Comparing Supervised Machine Learning Models for Covid-19 patient detection using a Combination of Clinical and Laboratory Dataset

Narges Mohebbi¹, Mehdi Tutunchian², Meysam Alavi³, Mehrdad Kargari⁴, Amir Behnam Kharazmi⁵

¹ Department of Information Technology, Tarbiat Modares University, Tehran, Iran; narges.mohebbi@modares.ac.ir, mehditutunchian@modares.ac.ir, meysamalavi@modares.ac.ir, M_kargari@modares.ac.ir
² Department of Internal Medicine, School of Medicine, Shahid Beheshti University of Medical Sciences, Imam Hossein Hospital, Tehran, Iran; Drkharazmy@sbmu.ac.ir

Abstract
COVID-19 is a new variant of SARS-COV-2 which can lead to mild to severe infection in humans. Despite the remarkable efforts to contain the epidemic, the virus spread rapidly around the world and its prevalence continued with different degrees of clinical symptoms in many countries. Although common strategies including prevention, diagnosis, and care are necessary to curb this epidemic, early and accurate diagnosis can play an important role in reducing the speed of the epidemic. In this regard, the use of technologies based on artificial intelligence can be of great help. For this reason, since the outbreak of COVID-19, many researchers have tried to use machine learning techniques as a subset of artificial intelligence for the early diagnosis of COVID-19. Considering the importance and role of using clinical and laboratory data in the diagnosis of people with covid-19, in this paper K-NN, SVM, decision tree, random forest, Naive Bayes, neural network and XGBoost models are the most common machine learning models, and a dataset containing 1354 records consisting of clinical and laboratory data of patients in Imam Hossein Hospital in Tehran has been used to diagnose patients with covid-19. The results of this research indicate that based on the evaluation criteria, XGBoost and K-NN models have the most accuracy among the mentioned models and can be considered suitable predictive models for the diagnosis of COVID-19.

Keywords: COVID-19; Coronavirus; Early Detection; Machine Learning Techniques; Supervised Model.

1. Introduction
The first records of the coronavirus date back to 1930 [1]. But in the last days of 2019, a new type of this virus, which is also highly contagious, was reported from the city of Wuhan, Hubei Province, China, and was named sars-cov-2 [2], [3]. The names of the virus include Covid-19, 2019nCov, and 2019 Novel Coronavirus. It reports the sequence of the bat gene, but the intermediate hosts between the bat and the human are still being studied. The disease is transmitted through the respiratory system of a patient that is inhaled, coughing, or sneezing in the air and is transmitted to a healthy person's respiratory system [4]. Covid-19 has spread rapidly worldwide due to its high transmission rate and contagiousness [5], making it the fifth deadliest virus in history [6], with the World Health Organization declaring the virus to be the sixth leading cause of global health emergency [7]. The virus has a similar impact on the global economy [4], [5], and according to the World Economic Cooperation and Development Organization, the emergence of Covid-19 has put economic conditions at their worst in history [8]. Some of the other impacts of the virus can be attributed to the impact on the education system and the impact on mental and social health [6]. Strategies for managing the disease worldwide include force lockdown, travel restrictions, social distancing, and quarantine [5]. Due to the damage that the virus has done to the world, early detection, improvement of screening, prevention and control of this disease are urgent and are highly encouraged [4-6], [9], [10]. All over the world, diagnosis of the disease is based on the patient’s history and physical examination, making the possibility of a mistake or late diagnosis of the disease inevitable. Once the disease is diagnosed quickly, the chances of successful treatment will increase, and the patient's rehabilitation costs will decrease. Most medical errors are due to a lack of medical tools when making decisions. For example, the RT-PCR test, the golden standard method for diagnosing the disease, is not efficient enough; it’s expensive, needs too much time and is not very sensitive at all [4], [10]. In recent years, after the start of Covid-19, artificial intelligence (AI) methods have become very popular due to features such as high accuracy, sensitivity and evaluation criteria, as well as the use of low resources and high speed in medicine and biology [2], [11]. Nowadays, very useful and popular tools are being made to help the decision-making process more accurately and with the least possible mistakes in the medical field, which are based on artificial intelligence techniques and algorithms, and machine learning has an undeniable role in health care and disease diagnosis [12-16].

The power of machine learning creates intelligent systems whose output can be used as a useful and efficient tool in finding the signs of disease in potential patients which will
lead to reducing the huge burden on the health care system in all countries of the world [15], [17]. Studies show that this case has been the intention and goal of many researchers who are active in this field [17], [18]. Most of these studies have focused on the diagnosis of patients with covid-19 based on CT scan and X-ray images of the lungs, and especially, deep learning algorithms are considered very successful in this field because they are unique in discovering patterns and high-level features from images. [19]. In addition, artificial intelligence techniques are widely used to predict whether people are sick or not based on their medical information. A person's medical information can include things such as symptoms and history of underlying diseases (clinical data), results of various tests such as blood or urine tests (laboratory data) and general information such as gender and age (demographic data) [18]. The diagnosis of covid-19 using clinical and laboratory findings is not an exception to this rule. Accordingly, in this study, a brand new and up-to-date dataset of the mentioned data of patients with covid-19 who were admitted to Imam Hossein Hospital in Tehran was collected and seven models of the most used supervised machine learning models were selected and trained by this dataset. Finally, the performance of these models has been examined and compared in terms of accurate disease diagnosis.

The structure of the continuation of the paper is organized as follows: In the second part, the literature review is briefly reviewed. In the third part, the research method is introduced. In the fourth part, the results are evaluated. The fifth part is related to conclusions and suggestions for future work.

2. Literature Review

As stated in the introduction of the study, several studies have been conducted in this field that has tried to use intelligent systems for the early diagnosis of diseases. In the following, we will categorize and review a selection of these studies. There are basically three types of disease diagnostic models; The first group is models whose basis is working on CT scan images or X-ray images of the patient's lungs. For example, the authors in [20] used a dataset consisting of 2481 CT scan images to train their models. Finally, the CNN-based model was able to register a better performance with 95% precision. Another research [21] has trained a CNN model to diagnose the disease with a dataset consisting of CT scans of 131 people with Covid-19 infection and 150 people without Covid-19 infection (a total of about 10,000 images), and its final precision had reached up to 90%. A study was performed on medical images of the lungs, which in the first stage analyzes the images and extracts important features using machine vision techniques, and in the second stage uses five machine learning models, including k Nearest Neighbor (k-NN), Decision Tree (DT), Multinomial Logistic Regression (MLR), Naive Bayes (NB) and Support Vector Machine (SVM) to diagnose Covid-19. They have reached the accuracy of 0.85, 0.79, 0.91, 0.61, 0.89, respectively [10]. X-ray images are also used to train disease diagnosis models, for example [22] used a set of 392 of these images to train their CNN model, which resulted in 98.4% precision. Also, among the research, one can find examples that have used both CT scan images and X-ray images to train their models at the same time [23-25].

In addition to the studies that used patient lung images to diagnose COVID-19, other studies used patients' clinical information. For example, the study [26] used a database of 1,624 patients (with a 52% positive COVID-19 rate) that included white blood cell count, monocytes (%), lymphocyte (%), and age features to build a stacking machine learning-based system for covid-19 diagnosis. The above system was measured by the parameters of accuracy, precision, recall, and F-Score, which achieved 91.45%, 91.44%, 91.44%, and 91.45%, respectively. In the final section, this system leads to another nomogram-based scoring system which scored 93.34%, 92.02%, 90.54%, and 93.73% for accuracy, precision, recall, and F-Score parameters, respectively. In [9], a dataset was formed based on routine blood values (RBV), which includes 28 features for two groups of patients named severely (192 patients) and mildly (4010 patients), and the writer claims that the effect of all 28 features on the diagnosis of Covid-19 in the patient is medically approved. In the next step, this dataset was analyzed using supervised machine learning algorithms. The algorithms used and their performance based on accuracy is (LWL)-97.86%, K-star (K*)-96.31%, Naive-Bayes (NB)-95.36%, and k-nearest-neighbor (KNN)-94.05%. Moreover, they have reached the AUC of LWL-0.95%, K*-0.91%, NB-0.85%, and KNN-0.75%. Another study [27] used four machine learning models, including RF, XGBoost, SVM, and logistic regression, to detect Covid-19. The data used in this study consists of the results of six clinical tests (CC), 14 blood tests (CBC), and PCR test results as the golden standard that were collected for 1391 patients. Using the mentioned algorithms, eight models have been made using CBC and combining CBC and CC. The RF model on the combined data performed best with 85.3% accuracy, 79.6% sensitivity, and 91.2% specificity, while the accuracy of this model was 82.8% on using only CBC. Also, the SVM model using combined data has the best performance on external data with 91.18% accuracy, 100% sensitivity, and 84.21% specificity. In [28], we also see that machine learning algorithms (NB, NLP, KNN, RF, LR & DT) were used to diagnose Covid-19 using clinical patient datasets. Nevertheless, the different point in this research is the attention paid to optimization of a dataset and recognizing the importance of different features (Feature selection) for detecting COVID-19, with different algorithms (Boruta (shadow features), Ridge Regression (RR) & Random Fourier features (RFF)) and the results of all models are reported for all cases. Although, in general, in all cases, we see a curacy of over 90%, feature selection has improved the performance of different models in terms of classification and processing time. For example, the accuracy of the NB model with the Boruta algorithm is 84%, with the RR algorithm 84%, and the RFF algorithm is 93%. Also, the best performance among all cases is the RF model with the Boruta algorithm with 95% accuracy, 94% precision, 95% recall, and 94 F-Score. In [29], the authors used the blood test results of 279 patients to form their dataset and train their models. Different machine learning models were trained using these data and sensitivity in the range of (92 to 95%) and accuracy in the range of (82 to 86%) were obtained. The authors in [30] have also tried to provide models for disease diagnosis using ML algorithms and the data of 235 patients. Finally, the SVM model showed the best performance among the five models built in this research. Another study [31] has focused on predicting the recovery of covid-19 patients. SVM, KNN, DT and RF models were used for this purpose, and the DT model performed better than the others with 99.85%
precision. The different performance of conventional machine learning models in different researches is also significant and in different conditions, different models have shown better performance. For example, in another study [32], which used the results of blood tests of patients to train the models, the NB model achieved the highest accuracy value and the lowest computational costs among other models. In the literature on the subject, one can find many other examples similar to what was stated [33-36].

To be comprehensive in the present report, we should also mention the studies that used a combination of image and clinical data to train COVID-19 diagnostic models. For example, [37] used a CXR image dataset and a clinical data dataset of Covid-19 patients to build a model that performed better than the image-based and the clinical data-based models. The first two independent models were trained and evaluated, then fusion techniques were used to combine the two models, and an accuracy of 0.970, a recall of 0.986, precision of 0.978, and an F-Score of 0.982 were obtained.

3. Methodology

The methodology of this research consists of four stages, which are data collection, data pre-processing, classifier model design and evaluation Figure 1. In the rest of this section, each stage will be briefly introduced.

3.1 Data collection

The dataset used in this research was collected from Imam Hossein Hospital in Tehran, which includes 1354 patients referred to this hospital for testing for covid-19. It took about 40 working days to collect this dataset and it was obtained from reviewing the files of 1354 patients who were referred to Imam Hossein Hospital between October 2019 and October 2014. In this data set, 754 samples of records with the label of confirmed infection with covid-19 and 600 samples with the label of non-confirmation of infection with covid-19 have been recorded.

3.2 Data Preprocessing

The dataset collected from patient files has 30 features, which are classified into three categories of clinical, laboratory and population data. Feature selection is considered as selecting the appropriate variable from a dataset, in other words, feature selection is a process in which essential information is extracted from a large dataset to reduce processing time with better performance [17]. According to the multiplicity of features and also the existence of a correlation between some features, with the opinion of a specialist doctor, essential features were selected and reduced from 30 basic features to 12 basic features, which include 4 clinical features and 8 laboratory findings. The details of the selected features can be seen in Table 1 and Table 2. In Table 1, row 13 is the record class label, where the numbers 0 and 1 are taken to mean being negative or positive of the Covid-19 disease.

3.3 Building model

The dataset used in this research has specific classes, so 7 of the most used supervised machine learning algorithms in medicine [38], [39] have been used for the early diagnosis of patients with covid-19. The models used are as follows: XGBoost, Neural Networks (NN), Naïve Bayes (NB),

![Figure 1. Research methodology stages](image)

<table>
<thead>
<tr>
<th>No.</th>
<th>Feature</th>
<th>Min</th>
<th>Max</th>
<th>Mean</th>
<th>Std</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Saturation o2</td>
<td>46</td>
<td>100</td>
<td>92.22</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>RR</td>
<td>8</td>
<td>80</td>
<td>16.42</td>
<td>4.01</td>
</tr>
<tr>
<td>3</td>
<td>PR</td>
<td>50</td>
<td>160</td>
<td>79.48</td>
<td>14.45</td>
</tr>
<tr>
<td>4</td>
<td>Clinical symptoms</td>
<td>0</td>
<td>8</td>
<td>2.10</td>
<td>1.58</td>
</tr>
<tr>
<td>5</td>
<td>CRP</td>
<td>0</td>
<td>357</td>
<td>39.9</td>
<td>49.9</td>
</tr>
<tr>
<td>6</td>
<td>CPK</td>
<td>14</td>
<td>5280</td>
<td>207.9</td>
<td>344.2</td>
</tr>
<tr>
<td>7</td>
<td>LDH</td>
<td>218</td>
<td>3094</td>
<td>564.5</td>
<td>319.4</td>
</tr>
<tr>
<td>8</td>
<td>D-Dimer</td>
<td>1</td>
<td>5000</td>
<td>675.9</td>
<td>996.1</td>
</tr>
<tr>
<td>9</td>
<td>WBC</td>
<td>0.8</td>
<td>123.8</td>
<td>7.55</td>
<td>4.78</td>
</tr>
<tr>
<td>10</td>
<td>Lymphocytes</td>
<td>11.62</td>
<td>10275.4</td>
<td>185.4</td>
<td>307.8</td>
</tr>
<tr>
<td>11</td>
<td>ESR</td>
<td>2</td>
<td>1545</td>
<td>357.5</td>
<td>456.3</td>
</tr>
<tr>
<td>12</td>
<td>Ferritin</td>
<td>9</td>
<td>1863</td>
<td>372.8</td>
<td>392.2</td>
</tr>
<tr>
<td>13</td>
<td>Result</td>
<td>0</td>
<td>1</td>
<td>0.55</td>
<td>0.49</td>
</tr>
</tbody>
</table>

![Table 1. Information About The Dataset](image)

<table>
<thead>
<tr>
<th>Type</th>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical Variable</td>
<td>Saturation o2</td>
<td>Percent of hemoglobin that is bound to molecular oxygen</td>
</tr>
<tr>
<td></td>
<td>RR(Respiratory Rate)</td>
<td>Number of person breaths in one minute</td>
</tr>
<tr>
<td></td>
<td>PR(Pulse Rate)</td>
<td>The average of heart beats in one minute</td>
</tr>
<tr>
<td></td>
<td>Clinical symptoms</td>
<td>Patient symptoms and signs, disease manifestation in a patient</td>
</tr>
<tr>
<td>Laboratory Variable</td>
<td>CRP(C-Reactive protein)</td>
<td>It is a substance that produces in the liver and rises in response to inflammation</td>
</tr>
<tr>
<td></td>
<td>CPK(Creatine phosphokinase)</td>
<td>It is an enzyme that is found mainly in the heart, brain and skeletal muscle and can rise in response to viral infections.</td>
</tr>
<tr>
<td></td>
<td>LDH(lactate dehydrogenase)</td>
<td>It is an enzyme that is found in all live cells and can rise in response to viral infections.</td>
</tr>
<tr>
<td></td>
<td>D-Dimer</td>
<td>It is a protein fragment that’s made when a blood clot dissolves in the body and rises in response to inflammation.</td>
</tr>
<tr>
<td></td>
<td>WBC(White Blood Cell)</td>
<td>It is the cells of the immune system and involve in protecting the body against infectious disease.</td>
</tr>
<tr>
<td></td>
<td>Lymphocytes</td>
<td>It is a type of white blood cell in the immune system</td>
</tr>
<tr>
<td></td>
<td>ESR(Erythrocyte Sedimentation Rate)</td>
<td>It is a blood test that measures the rate of settling red blood cells at the bottom of a test tube and rises in response to inflammation.</td>
</tr>
<tr>
<td></td>
<td>Ferritin</td>
<td>It is a blood protein that stores iron and rises in response to inflammation</td>
</tr>
</tbody>
</table>

![Table 2. Description of Data Set Features](image)
Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM) and K-Nearest Neighbor (KNN).

3.4 Model evaluation

In this paper, the K-Fold Cross Validation method has been used to evaluate the results of the mentioned models. In this research, the results are evaluated based on k=5, k=10

4. Comparison And Evaluation

A system with specifications of Intel(R) Core(TM) i5-3210M CPU @ 2.50 GHz, 8 GB memory and Windows 10 (64-bit) operating system was used in this research to implement and evaluate classification models.

4.1 Correlation Coefficient analysis

In order to clarify the nature of the collected dataset as much as possible, the analysis of the correlation coefficient between the 12 features in the dataset has been used. Figure 2 and Figure 3 show the results of these analyzes that show the relationship of each feature with all other features (and itself) in the numerical format in the range of 0 and 1 and the color spectrum from light to dark (both respectively meaning the lowest and highest value) are displayed. Independent features and dependent features can be identified as a result of this analysis.

4.2 Evaluation Criteria

The evaluation criteria of the models used in this research include Accuracy, Precision, Recall, F1-Score, and the method of calculating these criteria can be seen in Equ. (1) to (4).

\[
\text{Accuracy} = \frac{TP + TN}{TP + FP + FN}
\]

\[
\text{Precision} = \frac{TP}{TP + FP}
\]

\[
\text{Recall} = \frac{TP}{TP + FN}
\]

\[
F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

4.3 Results

The evaluation results of different models based on the evaluation criteria for the values of 5 and 10 k can be seen in Figure 2 and Figure 3.

As can be seen in Figure 2 and Figure 3, for both values of k, K-NN and XGBoost models have the best performance compared to other models. Another noteworthy point is that the neural network model has shown better performance for K equal to 5 compared to K equal to 10. Also, the confusion matrix of the seven models are presented in the Figure 4 and Figure 5.

In order to deepen the results of the ROC diagram, the AUC value of each model was calculated and as can be seen in Figure 6 and Figure 7, the ROC diagram for each model is presented separately. Accordingly, Naïve Bayes and K-NN show higher AUC values.

Based on the results of the XGBoost model, we found that in our prediction model, WBC, Lymphocytes, LDH Serum, D-Dimer, Ferritin, ESR, CPK, and saturation are the most important features.

According to the order of importance of features obtained in XGBoost modeling, in the continuation of our research, we re-evaluated all seven models with 9 important features of the XGBoost model. As can be seen in Figure 9 and Figure 10, the comparison of models using 9 features in two evaluation modes with different values of k are shown. The results indicate that in this case, K-NN and XGBoost models have shown the best performance compared to other models. After these two, the Naïve Bayes model is in third place.
Comparing Supervised Machine Learning Models for Covid-19 patient detection using a Combination of Clinical and Laboratory Dataset

Figure 4. Confusion matrix of prediction (5-Fold Cross Validation)

Figure 5. Confusion matrix of prediction (10-Fold Cross Validation)
Regarding the importance of features in the XGBoost model, we evaluated again the models using 9 of the most important features.

Figure 9 and Figure 10, show a comparison between the results of the seven models used in this paper in terms of evaluation criteria for 5 and 10-fold Cross-Validation using nine features.

Also, as seen in Figure 11 and Figure 12, the Accuracy criterion, as one of the important criteria in comparing models, has the same values in both K-NN and XGBoost models in two evaluation modes with 9 and 12 features. In other words, reducing the features from 12 to 9 in the two mentioned models did not decrease the accuracy.
Comparing Supervised Machine Learning Models for Covid-19 patient detection using a Combination of Clinical and Laboratory Dataset

As shown in Figure 8 and Figure 9, the XGBoost and K-NN in both experiments perform better than the other models based on 5 and 10-fold Cross-Validation. Also, Naïve Bayes and neural networks are in the next position. In addition, the results showed that the decision tree and the random forest in the dataset of this study do not perform well compared to other models. Figure 13 shows the Accuracy of all models in terms of evaluation criteria for 5 and 10-fold Cross-Validation. XGBoost and K-NN have the highest Accuracy.

As can be seen, based on the comparison of the results in 9 and 12 features, the XGBoost and KNN methods have performed better than other methods. In other words, by reducing the features from 12 to 9 while reducing the overhead, almost similar results can be obtained.
5. Conclusion

The COVID-19 pandemic has had a significant impact on the life of people and public health around the world, putting the world's health care systems at risk and causing irreparable damage. In order to prevent the spread of this disease, the first step is early diagnosis of it and, as a result, the reduction of morbidity and mortality. In this regard, the use of new technologies based on artificial intelligence, especially in the field of machine learning, can play a significant role in this matter.

In this research, seven supervised machine learning techniques, namely K-NN, SVM, Decision Tree, Random Forest, Naïve Bayes, Neural Network and XGBoost, were tested and evaluated for the diagnosis of Covid-19 based on a brand new Iranian dataset. The results of these evaluations were presented in detail in the form of evaluation criteria, RUC curves, confusion matrices, and other comparative charts in the text of the paper. Finally, the results of the research show that the K-NN and XGBoost models have the best performance compared to the rest of the models.

The dataset used in this research included 12 features for 1354 patients that one by one were reported from rereading the files and medical records of the patients in the hospital, it is obvious that the cost of collecting such a dataset is very high. Therefore, in the following, we tried to distinguish more important features from less important features based on the results of the XGBoost algorithm in order to reduce the
dataset collection costs as well as the system computing costs as well. Based on the results, 3 features, namely PR, RR and CRP-Quantitative were recognized as less important features and to prove the correctness of such a claim, all models were trained and evaluated again by removing these three features from the dataset (only 9 features remained) and the
performance results of this experiment were also reported in detail. The final results showed that there was no significant change in the performance of the models, and if necessary, the three mentioned features can be ignored to reduce costs. Also, as stated, the data used in this research were the clinical and laboratory data of the patients. The use of other types of data such as the images of the lungs of the patients (CT-scans or X-rays) in a combined manner along with these data and comparing the results of the models can be a good solution for further studies on this issue.

Declarations

Funding
This research did not receive any grant from funding agencies in the public, commercial, or non-profit sectors.

Authors’ contributions
NM: Study design, acquisition of data; MT: Study design, acquisition of data; MA: Statistical analysis, drafting the manuscript, revision of the manuscript; MK: Supervision, revision of the manuscript; AK: Study design, Interpretation of the results.

Conflict of interest
The authors declare that there is no conflict of interest.

References


Comparing Supervised Machine Learning Models for Covid-19 patient detection using a Combination of Clinical and Laboratory Dataset

Narges Mohebabi was born in 1996 in Tehran. She received her B.Sc. in Computer Engineering from Arak University in 2018. Currently, she is an M.Sc. student in Information Technology at Tarbiat Modares University. Her research interests include data mining and machine learning.

Mehdi Tutunchian was born in 1996 in Tabriz. He received his B.Sc. in Information Technology Engineering from Tabriz University in 2018. Currently, he is an M.Sc. student in Information Technology at Tarbiat Modares University. His research interests include medical image processing and machine learning.

Meysam Alavi was born in Hamadan, Iran. He received his B.Sc. and an M.Sc. degree in Computer Engineering in 2005 and 2010, respectively. Also, he received his Second M.Sc. degree in Information Technology at Tarbiat Modares University in 2020. His research interests include medical learning, data mining, and social network analysis.

Mehrdad Kargari received his PhD degree in industrial engineering from Tarbiat Modares University of Iran. He is currently an assistant professor at the department of Information Engineering, in Tarbiat Modares University. His research interests are in the fields of machine learning, artificial intelligence, IoT and their applications in health or Banking.

Amir behnam kharazni was born in 1985. He is an internal medicine specialist, pulmonary subspecialist, and assistant professor at Shahid Beheshti medical university. Also, he is a member of the European Respiratory Society(ERS) and American thoracic society(ATS). His research interest is in the field of the application of artificial intelligence in medicine.


