Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 11, No. 4, 2023, 573-585.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

Fast COVID-19 Infection Prediction with In-House Data Using Machine Learning Classification Algorithms: A Case Study of Iran

Ali Shabrandi[1*], Ali Rajabzadeh ghatari[1], Nader Tavakoli[2], Mahmoud Dehghan Nayeri[1] and Sahar Mirzaei[3]

1. Department of Industrial Management, Faculty of Management and Economics, Tarbiat Modares University, Tehran, Iran.
2. Department of Emergency Medicine, Trauma and Injury Research Center, Iran University of Medical Sciences, Tehran, Iran.
3. Department of Health and Environment, Iran University of Medical Sciences, Tehran, Iran.

Article Info

Article History:
Received 27 July 2023
Revised 18 September 2023
Accepted 28 October 2023

DOI:10.22044/jadm.2023.13291.2458

Keywords:
Covid-19, Symptomatic, Machine Learning, Classification, Artificial Intelligence.

**Corresponding author: A.Shabrandi@Modares.ac.ir (A. Shabrandi).*

Abstract

To mitigate COVID-19's overwhelming burden, a rapid and efficient early screening scheme for COVID-19 in the first-line is required. Much research has utilized laboratory tests, CT scans, and X-ray data, which are obstacles to agile and real-time screening. In this study, we propose a user-friendly and low-cost COVID-19 detection model based on self-reportable data at home. The most exhausted input features were identified and included in the demographic, symptoms, semi-clinical, and past/present disease data categories. We employed Grid search to identify the optimal combination of hyperparameter settings that yields the most accurate prediction. Next, we apply the proposed model with tuned hyperparameters to 11 classic state-of-the-art classifiers. The results show that the XGBoost classifier provides the highest accuracy of 73.3%, but statistical analysis shows that there is no significant difference between the accuracy performance of XGBoost and AdaBoost, although it proved the superiority of these two methods over other methods. Furthermore, the most important features obtained using SHapely Adaptive explanations were analyzed. "*Contact with infected people*," "*cough*," "*muscle pain*," "*fever*," "*age*," "*cardiovascular comorbidities*," "PO_2," and "*respiratory distress*" are the most important variables. Among these variables, the first three have a relatively large positive impact on the target variable, whereas "*age*," "PO_2", and "*respiratory distress*" are highly negatively correlated with the target variable. Finally, we built a clinically operable, visible, and easy-to-interpret decision tree model to predict COVID-19 infection.

1. Introduction

In December 2019, a novel coronavirus was reported in Wuhan, China, caused by the new "severe acute respiratory syndrome coronavirus 2", later named COVID-19 by the World Health Organization (WHO) in February 2020. Within a short period, this epidemic spread from China to more countries, and less than three months after the epidemic began; the WHO declared it a global pandemic [1]. COVID-19 is a highly contagious respiratory infection that can be considered the greatest challenge faced by humanity since World War II. As of today (June 21, 2023), there are over 768.2 million confirmed cases, and the number of people infected is probably much higher. There are also more than 6,954,731 confirmed deaths, according to the WHO [2]. The COVID-19 sudden outbreak imposes an overwhelming burden on countries' medical systems through an increase in the demand for hospital beds and a shortage of medical equipment, while medical staff themselves can also become infected [3]. To mitigate this burden, countries worldwide have taken various measures to cope with the spread of COVID-19, including curfews, lockdowns, and travel restrictions. These actions are needed, but are not

sufficient to cut off the source of infection. Therefore, a rapid and efficient early screening scheme for COVID-19 in the first-line is momentous in several ways. First, it prevents the source of infection, helps government agencies prevent its spread, and saves lives. Secondly, early treatment can be initiated immediately, and limited treatment resources are prioritized for patients at a higher risk of mortality. Thirdly, it helps optimize the allocation of limited health system resources [4, 5].

A high volume of research is carried out on how to detect COVID-19 coronavirus in the first-line through an accurate, real-time, and fast-screening scheme to take emergency or suitable preventive actions. Methods based on Reverse Transcriptase Polymerase Chain Reaction (RT-PCR) have become the gold standard for confirming that individuals with COVID-19 have active shedding of SARS-CoV-2 [6]. Although RT-PCR is considered an efficient and competent test kit, its wide-scale screening of patients at the screening or testing centers has several limitations such as the high mounting demand for testing kits in the initial phases of pandemic emergence; inequitable testing kit distribution, especially in developing countries [7]; test results ranging from one to more days, especially in rural areas [8]; the need for specialized laboratories with specialized equipment and trained staff [9]; and the high cost of RT-PCR, especially in single versus batches of samples run. In addition, the use of public transportation to visit screening centers increases the vulnerability to COVID-19 spread, as it increases the risk of infection due to the interaction between COVID-19-negative and COVID-19-positive patients.

Despite public health efforts aimed at improving testing [10], prevention strategies [11], and scientific advances in vaccination programs [12], the severe situation has not yet been effectively controlled. Smart technologies including machine learning [13, 14], blockchain [15], and the Internet of Things [16] have been used to overcome the challenges posed by COVID-19. Machine learning (ML) is a branch of artificial intelligence (AI) that provides systems with the ability to automatically learn and improve from experience without being explicitly programming [17]. Here, we present a ML-based detection model that predicts a positive SARS-CoV-2 infection in an RT-PCR test by asking some basic questions that can be used in the frontline of fighting against COVID-19.

Considering these issues, we propose a model for detecting COVID-19 based on machine learning within minutes, and the results of the model can be used to assist doctors in finding suitable preventive measures. Different classification algorithms including Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Naive Bayes (NB), Support Vector Classifier (SVC), Decision Tree (DT), Random Forest (RF), K-Nearest-Neighbor (KNN), Adaptive Boosting (AdaBoost), Extreme Gradient Boosting (XGBoost), Logistic regression (LR), and Extra Tree classifier (ETC) are used to evaluate the model. We used grid optimization to tune the hyperparameters of the classification algorithms. Different performance metrics such as accuracy, sensitivity, and specificity were used to compare the classification performance of the above algorithms.

2. Related Works

ML, as a subset of AI, has been successfully applied in many fields such as healthcare [18]. AI and ML can be used to improve diagnosis, prognosis, monitoring, and treatment delivery to improve patient health outcomes [19]. Since the beginning of the COVID-19 pandemic, an increasing number of people have become interested in using ML to fight the pandemic. One of the areas where ML is used is the screening phase of the pandemic management of COVID-19. An effective screening scheme leads to rapid diagnosis of COVID-19, thereby reducing the burden on healthcare systems. In the literature, COVID-19 infection prediction models have used features such as CT scans [20-24], clinical symptoms [25], laboratory tests [26, 27], and/or a combination of these features [28]. Each detection method has its drawbacks; for example, CT-based models require expensive equipment and professional staff, exposure patients to unnecessary irradiation [29], and overwhelming use of the limited resources of the health system. However, most previous models were based on data from hospitalized patients and thus were not effective for the fast screening of SARS-CoV-2 in the general population. Therefore, all the studies whose input data was not self-reportable are outside the scope of this study. A comprehensive but not exhaustive review of the infection prediction of COVID-19 based on self-reported data is summarized in Table 1. Eight studies were compared in several aspects including the type and size of the dataset, number and type of input features, classification methods, and performance indicators.

Access to COVID-19 data is a complex process owing to different government policies and regulations regarding data sharing. Additionally, the lack of standardized formats and varying levels

of data transparency across countries further complicate the accessibility and analysis of the COVID-19 data. As can be seen in Table 1, only three studies ([30], [31], and [32]) have more than 20,000 records in their dataset, and only four studies ([30], [31], [32], and [33]) have a real dataset. The limited availability of large datasets hinders comprehensive analysis and modeling of the pandemic. Furthermore, variations in data collection methods and reporting standards across regions further complicate the comparison and aggregation of the COVID-19 data. Although there is no consensus in the literature regarding a sufficient number of samples [34], the results of small-sample studies are more susceptible to minor analytical errors that result in false-negative results [35, 36]. The researchers are advised to conduct large-scale studies that can produce statistically realistic effects, owing to their higher statistical power. Results from large studies are statistically more reliable than those from small studies because of the reduced risk of increased effect size and lower Type I error [37].

Our literature review categorizes self-reportable input features into three main groups: basic information (demographic data), symptoms, and past or current diseases. The extent to which these features are covered in each study is a key difference between this and previous studies. Studies [38] and [30] focused only on the symptoms, excluding features like "*age*," "*gender*," "*contact with an infected person*," and other variables related to past or current diseases. In [33], in addition to the variables in the basic information category, the focus was on symptoms, but features related to past or current diseases were not included in the model. However, in [31], the researchers expanded their analysis to include the influence of past and current diseases to gain a more comprehensive understanding of the factors contributing to the study outcomes.

Although Iranian researchers, like other researchers, have been at the forefront of research related to COVID-19 [39], a limited number of studies are related to the topic of the current research. Rezaei *et al*. [40] have classified COVID-19 patients by using 767 chest images through a pre-trained convolutional neural network. Heydari *et al*. [41] clustered patients using self-organizing mapping. In a retrospective study, Sobhani *et al*. [42] investigated the relationship between clinical characteristics and laboratory findings for patients using statistical methods. In the most relevant research work, Jamshidi *et al*. [31] studied two models: the symptom prediction model and the mortality prediction model due to infection with COVID-19. They created a symptom prediction model with ROC-AUCs of 0.53–0.78.

Many studies have been carried out recently, and we refer the reader to the recent reviews of AI in combating COVID-19, but what distinguishes our study from other studies is the type of the data we need. The contributions of our paper are as what follows. First, it has the most thorough coverage of features compared to earlier studies by identifying a comprehensive list of input features in the three major categories of demographics, symptoms, and concurrent diseases. Secondly, the proposed model can easily be used with in-house data, so patients do not need to go to COVID-19 screening centers. Thirdly, the health system can optimize its limited resources by minimizing unnecessary exposure to irradiation [29] for COVID-19 detection. Fourthly, the performances of 11 state-of-the-art binary classification machine-learning techniques were compared. Fifthly, the SHapely Adaptive Explanations (SHAP) analysis is used to identify important features and interpret models, resulting in a clinically actionable decision tree that is beneficial to the assigned staff.

Up to now, we have reviewed the literature and specified where our study is located. The remainder of this paper is organized as what follows. In Section 3, we provide a detailed description of the materials and methods used in our study including the data collection process and algorithms employed. Section 4 presents the experimental results obtained from our analysis, and highlights key findings and insights. Section 5 describes the validation methods used to evaluate the generalization of the statistical analysis results. In Section 6, we engage in a comprehensive discussion, and compare our work with various machine learning algorithms employed in similar studies. Finally, in Section 7, we draw meaningful conclusions based on our research findings and propose potential avenues for future research.

Table 1. A literature review of COVID-19 detection model based on self-reportable input data.

Ref.	Year	Dataset	Methods	# of features				Dataset records	AUROC	Drawbacks
				G1	G2	G3	Sum			
[38]	2021	created dataset	AdaBoost, BT, NB, KNN, RF, SVM	0	6	2	8	200	0.97	D1, D2, D3, D4, D5
[43]	2022	created dataset	AdaBoost, DT, KNN, LR, NN, RF	3	9	3	15	200	0.98	D2, D3, D5
[44]	2022	Kaggle	KNN, SVM, LR, MLPNN, GRU and LSTM	1	8	5	14	5,434	0.98	D2, D3
[32]	2021	Real	Gradient-boosting machine model built with DT base-learners	3	5	0	8	99,232	0.90	D4, D5
[45]	2021	Kaggle	J48 DT, RF, NB k-NN, SVM*	1	8	5	14	5434	0.98	D2, D3
[30]	2021	Real	DT, LR, NB, KNN, SVM	0	7	0	7	199,334	0.90	D3, D5
[33]	2022	Real	BT, LR, MLP, RF, SVM	3	15	1	19	4434	0.65	D6
[31]	2021	Real	LR, RF, ANN, KNN, LDA, NB	2	12	14	28	26,189	0.78	-
This study		Real	AdaBoost*, XGBoost, LR, LDA, DT, RF, ETC, NB, KNN, QDA, SVC	6	18	14	38	76,324	0.73	-

Abbreviations: D1: small dataset, D2: unreal dataset, D3: limited attributes in G1, D4: limited attributes in G2, D5: limited attributes in G3, G1: Demographics, G2: Symptoms, G3: Past/current disease, GRU: Gated Recurrent Unit, MLPNN: Multi-layer Perceptual Neural Networks machine learning, LSTM: Long Short-Term Memory deep learning algorithms.

2. Materials and Method

The flow of the research, which refers to our data mining methodology, is adjusted to the step approach in the CRISP-DM method [46]. The overall COVID-19 classification process is illustrated in Figure 1. In Section 1, the problem is well-described, and how it can be converted into a data-mining project is clarified. First, the most time-consuming step is collecting and aggregating data, followed by pre-processing. As we will discuss in more detail later, in the pre-processing step, we cleaned the data. In some records in HIS databases, many useful data were stored in inappropriate data structures; for example, symptoms were typed in the Description field of HIS software by end users.

We used a mini-text mining process to extract the data and import them into the corresponding relevant fields. In step 3, based on an extensive literature review, we chose eleven state-of-the-art classifiers. Next, in a time-consuming step, the hyperparameters of the classifiers are tuned using Grid search. Data mining models are constructed, trained, tested, and their performance criteria are compared by the using Python programming language. Finally, we drew the overall top 20 features' importance in the COVID-19 classification, and traced and explained them well using SHAP analysis.

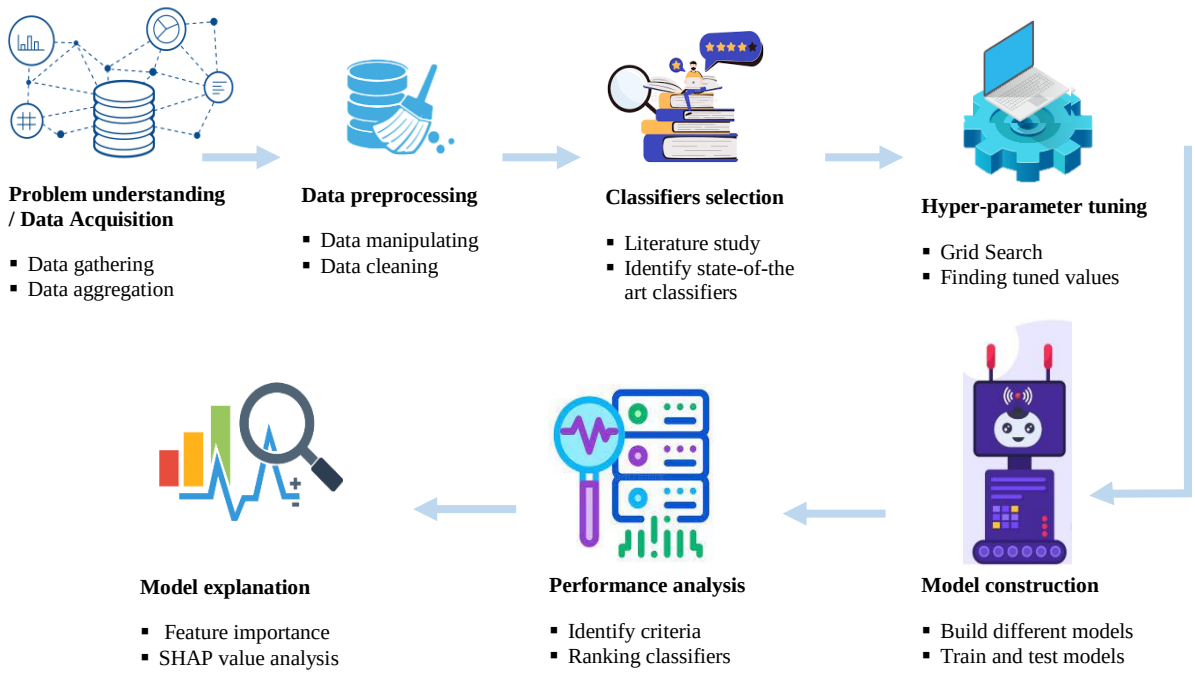

Figure 1. The overall process of the classification COVID-19.

2.1. Data Acquisition

We collected raw data on confirmed or suspected SARS-CoV-2 infections in 145 public and private hospitals under the supervision of Iran University of Medical Sciences, Tehran, Iran. Data was gathered between February 1, 2020, and September 30, 2020, and all ages were included in the data. The criteria to confirm COVID-19 was RT-PCR results. All individuals had undergone an RT-PCR assay of a nasopharyngeal swab at hospital admission. The dataset contains 217,436 records with 40.3% of patients' COVID-19 positive and 29.6% COVID-19 negative tests, and 30.1% missing values. The database has four categories of features: basic information (six features), symptoms (17 features), clinical data (one feature), (1 feature), and comorbidities (14 features). Table 2 provides detailed information regarding these features. Unlike most studies in the COVID-19 detection domain, radiological information was not utilized because most patients did not have radiological details, and it also required time and special equipment. Although in this research, the PO_2 measured in hospitals was used as the only clinical data; it was considered as data that can be measured at home easily. The raw data included positive and negative results for symptomatic and asymptomatic patients. The only integer variable was "*age*", and all others were almost Boolean. It can be further added that percentage of missing value is 12%, and the PO_2 and PCR results have the least and the most missing values, respectively.

2.2. Data pre-processing

After data acquisition, unnecessary fields were eliminated. Among the raw dataset attributes, 37 attributes plus one target attribute RT-PCR test result were required for this research, and others were removed. The dataset contains 217,436 records. The process of cleaning the dataset is shown in Figure 2.

Table 2. Dataset characteristics.

Features	Variable type	Total n=76,324		PCR (+) n=26,945		PCR (-) n=49,379	
		n	%	n	%	n	%
Basic information:							
Sex (Male)	Bionomical	39,632 (51.92)		14,212 (52.74)		25,420 (51.48)	
Age, median (IRQ)	Integer	55 (40-66)		57 (38-68)		54 (40-65)	
Smoker	Binary	1,710 (2.24)		1058 (3.92)		652 (1.32)	
Drug addiction	Binary	1,127 (1.48)		802 (2.97)		325 (0.66)	
Contact with infected people	Binary	53,490 (70.08)		13,514 (50.15)		39,976 (80.96)	
Current pregnancy	Binary	1,216 (1.59)		548 (2.48)		668 (1.11)	
Symptoms:							
Cough	Binary	39,239 (51.41)		10,317 (38.29)		28,922 (58.57)	
Fever[1]	Binary	30,912 (40.5)		21,878 (44.30)		9,034 (33.53)	
Convulsion	Binary	240 (0.31)		160 (0.59)		80 (0.16)	
Respiratory Distress	Binary	36,462 (47.77)		12,619 (46.83)		23,843 (48.29)	
Muscular pain	Binary	27,730 (36.33)		6,634 (24.62)		21,096 (42.72)	
Reduction or loss of smell	Binary	1,219 (1.6)		339 (1.26)		880 (1.78)	
Reduction or loss of taste	Binary	820 (1.07)		206 (0.77)		614 (1.24)	
Stomachache	Binary	2,316 (3.03)		1,312 (4.87)		1,004 (2.03)	
Anorexia	Binary	5,163 (6.76)		1,634 (6.06)		,3529 (7.15)	
Diarrhea	Binary	2,686 (3.52)		1,452 (5.40)		1,234 (2.50)	
Nausea	Binary	4,969 (6.51)		2,078 (7.71)		2,891 (2.85)	
Vomit	Binary	2,898 (3.8)		1,255 (4.66)		1,643 (3.33)	
Vertigo	Binary	1,854 (2.43)		522 (1.94)		1,332 (2.70)	
Headache	Binary	8,009 (10.49)		2,583 (9.56)		5,471 (11.08)	
Paresis or Paralysis	Binary	275 (0.36)		140 (0.52)		135 (0.27)	
Inflammation of the skin lesion	Binary	76 (0.1)		41 (0.15)		35 (0.07)	
Loss of consciousness	Binary	2,495 (3.27)		1,455 (5.4)		1,040 (2.10)	
Semi-clinical data:							
PO_2>93%	Binary	36118 (47.32)		14,444(53.60)		21,674 (43.90)	
Past/current disease:							
Asthma	Binary	1218 (1.6)		674 (2.50)		544 (1.10)	
Diabetes	Binary	13555 (17.76)		6082 (22.57)		7473 (15.13)	
Dialysis	Binary	925 (1.21)		642 (2.38)		283 (0.57)	
Hypertension	Binary	679 (0.89)		423 (1.57)		256 (0.52)	
Cancer	binary	2671 (3.5)		1800 (6.68)		871 (1.76)	
HIV/AIDS	Binary	46 (0.06)		21 (0.08)		25 (0.05)	
Cardiovascular	Binary	12722 (16.67)		7140 (26.50)		5582 (11.30)	
Liver disease	Binary	706 (0.93)		445 (1.65)		261 (0.53)	
Blood disease	Binary	822 (1.08)		496 (1.84)		326 (0.66)	
Kidney disorders	Binary	2715 (3.56)		1788 (6.63)		927 (1.88)	
Neurological disorders	Binary	1391 (1.82)		791 (2.94)		600 (1.21)	
Acquired/Congenital immunodeficiencies	Binary	656 (0.86)		407 (1.51)		249 (0.50)	
Other chronic diseases	Binary	3140 (4.11)		1692 (6.30)		1448 (2.94)	
Other respiratory diseases (except Asthma)	Binary	1423 (1.86)		795 (2.95)		628 (1.27)	

[1] body temperature ≥ 37.8°C

First, records with missing IDs (11 records) were removed because the attribute values were not appropriate to their corresponding attribute type, which may be due to inappropriate data extraction from the original database.

To avoid any bias, nationals of other countries who did not have the Iranian National Code (9,555 records) were excluded from the study. In the next step, due to the appropriate size of dataset size, all records with missing symptoms (8,123), PO2 (1,400), and Sex (123) were removed from the main dataset. Since our research scope is limited to detecting symptomatic COVID-19 patients, all records that did not have any symptoms were ignored. The criterion for people infected with COVID-19 was positive RT-PCR test results. All the records without a PCR test or unknown result maybe for data collection problems (29,672) or the result of the test was not yet known (1,678) are excluded from the main dataset.

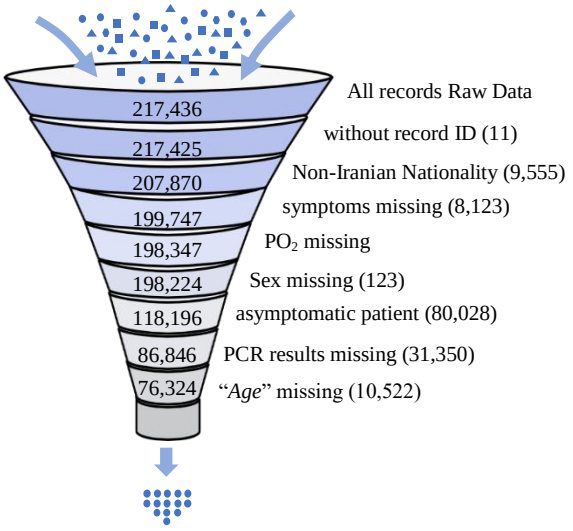

Figure 2. The process of raw dataset cleaning .

Unfortunately, approximately 10,522 records did not have age values, and we ignored them. Finally, we generated our final dataset containing the following features in four main categories: 1) Basic information: *sex, age, smoker, drug addiction, contact with infected people, current pregnancy* 2) Symptoms: *cough, fever, convulsion respiratory distress, muscular pain, reduction or loss of smell, reduction or loss of taste, stomachache, anorexia, diarrhea nausea, vomit, vertigo, headache, paralysis, inflammation of the skin lesion, loss of consciousness* 3) Semi-clinical data: $PO_2 > 93\%$, *and* 4) Concurrent disease: *asthma, diabetes, dialysis, hypertension, cancer, HIV/AIDS, cardiovascular, liver disease, blood disease, kidney disorders, neurological disorders, acquired/congenital immunodeficiencies,* and *other chronic or respiratory diseases (except asthma).*

2.3. Classifiers selection

Intelligent systems such as machine learning and deep learning have been widely used in various applications such as healthcare, especially in the fight against the worldwide COVID-19 pandemic. A recent study [47] provided an analysis of machine learning usage levels as well as deep learning techniques to address the COVID-19 challenge. The results show 31% text data input *vs.* 69% image and other type of data input in machine-learning research. Since our input data is a type of text, among machine learning techniques, we chose a top and recent ML algorithms to compare our COVID-19 prediction model and compare the results. Table 3 shows the list of selected state-of-the-art machine learning methods along with their main ideas. These methods are linear discriminant analysis, quadratic discriminant analysis, logistic regression, naive bayes, support vector classifier, decision tree, random forest, K-nearest-neighbor, extra tree classifier, adaptive boosting, extreme gradient boosting.

3. Experimental Results

This section covers the last four steps in Figure 1. Machine learning models are not intelligent enough to determine the hyperparameters that would lead to the highest possible accuracy on the given dataset. However, hyperparameter values when set correctly can build highly accurate models, allowing our models to try different combinations of hyperparameters during the training process and make predictions with the best combination of hyper-parameter values. Grid search is the most widely used strategy for hyperparameter optimization [48]. Regardless of the optimization method used, the hyperparameter optimization task is generally very expensive in terms of computational costs. The optimization process requires the creation of a search space. Geometrically, this can be compared with an n-dimensional volume, where each hyper-parameter stands for a distinct dimension and the scale of the dimension is represented by the possible values of the hyper-parameter such as real-valued, integer-valued or categorical. For each value of each hyperparameter, a point in the search space is a vector with a distinct value. Finding a vector that gives the model after learning the best accuracy is the aim of the optimization process. With the help of Python, we tuned the classifiers' hyperparameters, and the results are shown in

Table 4. For example, the AdaBoost classifier has two hyperparameters: the "learning_rate" is the weight assigned to each classifier at each iteration of boosting, and "n_estimator" is the maximum number of estimators at which boosting is terminated.

Table 3. Machine learning algorithms and their main ideas.

ML algorithm	Main idea
LDA	Finds a linear hyperplane that separates the classes and maximizes the between-class to within-class variance ratio
QDA	A variation of LDA that allows each class to have its own covariance matrix and fits a quadratic hyperplane to separate the classes.
NB	A probabilistic method that applies Bayes' theorem and assumes conditional independence among the features.
LR	A method that models the probability of an outcome using a logistic function and estimates the parameters by maximizing the likelihood function.
SVM	Finds an optimal hyperplane that maximizes the margin between the classes and uses kernel functions to map the data to a higher-dimensional space.
DT	A method that splits the data into subsets based on feature values and creates a tree-like structure of rules to classify the data.
RF	Builds multiple decision trees using bootstrap samples of the data and random subsets of the features and aggregates their predictions by majority voting or averaging
ETC	An ensemble method similar to RF but uses the entire dataset to train each decision tree and splits the nodes randomly rather than optimally.
KNN	A non-parametric method that assigns a label to a new instance based on the labels of its k closest neighbors in the feature space.
AdaBoost	An ensemble method that iteratively trains weak learners, such as decision trees, and assigns higher weights to misclassified instances to improve their performance.
XGBoost	An ensemble method that uses gradient boosting to train decision trees and optimizes a loss function using gradient descent.

After carrying out the Grid Search, their hyperparameter values are 70 and 0.9, respectively. Box-plot diagrams are presented using cross-validation accuracy to assess statistical significance. Finally, utilizing SHAP and SHAP analysis, key features for tree-based models have been determined.

Now, the classification algorithms could be applied to the cleaned dataset with hyperparameters tuned. After constructing, training, and testing the models, the popular boosting algorithm *AdaBoost* provides the best binary classification performance. However, its performance is comparable with that of XGBoost, which has a large performance difference; the *SVC* classifier shows poor performance among the different classifiers, as shown in Table 4. To demonstrate the results, we constructed an ROC curve (Figure 3). The performance criteria of the classifiers are listed in Table 5. It appears that XGBoost obtained the highest Kappa, MCC, and AUC values of 37.96%, 39.15%, and 67.78%, respectively.

4. Model Validation

A methodological fault is learning a prediction function's parameters and then testing them with the same data. A model with a perfect score that simply repeated the labels of the samples it had just examined would be unable to make any useful predictions regarding data that had not yet been observed. Overfitting is a term used in this context. It is a standard procedure to hold out a portion of the available data as a test set while conducting a (supervised) machine learning experiment to avoid this mistake. This is called K-fold cross-validation (CV), and is used as a technique to test the effectiveness of machine learning models. In this study, we set K = 10, we divided the dataset into 10 folds and used one-fold for testing and the remaining nine folds for training in each iteration. We repeated this process 10 times, using a different fold for testing each time. Table 6 displays the cross-validation results of the classifiers, with the classification results for each fold provided for all the classifiers. The last row provides the average of the accuracy performance metrics. From Table 6, it can be seen that SVC yielded the lowest score, whereas the AdaBoost and XGBoost have almost achieved the same average accuracy, with a mean of 73.31% and 73.30%, respectively. The results in Table 6 are also depicted in the ROC curve (Figure 3) and a box plot (Figure 4).

5. Statistical Test Results

To compare the cross-validation performance results of the machine learning algorithms on the same dataset, statistical significance tests can assist in dealing with the challenge of choosing the best machine learning method. To use the ANOVA test, the normality assumption should be checked before its application. The Anderson-Darling normality test results on shuffle 10-fold cross-validation indicate that the p-value is less than 0.05 ($\alpha = 0.05$) for all; therefore, the null hypothesis of normality, is not rejected. Because the normality assumption is not violated, the ANOVA test is applied at two levels. At the first level, the hypothesis that there is no difference between the means of the algorithms is rejected. At the second level, multiple comparisons with Bonferroni correction method are used to show the superiority of AdaBoost and XGBoost over the other methods. Tables 7 and 8 show the results of the Anderson-Darling normality test and ANOVA results.

6. Discussion

Machine learning models are frequently "*black boxes*," which makes it challenging to analyze them. We require explainable machine learning algorithms that reveal some of these qualities in order to identify the main characteristics that affect the output of the model.

Table 4. Classifier's hyperparameter values based on Grid search.

Classifiers	# Hyperparameters	Hyperparameters name	Hyperparameters value(s)
LDA	1	Solver	{'svd', 'lsqr', 'eigen}
QDA	1	param_reg _param	0.087
LR	2	Penalty, Solver	None, lbfgs
NB	1	Alpha	0.7
DT	3	Criterion, max_depth, max_features	Entropy,12,8
RF	4	Criterion, max_depth, max_features, n_estimator	Gini, 12, 10
ETC	2	Max Depth, Max Features	12, 10
KNN	1	Number of neighborhoods	19
XGBoost	5	N_estimator, Learning_rate, max_depth, colsample_by_tree	500, 0.6, 12, 0.7
RBF-SVM	2	Cost (C), Gamma (γ)	0.002724, 57.51
AdaBoost	2	n_estimator, Learning rate	70, 0.9

Table 5. Classification performance (%)

Classifiers	ACC	MSE	F1_Score	Kappa	MCC	Precision	SE	SP	AUC	Jaccard score
AdaBoost	73.31	26.69	56.14	37.71	39.00	67.70	48.32	86.95	67.63	39.40
XGBoost	73.30	26.70	56.56	37.96	39.15	67.53	49.00	86.56	67.78	39.78
LR	72.67	27.33	53.87	35.54	37.23	67.67	45.28	87.62	66.45	37.20
LDA	72.55	27.45	54.17	35.54	37.04	66.94	46.00	87.04	66.52	37.48
DT	72.35	27.65	55.12	35.82	36.94	65.66	47.88	85.71	66.79	38.40
RF	72.30	27.55	55.11	35.45	36.62	65.64	47.15	85.05	66.71	38.42
ETC	72.25	27.75	54.55	35.34	36.59	65.82	47.02	86.01	66.52	37.81
NB	71.67	28.33	57.06	36.19	36.69	62.35	53.27	81.71	67.49	40.30
KNN	71.66	28.34	49.53	31.63	33.90	67.23	39.34	89.30	64.32	33.23
QDA	71.30	28.70	42.28	27.46	32.60	73.86	29.83	93.93	61.88	27.00
SVC	64.70	35.30	0.00	0.00	0.00	0.00	0.00	100.00	50.00	0.00

Table 6. Accuracy score of classifiers using 10-fold cross-validation.

K	AdaBoost	XGBoost	LR	LDA	DT	RF	ETC	NB	KNN	QDA	SVC
1	67.18	65.77	66.78	66.71	64.21	71.99	64.98	64.48	65.36	66.91	64.69
2	65.74	65.49	65.79	65.60	64.46	71.19	65.45	63.89	64.56	67.08	64.69
3	66.38	67.10	66.93	66.54	66.62	70.49	66.68	64.30	66.29	67.56	64.69
4	71.87	71.95	71.30	71.10	70.50	73.18	69.76	70.01	70.63	69.84	64.69
5	77.37	77.53	76.87	76.51	76.68	69.49	76.22	76.19	74.99	73.91	64.69
6	79.95	79.98	79.59	79.69	78.87	70.88	78.51	79.38	77.21	76.53	64.70
7	75.43	75.48	74.78	74.80	74.91	75.77	74.75	75.08	73.89	72.20	64.70
8	78.64	78.85	77.80	77.84	78.35	74.60	77.92	78.35	75.92	74.66	64.70
9	74.10	74.44	72.58	72.33	73.09	72.79	72.38	71.42	73.03	72.20	64.70
10	76.42	76.43	74.33	74.36	75.84	73.15	75.80	73.61	74.74	72.14	64.70
AVG	73.31	73.30	72.67	72.55	72.35	72.30	72.25	71.67	71.66	71.30	64.70

Table 7. The Anderson-Darling normality test results.

	AdaBoost	XGBoost	LR	LDA	DT	RF	ETC	NB	KNN	QDA
Test statistic	0.41	0.45	0.30	0.30	0.42	0.17	0.41	0.37	0.55	0.35
Critical value	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.68
Reject the null hypothesis	No	No	No	No	No	No	No	No	No	No

Table 8. Multiple comparisons. (ANOVA)

Pair	P-value
AdaBoost – XGBoost	0.108
AdaBoost – each other	<0.05
XGBoost – each other	<0.05

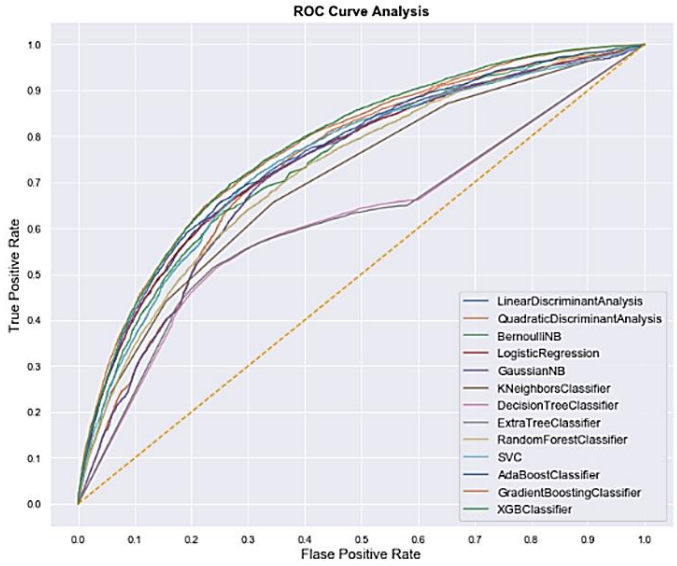

Figure 3. ROC curve for COVID-19 classification.

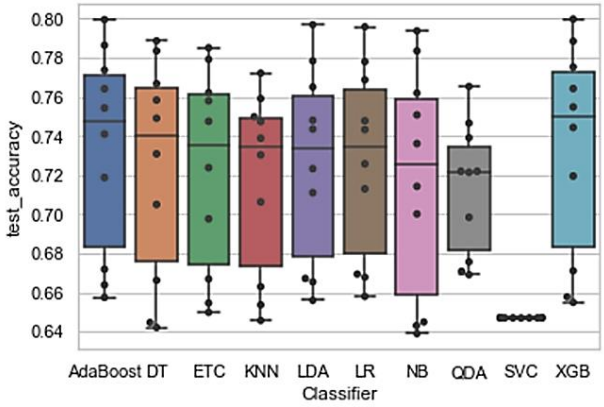

Figure 4. Box-plot for COVID dataset

The SHAP method is one of these techniques; it is used to describe how each feature impacts the model and allows local and global analysis for the dataset and problem at hand. Essentially, SHAP can display the local feature contribution for each instance of the problem using the scatterplot of the Beeswarm plot (Figure 5) and the global feature contribution using the feature importance (Figure 6). The absolute SHAP value in Figure 5 shows that the top 20 factors influence the model more. Variables are displayed in descending order of importance for each global feature, with the first variable being the most important, and the final one being the least important. By way of example, "*contact with infected people*", "*cough*", "*muscle pain*", "*age*", "*cardiovascular commodities*", "*fever*", "PO_2*", and* "*respiratory distress*" are the most important features. It should be noted that while SHAP values demonstrates the value or contribution of each feature to the model's

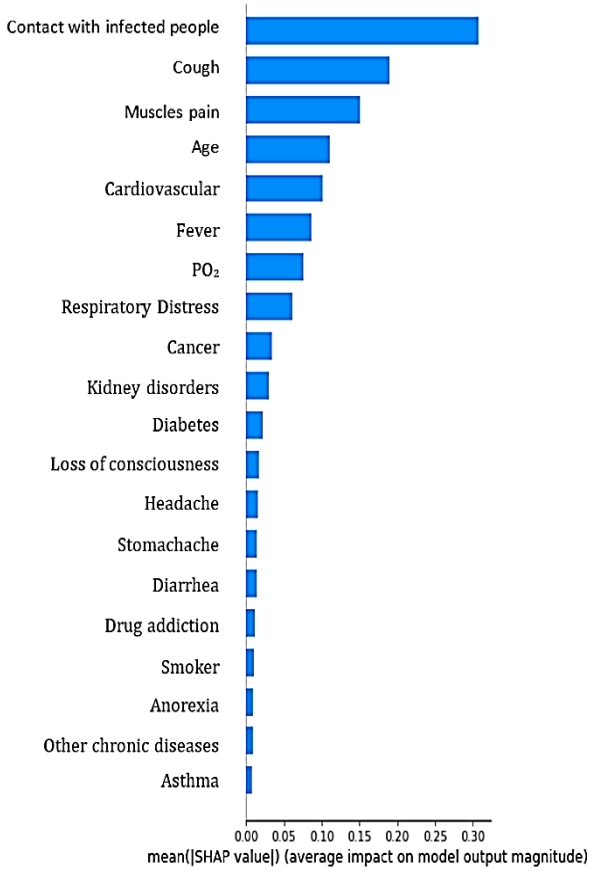

Figure 5. Graph of XGBoost SHAP feature importance

prediction, it does not assess the accuracy of the prediction itself.

However, SHAP values only describe the model behavior that is built from data, and it does not mean a causality relationship between features and targe variables [49]. As we know a prediction model can have false positives and false negatives errors, so the SHAP value can elucidate the results, and a summary plot will provide useful interpretation. According to Figure 6, we can conclude the direction of the feature impact on the target variable. As we can see, "*contact with infected people*", "*cough*", "*muscle pain*", "*cardiovascular*" commodities, and "*fever*" have a relatively large positive effects on the target variable. The red color represents the "*high*", while the X-axis displays the "*positive*" influence, whereas we conclude by mentioning that the features "*age*", "PO_2" ($\leq$ 93%), and "*respiratory distress*" are all strongly inversely related with the target variable. Also, variables from "*cancer*" to "*asthma*" have a small global contribution to the target variable, almost all of them except

"*headache*" are positively correlated to the target variables.

To distinguish between individuals who have COVID-19 infection and those who do not, we created a decision tree model that is clinically applicable, clear, and simple to understand. A DT as a simple classifier is a straightforward and user-friendly tool to understand the underlying process that is constructed in two stages, and the models that are produced can be visualized as binary trees. First, we investigate each variable to see how to most effectively divide the data into two groups. We choose important features including x1 = contact with infected people", x_2 = age, x_3 = muscle pain, x_4 = cough, x_5 = fever, x_6 = PO_2, x_7 = respiratory distress, and x_8 = headache. Figure 7 represents the corresponding DT.

7. Conclusion

This study proposes a simple and affordable model to detect COVID-19. The main feature of this model is its ability to quickly identify COVID-19 and assist physicians in treating COVID patients by guiding them to take appropriate safety precautions. Because the collection of self-reported symptoms can be performed remotely, and to maintain limited testing capacity for suspected cases, the ability to predict positive infections based on self-reported symptoms is important to reduce the need for labs that provide samples only for screening. The dataset consists of four main categories of features that can be easily accessed at home. In this research work, a Grid search-based machine learning model is used to identify COVID-19 with the in-home dataset. Several state-

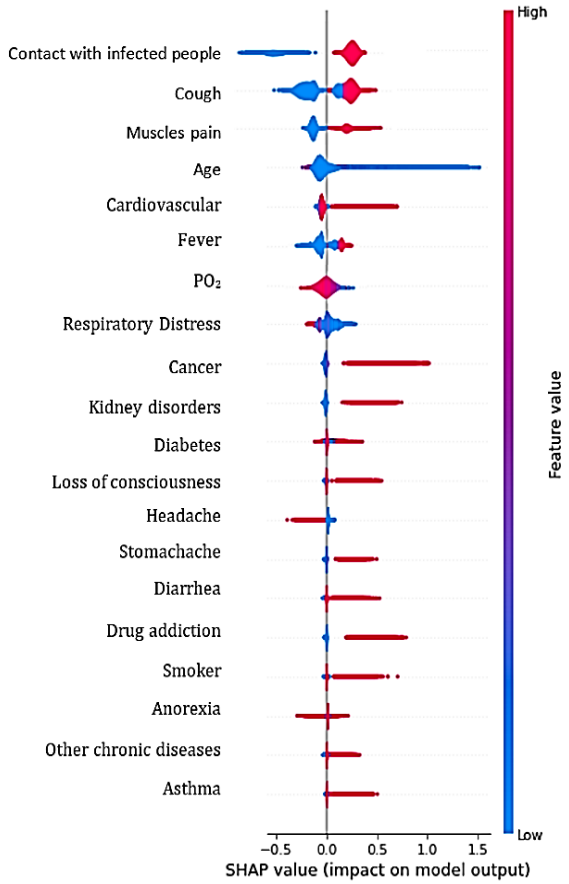

Figure 6. Graph of XGBoost SHAP variable importance for training data.

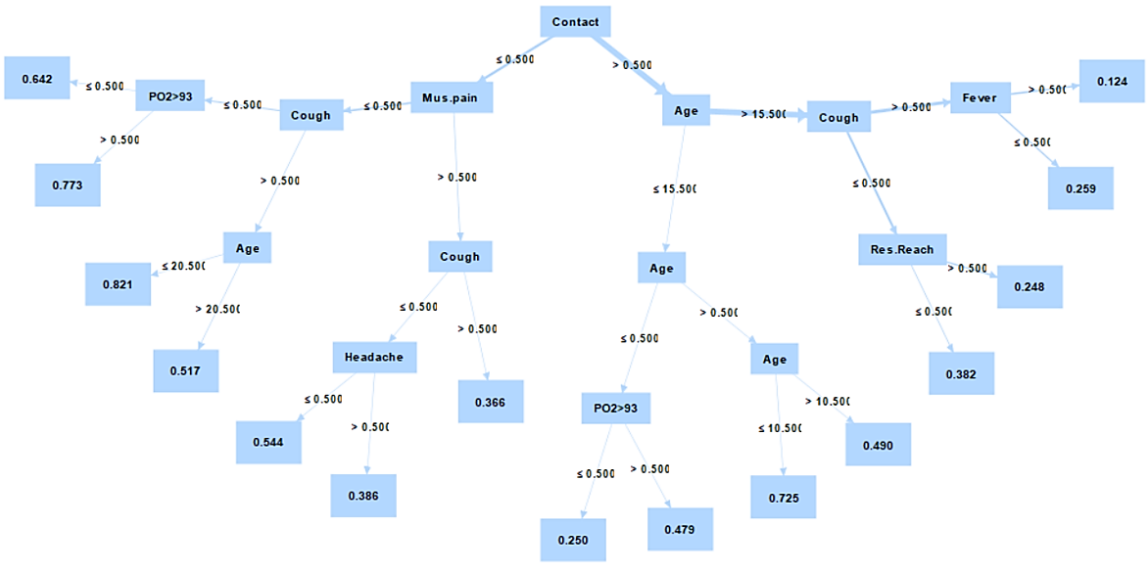

Figure 7. A decision rule for detecting COVID-19 with their thresholds in absolute value.

of-the-art classifiers including LDA, QDA, DT, RF, LR, NB, KNN, XGBoost, AdaBoost, ETC, and SVC are utilized to predict the COVID-19 patients. The models are validated using 10-cross-validation accuracy. Various classification measures including MSE, accuracy, Kappa index, specificity, sensitivity, Matthew's correlation coefficient, and Jaccard score are utilized to describe the classification performance from a different perspective. With interpretability becoming an increasingly important requirement for machine learning projects, we use SHAP value, the most powerful method for explaining how machine learning models make predictions. AdaBoost achieved the best classification performance (73.31%) in terms of accuracy. XGBoost with a slight difference has a high classification performance (72.30) with regards to MCC and Kappa. However, AdaBoost's cross-validation (10-fold) accuracy offers the greatest value. Integration of more clinical data such as blood samples and X-rays or/and CT-scan will improve the classifiers' accuracy. It is necessary to consider that our criteria for confirming COVID-19 infection are RT-PCR that in turn has false positive and negative results. That is maybe one of the reasons that the classification performance may not be satisfactory at the first glance. If we help other diagnostic tools including CT-scan or blood tests, the classification permeance will improve. Finally, a potential application of this research result could be able to integrate it into mobile devices would be extremely helpful to achieve all of our research goals.

Note

This research work was endorsed by the Research Ethics Committees of Tarbiat Modares University, Tehran, Iran.

References

[1] WHO. "WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020," Mar. 11, 2020. [Online]. Available: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020. [Accessed: June. 1, 2023].

[2] WHO. "WHO Coronavirus (COVID-19) Dashboard. 2022,". [Online]. Available: https://covid19.who.int/. [Accessed: 12:15pm CEST, 21 June 2023].

[3] Hafezi, F. and M. Khodabakhsh, "Coronavirus Incidence Rate Estimation from Social Media Data in Iran." *Journal of AI and Data Mining*, 2023. vol. 11, no. 2, pp. 315-329.

[4] Emanuel, E.J. *et al*., "Fair allocation of scarce medical resources in the time of Covid-19," *Mass Medical Soc.*, Vol. 382, no. 21, pp. 2049-2055, 2020.

[5] Fauci, Anthony S., H. Clifford Lane, and Robert R. Redfield. "Covid-19—navigating the uncharted." *New England Journal of Medicine*, vol. 382, no.13, pp. 1268-1269, 2020.

[6] Hong, K.H. *et al*., "Guidelines for laboratory diagnosis of coronavirus disease 2019 (COVID-19) in Korea," *Annals of laboratory medicine*, vol. 40, no. 5, pp. 351-360, 2020.

[7] Bai, H.X. *et al*., "Performance of radiologists in differentiating COVID-19 from non-COVID-19 viral pneumonia at chest CT," *Radiology*, vol. 296, no. 2, pp. E46-E54, 2020.

[8] Rajaraman, Sivaramakrishnan, and Sameer Antani. "Weakly Labeled Data Augmentation for Deep Learning: A Study on COVID-19 Detection in Chest X-Rays," *Diagnostics*, vol. 10, no. 6, p. 358, 2020.

[9] Cabitza, F. *et al*., "Development, evaluation, and validation of machine learning models for COVID-19 detection based on routine blood tests," *Clinical Chemistry and Laboratory Medicine (CCLM)*, vol. 59, no. 2, pp. 421-431, 2021.

[10] Yan, Y., L. Chang, and L., "Wang, Laboratory testing of SARS-CoV, MERS-CoV, and SARS-CoV-2 (2019-nCoV): Current status, challenges, and countermeasures," *Reviews in medical virology*, vol. 30, no. 3, pp. 1052-9276, 2020.

[11] Hellewell, J. *et al*., "Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts," *The Lancet Global Health*, vol. 8, no. 4, pp. e488-e496, 2020.

[12] Lurie, N. *et al*., "Developing Covid-19 vaccines at pandemic speed," *New England Journal of Medicine*, vol. 382, no. 21, pp. 1969-1973, 2020.

[13] Alimadadi, A. *et al*., "Artificial intelligence and machine learning to fight COVID-19," *American Physiological Society Bethesda, MD.* vol. 52, no. 4, pp. 200-202, 2020.

[14] Vaishya, R. *et al*., "Artificial Intelligence (AI) applications for COVID-19 pandemic," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 14, no. 4, pp. 337-339, 2020.

[15] Chang, M.C. and D. Park, "How can blockchain help people in the event of pandemics such as the COVID-19?," *Journal of medical systems*, vol. 44, no.5, pp. 1-2, 2020.

[16] Nasajpour, M., Pouriyeh, S., Parizi, R.M. *et al*. "Internet of Things for Current COVID-19 and Future Pandemics: An Exploratory Study," *J Healthc Inform Res*, vol. 4, no. 1, pp. 325–364, 2020.

[17] Dargan, S., Kumar, M., Ayyagari, M.R. *et al*. "A Survey of Deep Learning and Its Applications: A New

Paradigm to Machine Learning," *Arch Computat Methods Eng*, vol. 27, no. 4, pp. 1071–1092, 2020.

[18] O. Rajabi Shishvan, D. -S. Zois, and T. Soyata, "Machine Intelligence in Healthcare and Medical Cyber Physical Systems: A Survey," in *IEEE Access*, vol. 6, pp. 46419-46494, 2018.

[19] Collins, G.S. and K.G. Moons, "Reporting of artificial intelligence prediction models." *The Lancet*, vol. 393, no. 10181, pp. 1577-1579, 2019.

[20] Gozes, O. *et al*., "Rapid ai development cycle for the coronavirus (covid-19) pandemic: Initial results for automated detection & patient monitoring using deep learning ct image analysis." arXiv preprint arXiv:2003.05037, 2020.

[21] Jin, C., Chen, W., Cao, Y. *et al*. "Development and evaluation of an artificial intelligence system for COVID-19 diagnosis," *Nat Commun*, vol. 11, no. 5088, 2020.

[22] Punn, N.S. and S. Agarwal, "Automated diagnosis of COVID-19 with limited posteroanterior chest X-ray images using fine-tuned deep neural networks". *Applied Intelligence*, vol. 51, no. 5, pp. 2689-2702, 2021.

[23] Song, Y. *et al*., "Deep learning enables accurate diagnosis of novel coronavirus (COVID-19) with CT images". *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 18, no. 6, pp. 2775-2780, Nov.-Dec. 2021.

[24] Wang, S., Kang, B., Ma, J. *et al*. "A deep learning algorithm using CT images to screen for Corona virus disease (COVID-19)," *Eur Radiol*, vol. 31, pp. 6096–6104, Aug. 2021.

[25] Tostmann, A. *et al*., "Strong associations and moderate predictive value of early symptoms for SARS-CoV-2 test positivity among healthcare workers", the Netherlands, March 2020. *Eurosurveillance*, 2020., vol. 25, no. 16, p. 2000508, Apr. 2020.

[26] Punn N, Sonbhadra S, and Agarwal S. "COVID-19 Epidemic Analysis using Machine Learning and Deep Learning Algorithms," medRxiv, Preprint posted online on June 01, 2020.

[27] Feng, C. *et al*., "A novel artificial intelligence-assisted triage tool to aid in the diagnosis of suspected COVID-19 pneumonia cases in fever clinics," *Ann Transl Med*, vol. 9, no. 3, p. 201, 2021.

[28] Mei, X., Lee, HC., Diao, Ky. *et al*. "Artificial intelligence–enabled rapid diagnosis of patients with COVID-19," *Nat Med*, vol. 26, no. 1, pp. 1224–1228, 2020.

[29] Shaverdian, N. *et al*., "Need for caution in the diagnosis of radiation pneumonitis during the covid-19 pandemic," *Advances in radiation oncology*, vol. 5, no. 4, pp. 617-620, 2020.

[30] Malik, M. *et al*., "Determination of COVID-19 Patients using Machine Learning Algorithms," *Intelligent Automation & Soft Computing*, vol. 31, no.1, 202.

[31] Jamshidi, E. *et al*., "Symptom prediction and mortality risk calculation for COVID-19 using machine learning," *Frontiers in artificial intelligence*, vol. 4, no. 1, p. 72, 2021.

[32] Zoabi, Y., S. Deri-Rozov, and N. Shomron, "Machine learning-based prediction of COVID-19 diagnosis based on symptoms," *npj digital medicine*, vol. 4, no.1, p. 3, 2021.

[33] Antoñanzas, J.M. *et al*., "Symptom-based Predictive Model of COVID-19 Disease in Children," *Viruses*, vol. 14, no.1, p. 63, 2022.

[34] Rajput, D., W.-J. Wang, and C.-C. Chen, "Evaluation of a decided sample size in machine learning applications," *BMC bioinformatics*, vol. 24, no.1, p. 48, 2023.

[35] Ioannidis, J.P., "Why most discovered true associations are inflated," *Epidemiology*, vol. 19, no. 5, p. 640-648, 2008.

[36] Carp, J., "The secret lives of experiments: methods reporting in the fMRI literature" *Neuroimage*, vol. 63, no.1, pp. 289-300, 2012.

[37] Ingre, M., "Why small low-powered studies are worse than large high-powered studies and how to protect against "trivial" findings in research: comment on Friston (2012)," *Neuroimage*, vol. 81, no. 1, pp. 496-498., 2013.

[38] Guhathakurata, S. *et al*., "A novel approach to predict COVID-19 using support vector machine, in Data Science for COVID-19," *Data Science for COVID-19. Academic Press*, 2021, ch. 18, pp. 351-364.

[39] Shamsi, A. *et al*., "Contribution of Iran in COVID-19 studies: a bibliometrics analysis," *Journal of Diabetes & Metabolic Disorders*, vol. 19, no. 2, pp. 1845-1854, 2020.

[40] K. Rezaee, A. Badiei, and S. Meshgini, "A hybrid deep transfer learning-based approach for COVID-19 classification in chest X-ray images," *2020 27th National and 5th International Iranian Conference on Biomedical Engineering (ICBME), Tehran, Iran*, 2020 pp. 234-241.

[41] Heydari, M.H. *et al*. "Clustering of Infected Patients by COVID-19 using Self-Organized Mapping and Extracting the Most Important Clinical Features," *2020 6th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS), Mashhad, Iran*, 2020, pp. 1-6.

[42] Sobhani, S. *et al*., "Association between clinical characteristics and laboratory findings with outcome of hospitalized COVID-19 patients: a report from Northeast Iran," *Interdisciplinary perspectives on infectious diseases, 2021*, 2021.

[43] Guhathakurata, S. *et al*., "A new approach to predict COVID-19 using artificial neural networks," *in*

Cyber-physical systems, Elsevier, 2022, Ch. 8, pp. 139-160.

[44] YALÇIN, N. and S. ÜNALDI, "Symptom-based COVID-19 Prediction using Machine Learning and Deep Learning Algorithms," *Journal of Emerging Computer Technologies*, vol. 2, no.1, pp. 22-29, 2022.

[45] Villavicencio, C.N. *et al.*, "COVID-19 Prediction applying supervised machine learning algorithms with comparative analysis using WEKA," *Algorithms*, vol. 14, no.7, p. 201, 2021.

[46] Wirth, R. and J. Hipp. "CRISP-DM: "Towards a standard process model for data mining," *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*," vol. 1, Manchester, Apr. 2000, pp. 29-39.

[47] Nayak, J., Naik, B., Dinesh, P. *et al*. "Intelligent system for COVID-19 prognosis: a state-of-the-art survey," *Appl Intell*, vol. 51, no.5, pp. 2908–2938, 2021.

[48] Bartz-Beielstein, T. and M. Zaefferer, "Hyperparameter Tuning Approaches, in Hyperparameter Tuning for Machine and Deep Learning with R: A Practical Guide," *Springer Nature Singapore*, Singapore, ch. 4, pp. 71-119, 2023

[49] Baptista, M.L., K. Goebel, and E.M. Henriques, "Relation between prognostics predictor evaluation metrics and local interpretability SHAP values," *Artificial Intelligence*, vol. 306, no.1, pp. 103667, 2022.

پیش‌بینی سریع ابتلای کووید-۱۹ با داده‌های خانگی و الگوریتم‌های طبقه‌بندی یادگیری ماشین: مطالعه موردی ایران

علی شبرندی[1*]، علی رجب زاده قطرمی[1] ، نادر توکلی[2] ، محمد دهقان نیری[1] و سحر میرزائی [3]

[1] گروه مدیریت صنعتی، دانشکده مدیریت و اقتصاد، دانشگاه تربیت مدرس، تهران، ایران

[2] گروه طب اورژانس، مرکز تحقیقات تروما، دانشگاه علوم پزشکی ایران، تهران، ایران.

[3] گروه بهداشت و محیط زیست، دانشگاه علوم پزشکی ایران، تهران، ایران.

ارسال ۲۰۲۳/۰۷/۲۷؛ بازنگری ۲۰۲۳/۰۹/۱۸؛ پذیرش ۲۰۲۳/۱۰/۲۸

چکیده:

برای کاهش بار شدید مبتلایان به کووید-۱۹ به سیستم های بهدا شتی و درمانی، طرح غربالگری سریع و کارآمد در خط مقدم مبارزه با این بیماری مورد نیاز است. بسیاری از تحقیقات گذشته از نتایج آزمایشگاهی، سی‌تی اسکن و اشعه ایکس استفاده برای این موضوع استفاده نموده اند که مانعی جدی برای غربالگری چابک است. در این مطالعه، یک مدل تشخیص کووید-۱۹ کاربرپسند و کم‌هزینه را بر اساس داده‌های خانگی در قالب سه دسته داده، جمعیت شناختی، علائم و سوابق بیماری ارائه شده است. در این مطالعه از روش جستجوی گرید برای شناسایی ترکیب بهینه هایپرپارامترهایی که دقیق‌ترین پیش‌بینی را ارائه می‌دهد، استفاده شده است و عملکرد ۱۱ الگوریتم طبقه بندی یادگیری ماشین مقایسه شده است. نتایج نشان می‌دهد که الگوریتم XGBoost بالاترین صحت، ۷۳,۳٪ را ارائه می‌کند، اما تحلیل‌های آماری نشان می‌دهد که تفاوت معنی‌داری بین عملکرد دقت XGBoost و AdaBoost وجود ندارد، اگرچه برتری این دو روش را نسبت به سایر روش‌ها اثبات کرد. علاوه بر این، مهمترین ویژگی های به دست آمده با استفاده از SHapely Adaptive explanations مورد تجزیه و تحلیل قرار گرفت. «تماس با افراد آلوده»، «سرفه»، «درد عضلانی»، «تب»، «سن»، «مشکلات قلبی عروقی»، «PO2» و «دیسترس تنفسی» مهم‌ترین متغیرها هستند. در بین این متغیرها، سه متغیر اول تاثیر مثبت نسبتا زیادی بر متغیر هدف دارند. در حالی که «سن»، «PO2» و «دیسترس تنفسی» به شدت با متغیر هدف همبستگی منفی دارند. در نهایت، یک مدل درخت تصمیم قابل اجرا، قابل مشاهده و تفسیر آسان برای پیش‌بینی ابتلای کووید-۱۹ ارائه شده است.

کلمات کلیدی: کووید-۱۹، علائم بیماری، یادگیری ماشین، طبقه بندی، هوش مصنوعی.