farnaz.nafarieh@ut.ac.ir

Fariborz Jolai
fjolai@ut.ac.ir

[1] School of Industrial and Systems Engineering, College of Engineering, University of Tehran, P.O. Box, Tehran 11155-4563, Iran

[2] School of Industrial Engineering, K. N. Toosi University of Technology (KNTU), Tehran, Iran

algorithms use historical data and predict events. Predicting confirmed cases, diagnosing a disease by CT scan of the lungs and coughing sound, predicting intubation for the patient, and predicting and influencing climate parameters on the spread of the disease are some of the machine learning applications during the epidemic [6].

This study uses a collection of blood samples from 306 patients with COVID-19 in Wuhan, China, approved by the ethics committee of Tangji Hospital [7]. LD, a highly sensitive C-reactive protein (hs-CRP), lymphocyte, leukocytes, percentage lymphocytes, and age are six biomarkers whose variations in blood levels can indicate COVID-19 infection and disease progression. We use these biomarkers to predict and analyze a patient's likelihood of survival and death. Due to the data's heterogeneity, we balanced the data using the available methods and analyzed the relationship and correlation of biomarkers. Then, predictions were made by support vector machine (SVM), decision tree (DT), random forest (RF), K-nearest neighbor (KNN), logistic regression (LR), stochastic gradient descent (SGD), bagging classifier (BC), and adaptive boosting (AdaBoost). Finally, the performance of machine learning algorithms was compared, and the best ones with the most accuracy were determined by conducting a statistical hypothesis test.

2 Literature review

Since the advent of COVID-19 disease, many studies have been conducted to analyze and detect patterns in datasets related to COVID-19 patients. Some of these studies focused on predicting the deterioration of patients with COVID-19. Assaf et al. [8] used three different machine learning algorithms to identify patients' risk during hospitalization and predict the patients' condition before they undergo critical condition. This will lead to the effective management of the hospitals' intensive care sector. Arvind et al. [9] examined the clinical information of 4087 patients admitted to 5 hospitals. They used a machine-learning algorithm to provide a tool for better evaluation of patients who needed intubation and mechanical ventilation. Their proposed algorithm is significantly better than the ROX index for the risk of blockage and intubation. Several artificial intelligence (AI) methods were used to predict mortality in critically ill patients with COVID-19. For this purpose, Chaurasia and Pal [10] used data from the WHO, including information about the date, origin, country, and the latest COVID-19 updates over five months. Among the simple mean methods, moving means, naive, ARIMA method was introduced as the most appropriate method. Li et al. [11] applied machine learning algorithms to derive prognostic models for predicting patients' mortality with COVID-19. Predicting patients' recovery period with machine learning algorithms was done by Muhammad et al. [12]. They predicted the recovery of patients with COVID-19 using the epidemiological dataset of COVID-19 patients in South Korea and data mining models. They predicted the minimum and maximum number of days for the patient to recover, as well as patients who were unlikely to recover. They used DT algorithms, naive Bayes, SVM, LR, RF, and nearest neighbor directly on the dataset. They introduced the DT algorithm as the most effective way to predict patients' recovery.

One of the most important studies that have been done is the diagnosis of positive cases of COVID-19. Brinati et al. [13] diagnosed COVID-19 by presenting two classification methods and hematochemical routine blood tests. Their proposed model can replace the polymerase chain reaction (PCR) test. Additionally, they demonstrated that between LR and RF, RF has better performance for blood test samples. The ability to predict the number of new cases for 5 consecutive days was provided by Khakharia et al. [14]. They have developed a prediction system for COVID-19 outbreaks in the top 10 highly and densely populated countries. The proposed prediction models forecast the number of new cases likely to arise for five successive days using 9 different machine learning algorithms. For example, the auto-regressive moving average (ARMA) performed best for Germany and India, and the XGB model performed better for China. One of the notable capabilities of machine learning algorithms in the field of pathology is a diagnosis by CT scan and visual clinical data. Hussain et al. [15] categorized lung images into four categories: COVID-19, bacterial pneumonia, non-COVID-viral pneumonia, and normal, using patient chest X-ray (CXR) imaging data and five different machine learning algorithms. Their proposed system distinguished the morphological features of CRV-19 pulmonary infection CRX from the rest of the data. A deep learning method called convolutional neural network (CNN) has been used to diagnose COVID-19 by lung scan of patients. For this purpose, Yasar and Ceylan [16] used lung scans of 1396 people and identified the patients. Sharma [17] classified CT scans of patients' lungs into two categories: patients with pneumonia and patients with COVID-19 using machine learning techniques. This technique has been used in hospitals in China, Italy, Moscow, and India. Khanday et al. [18] used 212 clinical textual data provided by Johns Hopkins University and employed supervised machine learning techniques to classify the data into four disease categories. The results showed that LR and naive Bayes classifier algorithms provided more accuracy. Identifying COVID-19 patients and predicting Acute Respiratory Distress Syndrome (ARDS) is a study conducted by Jiang [19]. They used historical data from two hospitals in Wenzhou and Zhejiang, China, and AI techniques. Vijayakumar and

Sneha [20] processed cough audio data using deep learning approaches. They recorded respiratory and non-respiratory patients' data and used SVM with RBF kernel and LSTM technique, which is a neural network, to classify them accurately. Finally, they divided them into four categories pertussis, pneumonia, COVID, and normal hack. Planning is needed for hospitals' capacity and the allocation of medical resources and supplies during the COVID-19 outbreak. Qian et al. [21] introduced the capacity planning and analysis system (CPAS) based on machine learning to plan hospital capacity on a national scale and successfully deployed this new system in various hospitals in the UK. CPAS is one of the first machine learning systems deployed nationwide to address COVID-19 in hospitals, helping manage and allocate medical resources in hospitals.

Estimating the prevalence of the disease nationwide will provide valuable assistance to the medical staff and anti-COVID-19 policies in countries. Sujath et al. [22] developed a machine learning-based prediction model to predict the prevalence of COVID-19 in India. They used linear regression, multilayer perceptron (MLP), and self-regression vector method to predict the disease's epidemiological sample and its incidence. Comparing the predicted cases with the Johns Hopkins University data, they concluded that the MLP method offers better results than other methods.

Shrivasav and Jha [23] used a gradient-based machine learning method to investigate the relationship between the COVID-19 transfer rate in meteorological parameters in India. They were able to implement an efficient method of predictive modeling. Albahri et al. [24] studied COVID-19 prediction algorithms based on AI, data mining, and machine algorithms. They found the lack of real-world studies and the lack of access to large-scale updated data as a significant gap in the field. They called for the full cooperation of AI, data mining professionals, and the medical community. Shuja et al. [25] provided a comprehensive review of the COVID-19 open-source dataset and organized it by data type. Medical images, textual data, and spoken data are the main types of this category. They identified the main challenge in this area as the lack of information and research methods. In a study in Iran, Behnam and Jahanmahin [26] discussed the prediction process and mortality rate using machine learning algorithms compared to the global level. The Gaussian function was used to find the best model for estimating the peak and end times of the disease in the short and long term. A review of the most important machine learning forecasting models for COVID-19 and a brief analysis of related literature is presented by Rahimi et al. [27].

There have not been many studies on clinical blood indices in the existing literature. In contrast, further studies on clinical blood indices help a lot to analyze the recovery process and deterioration of patients with COVID-19. In this study, considering the blood indicators and age, important analyses have been performed about these indicators' relationship. Also, the performance of the algorithms used is measured, and the best algorithm(s) is introduced.

3 Material and methods

This study pursues three main objectives. The first is analyzing every clinical indicator of patients' blood and their impact on the survival or death of patients. The second is to predict the recovery or mortality of patients using machine learning algorithms. Finally, the accuracy of the results obtained from the algorithms will be analyzed. In the first step, the data are cleared and balanced, and then the relationship between each of the datasheet features is examined. The most important factors affecting the mortality rate are examined. In general, a statistical analysis is performed on all numerical and non-numerical variables to find a relationship with the patient's deterioration, recovery, and mortality. The framework of this study is shown in Fig. 1.

3.1 Data exploration

In this section, a pair plot is drawn to identify the dataset's patterns and extract information from the dataset. The first row and column of the pair plot in Fig. 2 show that age distribution for both alive and deceased patients is close to normal distribution. In general, there are many ways to test the normality of data, so in this study, we used a graphical test (Probability Plot) in Minitab software to assess whether the sample data follows a normal distribution or not [28]. Using the probability plot, we have a more accurate analysis of how the data are scattered. In this visual test, placing the data during a straight line indicates a 100% fit with the normal distribution and a fit with the main regression line.

In Fig. 3, the discrepancy with the regression line indicates that the data distribution is not normal. For a more detailed examination, we assess the hypothesis that the data distribution is not significantly different than normal (H_0) versus the data are significantly different than normal (H_1). Obtained results illustrated in Fig. 3 indicate that the data do not follow normal distribution because the*p*-values are less than the confidence level ($\alpha = 0.05$), so the assumption is that the data are normal (H_0) is rejected.

Moreover, Fig. 2 indicates that mortality is more common in elderly patients with COVID-19 than in younger people. In the age–LD diagram, we see a direct relationship

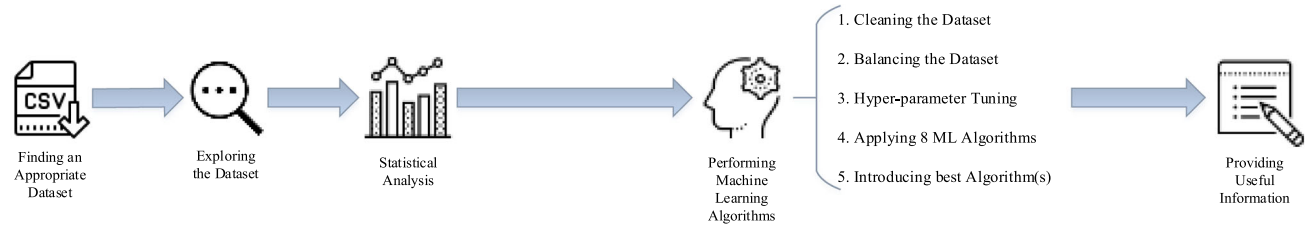


Fig. 1 The framework of this study

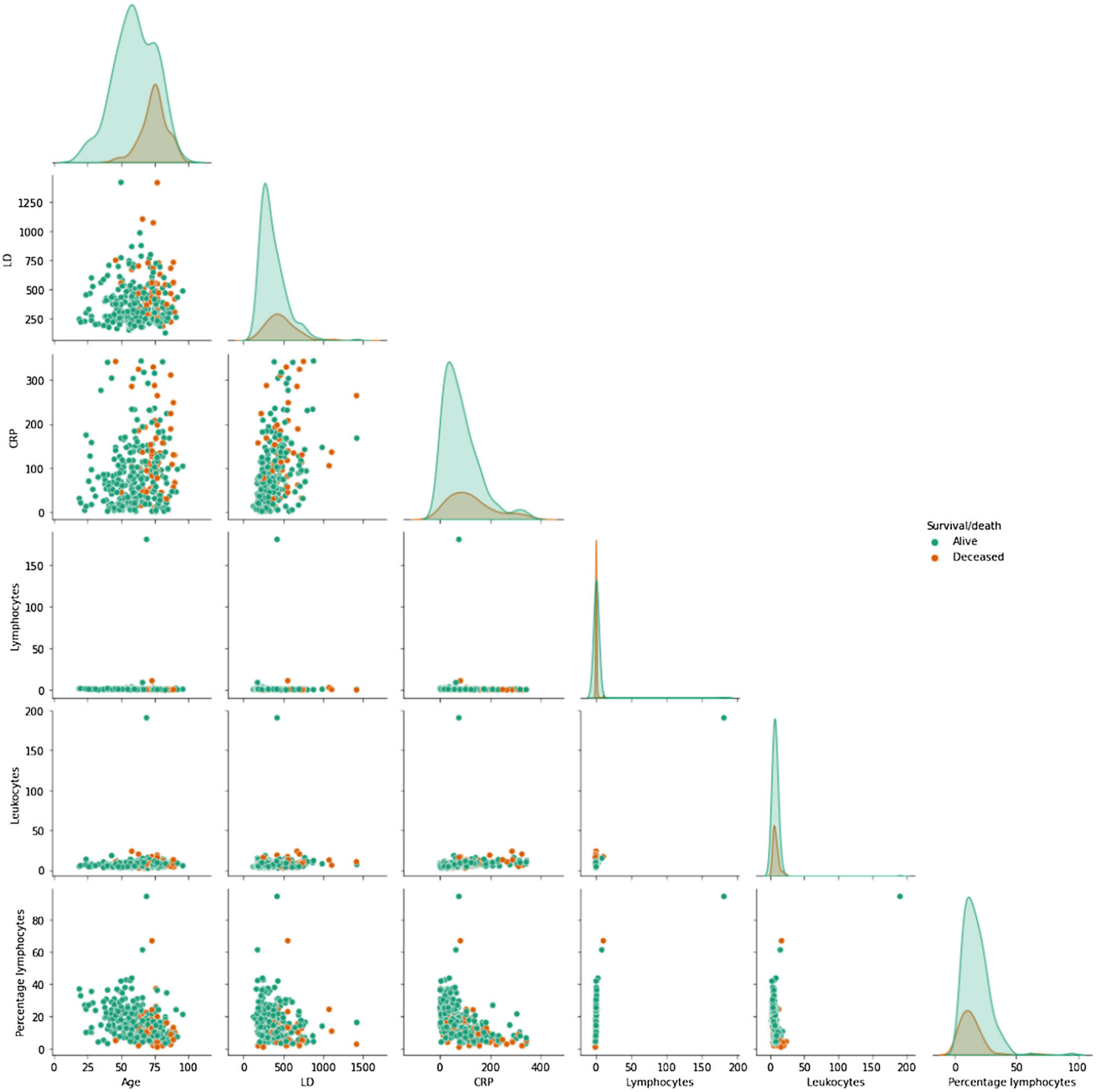


Fig. 2 Pair plot charts to show binary relationships of dataset's features

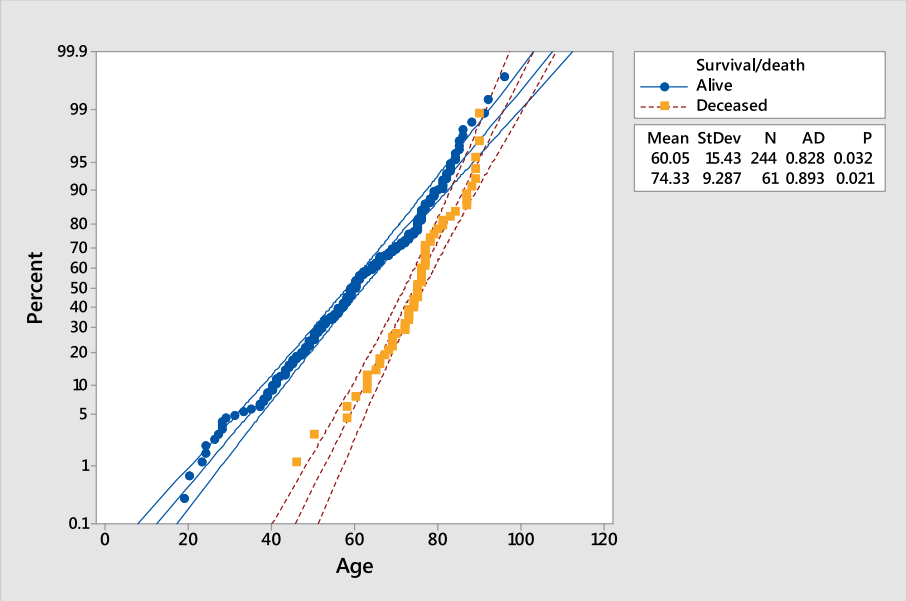


Fig. 3 Probability plot of age

between the increase in blood LD levels and age, which indicates the deterioration of older patients and damage to lung tissue. The age–CRP chart shows the effect of aging on the hs-CRP level of blood, which indicates an increase in infection and more tissue damage in older patients with COVID-19. Age–lymphocytes percentage chart shows the expected inverse relationship, and the decrease in percentage lymphocytes, which is a possible sign of disease and viral infection, is more common in older patients [29]. In the percentage lymphocytes–leukocytosis chart, we also see the effect of decreasing this clinical index on increasing mortality. Age–percentage lymphocytes show the inverse relationship between percentage lymphocytes and age.

According to Fig. 4, only the group of deceased women with a*p*-value of more than (0.05), (0.385) follows a normal distribution, and other groups in significance level of 0.05 do not follow the normal distribution.

Now, the correlation between the dataset's features will be examined. For this purpose, Spearman correlation, which is nonparametric, will be used for two reasons; the dataset does not follow a normal distribution, and the dataset consists of both ordinal and continuous variables. According to Fig. 3, there is a patient with a high level of

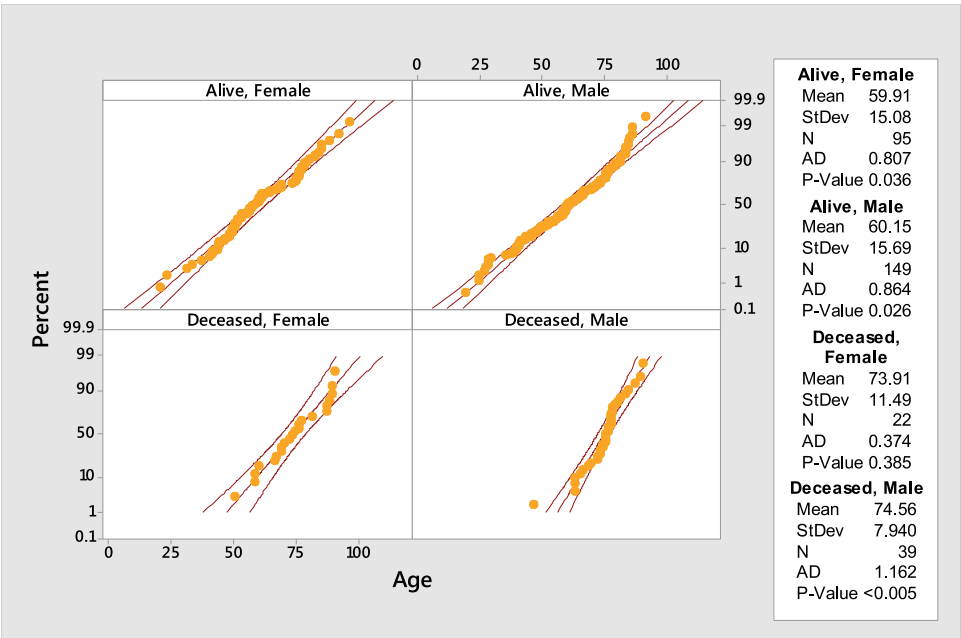


Fig. 4 Probability plot of age regarding alive/deceased in respect of gender

leukocytes and lymphocytes, which might affect the correlation between variables. We have calculated correlation for the variables by excluding and including the outlier. As shown in Figs.5 and6, it does not affect the results significantly.

The results show that the number of leukocytes and CRP in patients' blood tests with a correlation coefficient of 0.58 has a direct relationship. Each one can be used alone if the other one is not available because a high level of each of these two indicators indicates a high level of inflammation and infection with an increase in the number of white blood cells in the patient's body. According to scientific findings, there is a direct relationship between the rate of lung infection and CRP levels [30]. A high positive correlation in the results confirms this statement. It should be noted that the CRP test cannot definitively confirm the patient with COVID-19. This is because the CRP test measures the level of inflammation and infection by any type of bacteria or virus [31]. Also, the correlation between CRP and LD confirms the presence of inflammation. In particular, relatively high levels of LD alone can play an important role in diagnosing most cases that require immediate medical attention [32].

The negative correlation between percentage lymphocytes and LD, CRP, and age is as expected, but the inverse relationship between CRP and percentage lymphocytes is more important. We found that these two variables act oppositely in the rate of improvements and deaths with a negative correlation. By decreasing CRP and increasing the level of percentage lymphocytes, the number of survived people increases. Also, the number of dead people increases by increasing CRP and consequently decreasing the percentage of lymphocytes. Proof of this claim with a (− 0.62) correlation is evident in Fig. 5. Also, the negative correlation between CRP and lymphocytes (− 0.32) confirms the increasingly opposite relationship between CRP index and lymphocyte's index.

According to experts' opinion from Masih Daneshvari hospital in Tehran, although these indicators alone are not enough to confirm COVID-19 infection, they can be useful together. Very small correlations such as LD and lymphocytes indicate that having more blood's clinical indicators affected by COVID-19 is directly related to more accurate diagnosis and prediction of COVID-19.

According to Fig. 7, Tables1, and2, the CRP trend indicates that the CRP test with a range of (344–1) has an average of 83.26 in living people and an average of 122.6 in dead people. Also, the CRP level in women with an average of 74.28 is significantly lower than the average CRP in men with 101.6. This indicates a higher probability of mortality in men than women by comparing CRP levels. According to Table3, the p-value is less than the confidence level ($0.002 \leq 0.05$), so the assumption of the equality of means is rejected and the existence of a significant difference in the amount of CRP index based on gender is confirmed.

LD levels in adults and the elderly typically range from 140 to 280 U/L [33]. In Fig. 8, the LD level has increased dramatically, and most of the data have taken values from

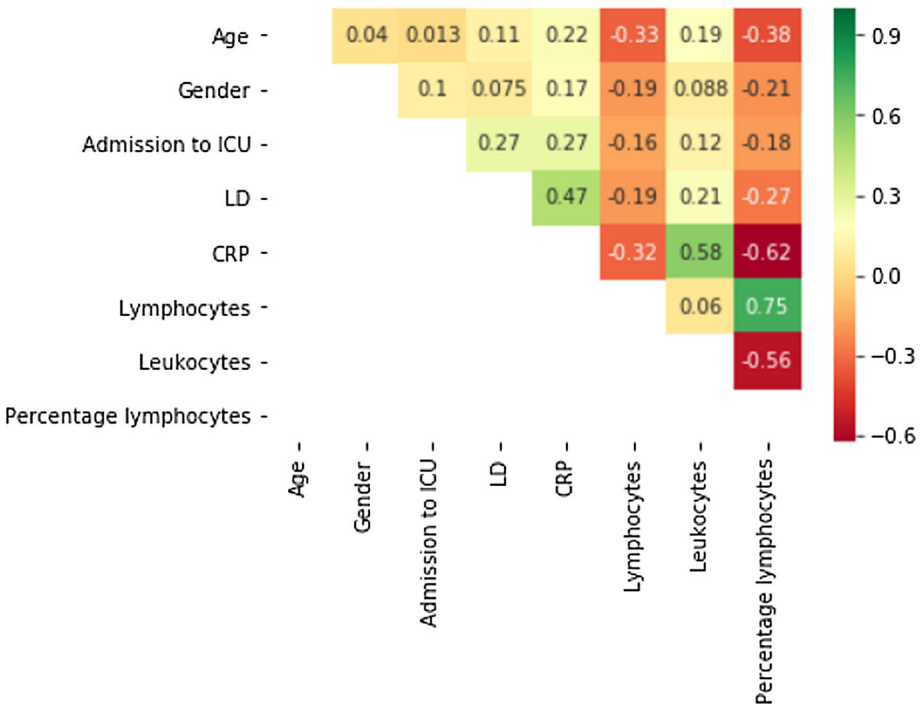


Fig. 5 Correlation matrix including the outlier

Fig. 6 Correlation map excluding the outlier

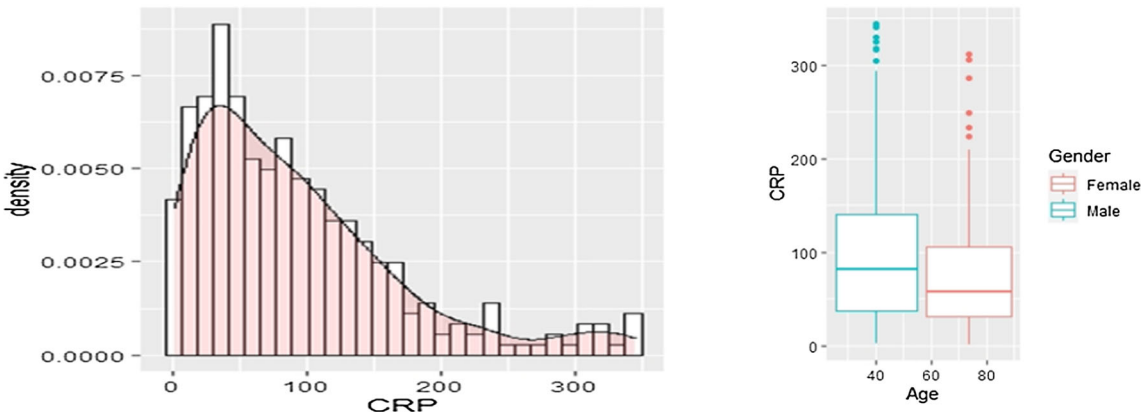


Fig. 7 Boxplot the effect of CRP on the age variable by gender

Table 1 Descriptive statistics of CRP

Variable	Gender	N	N*	Mean	SE ,mean	StDev	Minimum	Q1	Median	Q3	Maximum
CRP	Female	117	0	74.28	6.08	65.80	1.00	30.50	57.00	105.50	312.00
	Male	188	0	101.61	5.97	81.92	2.00	37.00	81.50	141.50	344.00

Table 2 Mean and median of CRP

Variable	Survival/death	Mean	Median
CRP	Alive	83.26	66.00
	Deceased	122.6	107.0

300 to 400. The LD, CRP, and percentage lymphocytes data distributions are close to the normal distribution.

According to the results obtained from Fig. 8 and Table4, the average LD in the deceased patients is higher than the average of the survived patients for both the survived and deceased groups.

Table 3 Result of the paired t-test of CRP

Variable	Gender	Difference	95% CI for difference	T-value	DF	*p*-value
CRP	Female–male	− 119.8	(− 44.11 = − 10.55)	−3.21	283	0.002

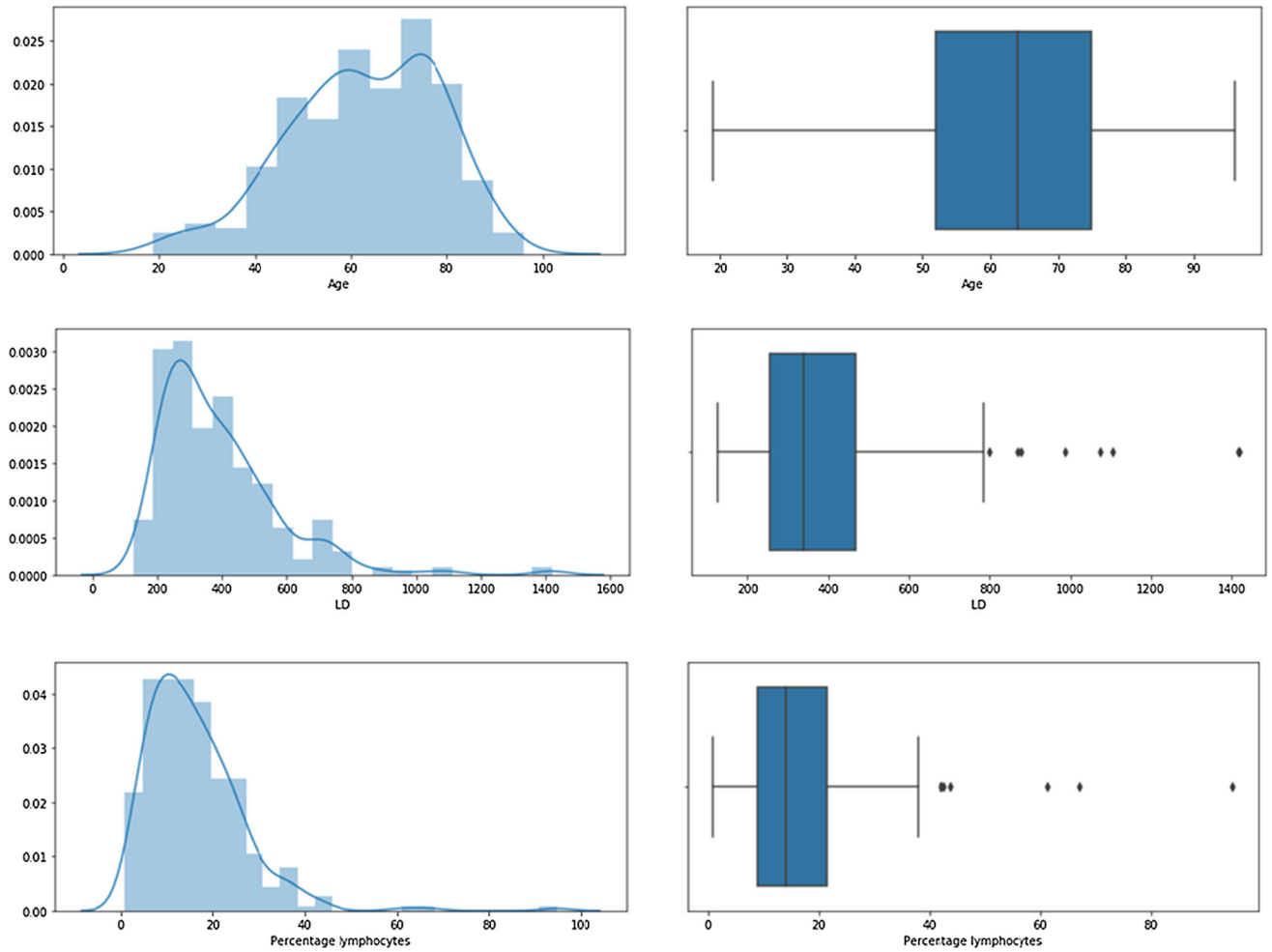


Fig. 8 Dispersion of indicators and boxplots

Table 4 Descriptive statistics of LD

Variable	Survival/death	Mean	Skewness	Kurtosis
LD	Alive	365.5	1.93	6.50
	Deceased	485.3	1.88	5.33

Table 5 Assumptions considered for LD based on survival/death

Null hypothesis	H_0: The population means are all equal
Alternative hypothesis	Not all means are equal
DF	78
95% CI for difference	(− 179.9, − 59.7)
p-value	$p < 0.001$
T-value	− 3.97

According to Table5, the*p*-value is less than the confidence level of alpha (0.05), which indicates the rejection of the null hypothesis and the existence of a significant difference in the mean of dead and living people in terms of LD value. It can be concluded that a higher LD level in blood samples increases the risk of death.

High levels of protein C have been observed in 86% of COVID-19 patients. Also, due to the deterioration of the patient's condition, a direct relationship was observed with the level of protein C [31]. In Fig. 8, we see an increase in

blood CRP levels in the patient's blood. Normal CRP levels are generally less than 10 mg /L [34]. Also, in Fig. 8 the level of most data in this section is between 10 and 20%, which decreases the percentage of lymphocytes due to the presence of disease and the involvement of white blood cells. In adults, the approximate percentage of lymphocytes is generally 20 to 40% [35].

3.2 Preprocessing

The raw dataset consists of fifteen columns. First, in order to prepare the dataset for training machine learning models, five columns which were related to dates, such as "Date of Presentation Emergency Room," "Date of Admission," "Date of Discharge," were removed for two reasons. First, a great number of rows had missing values (no data were recorded). Second, the main goal of this study is to predict patients' survival or death with respect to their blood characteristics, age, gender, and admission to ICU and removed columns were not useful. Second, the "Survival or Death" column is selected as a label. Third, several categorical data columns such as Gender, Admission to ICU, and Survival or Death were converted to numerical data. Forth, the dataset is imbalanced and should be balanced because it could affect machine learning algorithms' performance. Imbalanced data are a case where the dataset has a skewed proportion of each class. According to Fig. 9, the dataset is imbalanced, and even getting high accuracy is misleading because this accuracy stems from predicting the majority class correctly. Simultaneously, machine learning algorithms perform poorly in predicting the minority class; various methods tackle imbalanced data. This paper has implemented the synthetic minority oversampling technique for nominal and continuous (SMOTENC) method to balance the dataset because this technique has been successfully used in similar previous studies [36]. SMOTENC creates synthetic data for the minority class rather than oversampling with replacement [37]. Finally, dataset has been normalized since each column has a different range that might affect machine learning algorithms' performance.

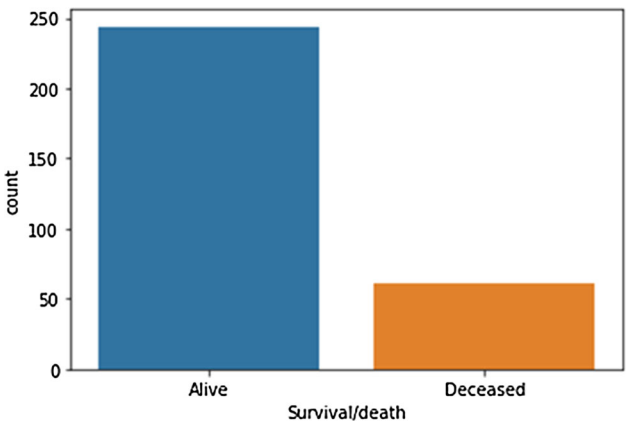


Fig. 9 Number of samples in each class

3.3 Training machine learning models

In similar previous studies, researchers have employed various machine learning algorithms to facilitate the decision-making process or extract useful information to provide better patient service in hospitals or clinics. Among machine learning algorithms, some are more popular and are widely used by researchers. In this paper, we use and compare the most adopted machine learning algorithms in the literature that proved to be suitable for datasets related to COVID-19 (e.g., [38–40]). We used algorithms such as KNN, DT, LR, SVM, RF, SGD, BC, and AdaBoost to find the best prediction model.

Here is a brief description of each model and a general comparison between them. KNN is one of the simplest supervised machine learning algorithms that rely on the hypothesis "things that look alike" [41]. KNN uses existing distance metrics to measure similarities between two data points. It also decides on a hypothetical observation based on the closest distance [42]. DT is generally drawn in reverse. An experiment is an internal node that occurs on a property, and the test result is called a branch. Finally, the tree leads to the leaf nodes that are the class tag [43]. LR uses some weights and coefficients on input values and combines them linearly to predict the output. [44]. The algorithm is used for classification to find out a single Boolean expression that predicts a binary outcome. In a regression, many Boolean expressions can be investigated and simultaneously embedded into a linear regression model [45].

The SVM finds the best superplane and separates the data points based on their superplane distance. A superplane with a maximum margin would be best [46]. RF and gradient boosting combine the results of a DT for better prediction. Also, gradient boosting is ensemble tree-based methods applying the principle of gradient descent [47]. The RF divides the data into some random subset and trains them in parallel, ultimately using the majority of votes for the final prediction. Gradient boosting works so that each model considers the previous model's mistakes and learns to predict them better [48]. SGD uses a small randomly-selected subset of the training samples to approximate the objective function's gradient [49]. It is a simple yet very efficient approach to fitting linear classifiers and regressors under convex loss functions such as (linear) SVM and LR [50].

Implementing machine learning models in python or other distributors like Jupyter and Spyder are almost the

same. Every machine learning model has one or several hyperparameters that should be adjusted. For this purpose, cross-validation with 10 folds is used. After tuning hyperparameters, the dataset is split, 0.8 for training and 0.2 for testing. Figure 10 is an illustration of both preprocessing and implementing machine learning algorithms.

4 Computational results

4.1 Accuracy, mean error, sensitivity, and specificity

After implementing machine learning algorithms, results will be examined in this part. The performance of machine learning algorithms is compared based on four metrics which are presented in Table6. Mean absolute error (MAE) is a measure of errors between paired observations expressing the same phenomenon. Accuracy is the proximity of measurements to a specific value. It is also used as a statistical criterion to assess whether a binary classification test correctly identifies or removes a condition [51]; in other words, ability of a machine-learning algorithm to predict or classify positive and negative samples correctly. Specifity refers to classifying negative samples correctly, while classifying positive samples correctly is called sensitivity.

Table 6 Analysis of machine learning algorithms

Algorithm	MAE	Accuracy	Sensitivity	Specificity
DT	0.089082	0.9163	0.98	0.85
AdaBoost	0.086118	0.9132	0.97	0.85
RF	0.09338	0.9080	0.98	0.84
KNN	0.092357	0.9078	0.97	0.84
SVM	0.091612	0.9063	0.96	0.84
BC	0.138188	0.8617	0.96	0.77
LR	0.212655	0.7888	0.86	0.72
SGD	0.230922	0.7697	0.82	0.72

As shown in Table6 and Fig. 11, the lowest error rate in all three calculated error categories is related to the AdaBoost model. Besides, all machine algorithms perform better in predicting recovered patients in comparison to predicting deceased patients. According to obtained results, it seems that DT, AdaBoost, RF, KNN, and SVM outperform other algorithms with respect to accuracy. However, this claim will be tested using statistical hypothesis tests in the next section.

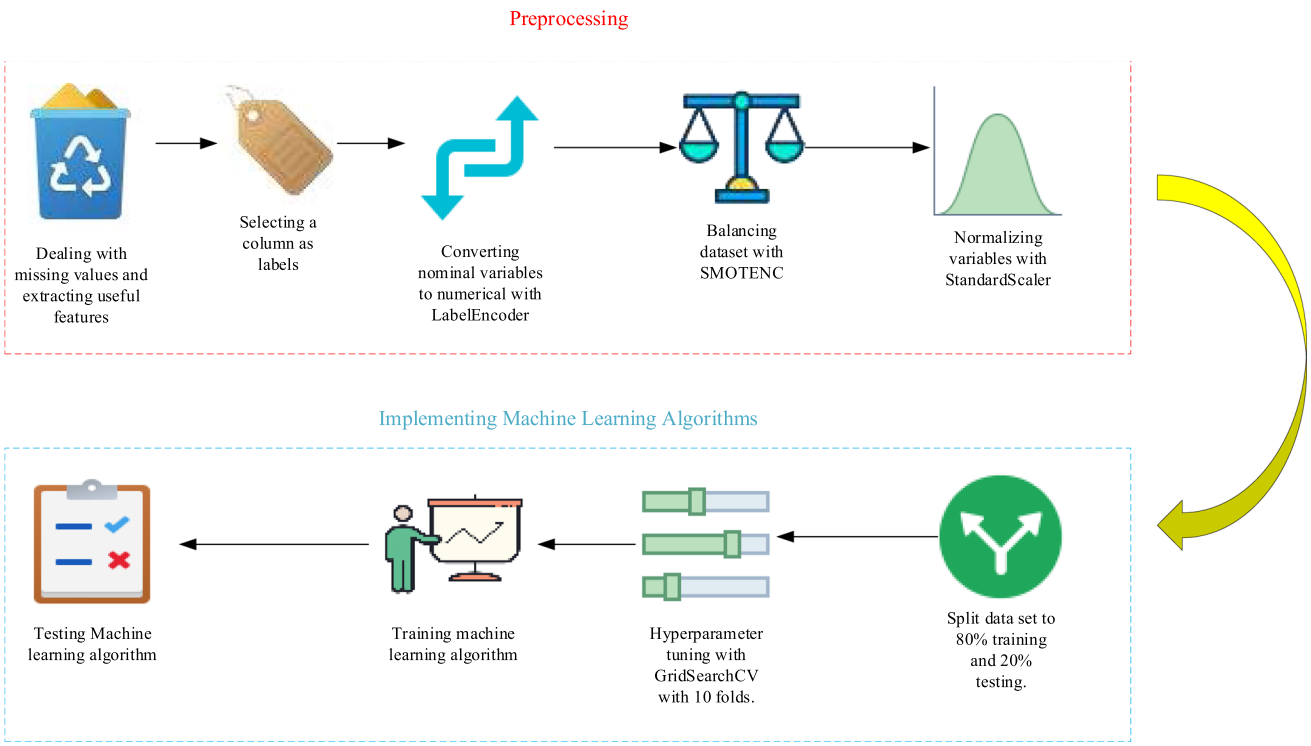


Fig. 10 Graphical view of preprocessing and implementation of machine learning algorithms

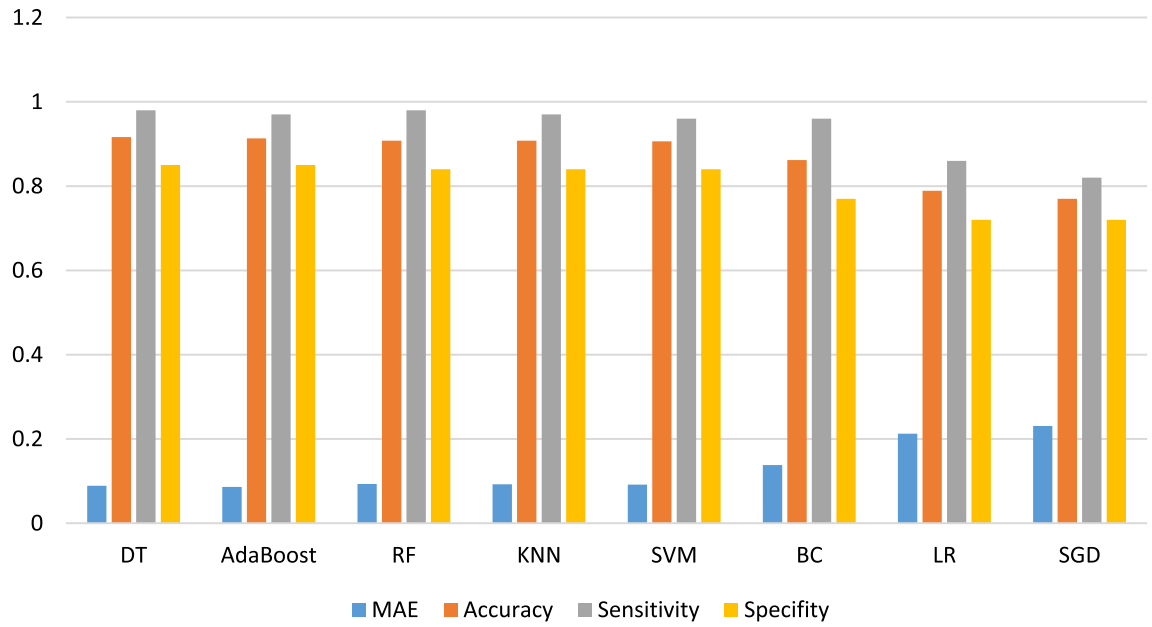
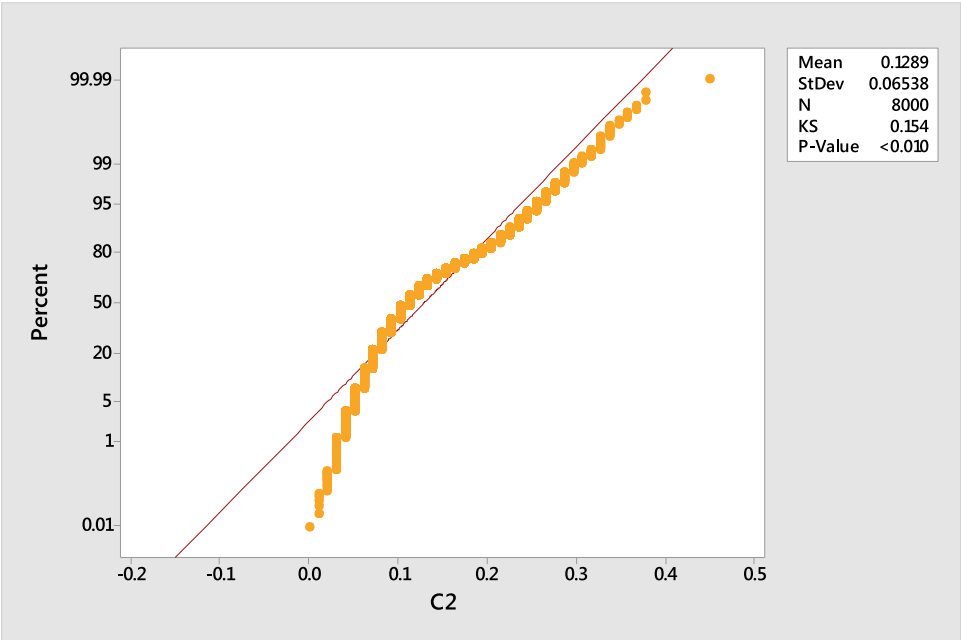


Fig. 11 Accuracy and mean absolute error of machine learning algorithms

Fig. 12 Probability plot



4.2 Results and comparative study

The dataset should follow a normal distribution to use the parametric test, and communities should be independent [52]. For this purpose, we used the Kolmogorov–Smirnov (K–S) method to test the normality. The results of this test are well illustrated in Fig. 12.

The p-value obtained is less than the significance level of 0.05. Therefore, the distribution of these data does not follow the normal distribution, so using parametric tests such as ANOVA is not correct. Accordingly, we use non-parametric tests because they do not have any specific assumption about the probability distribution of data [53].

The Kruskal–Wallis test is a nonparametric test which is an extension of the Mann–Whitney U test [54]. It is used to compare the mean of two or more populations to understand the difference or equality between the mean of populations. The result of the Kruskal–Wallis test is displayed in Table7.

Table 7 Assumptions considered for hypothesis testing

Null hypothesis	H_0: Mean of accuracies are equal
Alternative hypothesis	H_1: At least one of the accuracies are significantly different
Statistic	5079.4627
p-Value	$p < 0.001$

According to Table7, The*p*-value is$p < 0.001$, and the null hypothesis is rejected at a significance level of 0.05 which means that at least one machine learning algorithm's performance is significantly different.

Kruskal–Wallis test indicated that there is a difference in at least one machine learning algorithms' accuracy. Nevertheless, it does not specify which algorithms have different performances, so to identify algorithms differing from each other, we employed Dunn's test, a nonparametric multiple comparison test developed by Charles Dunnett [55]. This test is used to find significant differences between independent groups. There are no assumptions about the type of distribution of the data and groups can be equal or unequal in size [56].

The results in Table8 show that at the significance level of α equal to 0.05, RF, SVM, DT, AdaBoost, and KNN algorithms have statistically the same performance.

Figure 13 shows the importance of the features in DT. According to Fig. 13, the most important factors in assessing and analyzing whether people with COVID-19 survived or deceased were age, LD, and leukocytosis, respectively. However, LR coefficients show completely different results. Based on Fig. 14, admission to ICU is a decisive factor in predicting whether patients stay alive or pass away.

For better analysis, the DT diagram was drawn in Python with a depth of 19 (Max depth = 19) to measure the importance of the variables involved in this evaluation from a decision tree perspective. Age has the greatest impact on the classification and was considered as the main branch (Fig. 15).

Based on previous studies, not only can predictive systems, based on machine learning algorithms, effectively answer many complex medical questions in the shortest possible time, but also, they lead to policy adoption and proper planning. In this study, eight machine learning algorithms that are used in similar studies are compared with each other. It was proved that RF, SVM, DT, KNN, and AdaBoost outperform other machine learning algorithms, so these algorithms can be used to predict whether a patient survives or passes away with several features.

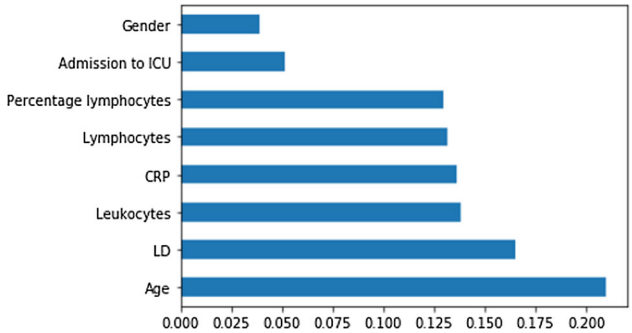


Fig. 13 Feature importance in the DT

By examining this dataset, it was found that the mortality rate due to COVID-19 is more common in older patients. Also, it was observed that LD and CRP have a higher rate in older patients than the younger ones. Since LD and CRP have a positive correlation with age, it could be the reason for the increase in lung infection and severity among older patients. Moreover, the average of LD level is significantly different for the deceased and recovered individuals. It was found that any increase in the percentage of lymphocytes is an important sign because a high percentage of lymphocytes was observed among deceased patients.

Table 8 Dunn's test result

Index	AdaBoost	RF	BC	KNN	LR	SGD	SVM	DT
AdaBoost	1	0.837	$p < 0.001$	0.0524	$p < 0.001$	$p < 0.001$	0.0697	1
RF		1	$p < 0.001$	1	$p < 0.001$	$p < 0.001$	1	1
BC			1	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$
KNN				1	$p < 0.001$	$p < 0.001$	1	1
LR					1	0.2316	$p < 0.001$	$p < 0.001$
SGD						1	$p < 0.001$	$p < 0.001$
SVM							1	1
DT								1

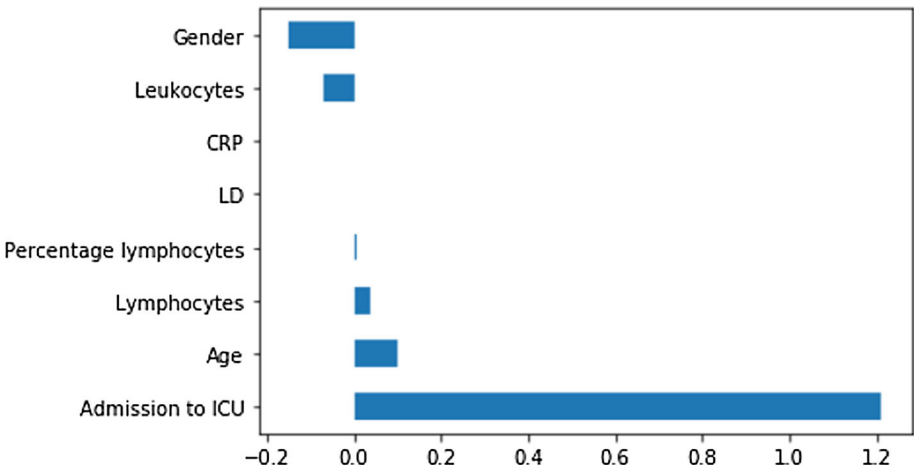


Fig. 14 Feature importance in LR

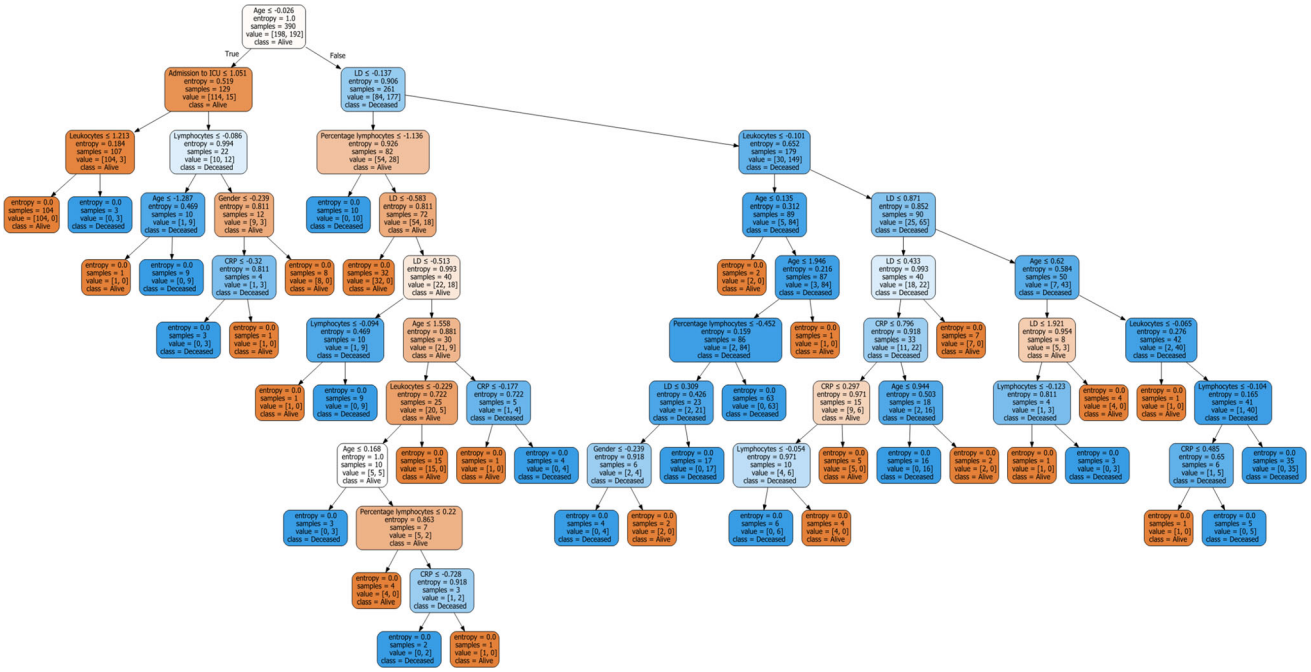


Fig. 15 A schematic view of DT

The existence of a high negative correlation between CRP and percentage lymphocytes showed that these two indicators work in opposite directions in recovered individuals; when CRP level decreases and percentage lymphocytes increase, there is a reduction in mortality rate. CRP in men and women was examined separately, and according to the statistical test, it is concluded that women have lower CRP levels than men.

5 Conclusion

The COVID-19 pandemic is the most crucial health disaster surrounding the world for the past two years. This study used a dataset of blood samples from 306 infected patients to analyze blood's clinical indicators and to compare the performance of the eight machine learning algorithms. For this purpose, the clinical parameters of patients' blood and their effect on patients' survival or death were analyzed. The results showed that the number of lymphocytes and leukocytes in the blood test has a very high effect on each other, and age was also the most influential variable among other factors.

Eight commonly used machine learning algorithms such as KNN, DT, LR, SVM, RF, SGD, BC, and AdaBoost to predict survival or death of COVID-19 patients were implemented on the dataset. According to the statistical hypothesis tests, it turned out that RF, SVM, DT, KNN, and AdaBoost produce more accurate results. This study's findings can be used to prioritize high-risk patients through the results of their blood data.

Declarations

Conflict of interest The authors declare that they have no conflict of interest

Human and animal rights This article does not contain any studies with human participants or animals performed by the author.

Consent for publications The undersigned authors declare that this manuscript is original, has not been published before, and is not currently being considered for publication elsewhere.

References

1. WHO (2020) Coronavirus disease 2019 (COVID-19) Situation Report - 43, 8 March
2. USAFacts. Nonpartisan Government Data. Available online:https://usafacts.org/ (accessed on 7 August 2020)
3. Zhang Y, Jiang B, Yuan J, Tao Y (2020) The impact of social distancing and epicenter lockdown on the COVID-19 epidemic in mainland China: A data-driven SEIQR model study. *MedRxiv*
4. Li Z, Yi Y, Luo X, Xiong N, Liu Y, Li S, Ye F (2020) Development and clinical application of a rapid IgM-IgG combined antibody test for SARS-CoV-2 infection diagnosis. J Med Virol 92(9):1518–1524
5. Ferrari D, Motta A, Strollo M, Banfi G, Locatelli M (2020) Routine blood tests as a potential diagnostic tool for COVID-19. Clin Chem Lab Med (CCLM) 58(7):1095–1099
6. Ahmad A, Garhwal S, Ray SK, Kumar G, Malebary SJ, Barukab OM (2021) The number of confirmed cases of covid-19 by using machine learning: Methods and challenges. Arch Comput Meth Eng 28(4):2645–2653
7. https://www.kaggle.com/plarmuseau/forecast-covid-death
8. Assaf D, Gutman YA, Neuman Y, Segal G, Amit S, Gefen-Halevi S, Tirosh A (2020) Utilization of machine-learning models to accurately predict the risk for critical COVID-19. Intern Emerg Med 15(8):1435–1443
9. Arvind V, Kim JS, Cho BH, Geng E, Cho SK (2021) Development of a machine learning algorithm to predict intubation among hospitalized patients with COVID-19. J Crit Care 62:25–30
10. Chaurasia V, Pal S (2020) Application of machine learning time series analysis for prediction COVID-19 pandemic. Research on Biomedical Engineering, 1–13
11. Li S, Lin Y, Zhu T, Fan M, Xu S, Qiu W, Xu S (2021) Development and external evaluation of predictions models for mortality of COVID-19 patients using machine learning method. Neural Computing and Applications, 1–10
12. Muhammad LJ, Islam MM, Usman SS, Ayon SI (2020) Predictive data mining models for novel coronavirus (COVID-19) infected patients' recovery. SN Comput Sci 1(4):1–7
13. Brinati D, Campagner A, Ferrari D, Locatelli M, Banfi G, Cabitza F (2020) Detection of COVID-19 infection from routine blood exams with machine learning: a feasibility study. J Med Syst 44(8):1–12
14. Khakharia A, Shah V, Jain S, Shah J, Tiwari A, Daphal P, Mehendale N (2021) Outbreak prediction of COVID-19 for dense and populated countries using machine learning. Ann Data Sci 8(1):1–19
15. Hussain L, Nguyen T, Li H, Abbasi AA, Lone KJ, Zhao Z, Duong TQ (2020) Machine-learning classification of texture features of portable chest X-ray accurately classifies COVID-19 lung infection. Biomed Eng Online 19(1):1–18
16. Yasar H, Ceylan M (2021) A novel comparative study for detection of Covid-19 on CT lung images using texture analysis, machine learning, and deep learning methods. Multim Tool Appl 80(4):5423–5447
17. Sharma S (2020) Drawing insights from COVID-19-infected patients using CT scan images and machine learning techniques: a study on 200 patients. Environ Sci Pollut Res 27(29):37155–37163
18. Khanday AMUD, Rabani ST, Khan QR, Rouf N, Din MMU (2020) Machine learning based approaches for detecting COVID-19 using clinical text data. Int J Inf Technol 12(3):731–739
19. Jiang X, Coffee M, Bari A, Wang J, Jiang X, Huang J, Huang Y (2020) Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity. Comput, Mater Contin 63(1):537–551
20. Vijayakumar DS, Sneha M (2021) Low cost Covid-19 preliminary diagnosis utilizing cough samples and keenly intellective deep learning approaches. Alexandria Eng J 60(1):549–557
21. Qian Z, Alaa AM, van der Schaar M (2021) CPAS: the UK's national machine learning-based hospital capacity planning system for COVID-19. Mach Learn 110(1):15–35
22. Sujatha R, Chatterjee J (2020) A machine learning methodology for forecasting of the COVID-19 cases in India
23. Shrivastav LK, Jha SK (2021) A gradient boosting machine learning approach in modeling the impact of temperature and humidity on the transmission rate of COVID-19 in India. Appl Intell 51(5):2727–2739
24. Albahri AS, Hamid RA, Alwan JK, Al-Qays ZT, Zaidan AA, Zaidan BB, Madhloom HT (2020) Role of biological data mining and machine learning techniques in detecting and diagnosing the novel coronavirus (COVID-19): a systematic review. J Med Syst 44:1–11
25. Shuja J, Alanazi E, Alasmary W, Alashaikh A (2021) COVID-19 open source data sets: a comprehensive survey. Appl Intell 51(3):1296–1325
26. Behnam A, Jahanmahin R (2021) A data analytics approach for COVID-19 spread and end prediction (with a case study in Iran). Model Earth Sys Environ 8(1):579–589
27. Rahimi I, Chen F, Gandomi AH (2021) A review on COVID-19 forecasting models. Neur Comput Appl.https://doi.org/10.1007/s00521-020-05626-8
28. Singh SA, Masuku BM (2014) Assumption and testing of normality for statistical analysis. Am J Math Math Sci 3(1):169–175
29. Stegeman I, Ochodo EA, Guleid F, Holtman GA, Yang B, Davenport C et al (2020) Routine laboratory testing to determine if a patient has COVID-19. Cochrane Data Sys Rev.https://doi.org/10.1002/14651858.CD013787
30. Wang L (2020) C-reactive protein levels in the early stage of COVID-19. Med et maladies infectieuses 50(4):332–334
31. Ali N (2020) Elevated level of C-reactive protein may be an early marker to predict risk for severity of COVID-19. J Med Virol.https://doi.org/10.1002/jmv.26097

32. Panteghini M (2020) Lactate dehydrogenase: an old enzyme reborn as a COVID-19 marker (and not only). Clin Chem Lab Med (CCLM) 58(12):1979–1981
33. Henry BM, Aggarwal G, Wong J, Benoit S, Vikse J, Plebani M, Lippi G (2020) Lactate dehydrogenase levels predict coronavirus disease 2019 (COVID-19) severity and mortality: a pooled analysis. Am J Emerg Med 38(9):1722–1726
34. Fischbach FT, Dunning MB (2009) A manual of laboratory and diagnostic tests. Lippincott Williams & Wilkins, Pennsylvania
35. Jacob EA (2016) Complete blood cell count and peripheral blood film, its significant in laboratory medicine: a review study. Am J Lab Med 1(3):34
36. Pahar M, Klopper M, Warren R, Niesler T (2021) COVID-19 Cough classification using machine learning and global smartphone recording. Comput Biol Mede.https://doi.org/10.1016/j.compbiomed.2021.104572
37. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357
38. Alballa N, Al-Turaiki I (2021) Machine learning approaches in COVID-19 diagnosis, mortality, and severity risk prediction: a review. Info Med Unlock.https://doi.org/10.1016/j.imu.2021.100564
39. Pourhomayoun M, Shakibi M (2021) Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making. Smart Health 20:100178
40. Sharma S, Gupta YK (2021) Predictive analysis and survey of COVID-19 using machine learning and big data. J Interdisc Math 24(1):175–195
41. Mathkunti NM, Rangaswamy S (2020) Machine learning techniques to identify dementia. SN Comput Sci 1(3):1–6
42. Subramanian D (2019) A simple introduction to K-Nearest Neighbors Algorithm. Towards Data Science.https://towardsdatascience.com/a-simple-introduction-to-k-nearest-neighborsalgorithm-b3519ed98e
43. Brid RS. Decision Trees-A simple way to visualize a decision.https://medium.com/greyatom/decision-trees-a-simple-way-to-visualize-a-decisiondc506a403aeb
44. Logistic regression for machine learning. Retrieved October 1, 2019, from https:// machinelearningmastery.com/logistic-regression-for-machine-learning
45. Schwender H, Ruczinski I (2010) Logic regression and its extensions. Adv Genet 72:25–45.https://doi.org/10.1016/B978-0-12-380862-2.00002-3
46. Support vector machine – Introduction to machine learning algorithms. Retrieved October 1, 2019, fromhttps://towardsdatascience.com/support-vector-machineintroduction-to-machine-learning-algorithms-934a444fca47 .
47. Samieinasab M, Torabzadeh SA, Behnam A, Aghsami A, Jolai F (2022) Meta-health stack: a new approach for breast cancer prediction. Healthc Analyt 2:100010
48. Basic ensemble learning (Random Forest, AdaBoost, Gradient Boosting) – Step by step explained. Retrieved October 1, 2019, fromhttps://towardsdatascience.com/basicensemble-learning-random-forest-adaboost-gradient-boosting-step-by-stepexplained-95d49d1e2725 .
49. Tsuruoka Y, Tsujii JI, Ananiadou S (2009) Stochastic gradient descent training for l1-regularized log-linear models with cumulative penalty. In Proceedings of the Joint Conference of the 47th o International Joint Conference on Natural Language Processing of the AFNLP, pp 477–485
50. "Stochastic Gradient Descent" L. Bottou - Website, 2010.
51. YK Mohammed (2015) Re: how can I calculate the accuracy? Retrieved from:https://www.researchgate.net/post/How_can_I_calculate_the_accuracy/56784cb96225ff47c88b45c3/citation/download
52. Kayvanfar V, Husseini SM, Karimi B, Sajadieh MS (2017) Bi-objective intelligent water drops algorithm to a practical multi-echelon supply chain optimization problem. J Manuf Syst 44:93–114
53. Teymourian E, Kayvanfar V, Komaki GM, Zandieh M (2016) Enhanced intelligent water drops and cuckoo search algorithms for solving the capacitated vehicle routing problem. Inf Sci 334:354–378
54. Muenchen RA (2009) R for SAS and SPSS users. Springer, Berlin
55. Dunnett CW (1955) A multiple comparison procedure for comparing several treatments with a control. J Am Stat Assoc 50(272):1096–1121
56. Dinno A (2015) Nonparametric pairwise multiple comparisons in independent groups using Dunn's test. Stand Genomic Sci 15(1):292–300