RESEARCH ARTICLE **SPECIAL ISSUE**

RSP Science Hub

International Research Journal on Advanced Science Hub
2582-4376
Vol. 05, Issue 05S May
www.rspsciencehub.com

Check for updates

http://dx.doi.org/10.47392/irjash.2023.S046

International Conference on intelligent COMPUting TEchnologies and Research (i-COMPUTER) 2023

Prediction of COVID-19 using Machine Learning Models based on Clinical Blood Test Data

Hari Priya N [1], Rajeswari S [2]

[1]Research Scholar, Department of Computer Science, Sree Saraswathi Thyagaraja College, Pollachi, Tamil Nadu, India.
[2]Associate Professor, Department of Computer Science, Sree Saraswathi Thyagaraja College, Pollachi, Tamil Nadu, India.

Email: haripriyanarasimma@gmail.com

Article History

Received: 1 March 2023
Accepted: 22 March 2023

Keywords:

COVID19;
Blood test;
Machine Learning;
Random Forest;
Feature Selection;
Recursive Feature Elimination

Abstract

The global pandemic of Coronavirus Disease 2019 (COVID-19) has caused serious problems and threatened the lives of many people. To effectively combat the disease, early and precise screening of infected individuals is essential. The study uses blood test data which comprises 1736 instances and 35 features that have been collected from the patients who were admitted to the emergency department at the San Raffaele Hospital. For predicting COVID-19 in patients, RT-PCR tests a-re widely used. Once a patient has been identified with the presence of COVID-19, the patient should approach a healthcare professional to determine the severity of the virus and appropriate medical treatment and supportive care should be provided. The patient's condition should be closely monitored to ensure that their health is improving and to detect any complications that may arise. For this purpose, blood test samples taken from the patient will help to diagnose his condition and the severity of the virus. In this work, a feature selection technique known as Recursive Feature Elimination (RFE) has been used to find out the optimal set of features that are highly related to the existence of COVID-19 in patients. The features obtained using RFE are then applied with a machine learning model and the best results are achieved using a Random Forest classifier with an accuracy of 89%.

1. Introduction

A highly contagious illness known as Covid-19 (Coronavirus Disease) has spread throughout the entire world. With a high morbidity and fatality rate, the COVID-19 pandemic has had a huge impact on the world's healthcare system. In order to prevent the disease from spreading and to provide effective treatment, early diagnosis of COVID-19 cases is essential. Using clinical data, including the results of blood tests, machine learning models have emerged as a viable method for diagnosing the COVID-19 severity in patients. Findings will be essential in the first screening for COVID-19 as recent clinical studies have indicated that the blood parameters of COVID-19 patients fluctuate significantly. Initial screening offers a preliminary probabilistic indicator of the disease's presence, whereas diagnosis verifies the disease's presence or absence. However, machine learning algorithms are capable of identifying and differentiating different patterns present in common blood tests parameters.

A technique based on machine learning should be used to determine the latent relationships between the blood test parameters and the existence of

COVID-19 (Hany et al.). After finding out that a person has been affected with COVID-19 using RT-PCR, further diagnosis is necessary for examining the patient's condition thereby knowing the prevailing health status of the patient to provide better treatment.

The objective of the work is to share the results of the blood test report which includes complete blood count (CBC) of COVID-19 patients as additional information about this virus. This would help in supplying physicians with crucial information on the changes that may be anticipated from the blood test report of COVID-19 affected patients. In such a case, routine blood test samples of the patient can be used to diagnosis his medical condition and the severity of the disease. Hence, blood test data is considered for this study.

The motivation of this research is to examine the potential of machine learning models for detecting COVID-19, after predicting its severity using data from blood tests reports. It is essential to identify and anticipate COVID-19 patients as early as possible in order to stop the disease's spread and provide appropriate care. Blood tests are a frequently utilised diagnostic procedure that can yield useful details about a patient's health status, including indicators of immunological response that are pertinent to COVID-19. On the basis of this data, machine learning algorithms can be trained to find patterns and forecast the probability of COVID-19 infection.

Big data creates new possibilities for studying how this pathogen behaves. Due to its similarity to other respiratory disorders like influenza, early identification is difficult. The RT-PCR test is currently the gold standard for COVID-19 diagnosis (O Abayomi-Alli et al.). However, it is susceptible to false negatives and inaccurate outcomes. As a result, and as an alternative method X-rays, blood tests, CT scans and sound analysis can all be used to accurately diagnose COVID-19. The preceding methods can be used in situations such as pandemic peaks (Chadaga et al.).

In the existing study (Hany et al.), they have considered RF, SVM and Naïve Bayes and the accuracy of each classifier were 76%, 88%, 85%. The drawback which we have identified in the existing study was that they haven't adopted any feature selection technique before training the model. The removal of redundant and irrelevant features that would not improve the model's predictive ability is aided by feature selection. Reducing the number of features makes the model more accurate and efficient, which also improves the performance of the model.

The main contribution of this work is to identify the most crucial blood test parameters that potentially leads to the presence of COVID -19 in a patient. It also aims to achieve a higher rate of accuracy in a machine learning algorithm by adopting a feature selection technique before training a model.

2. Background Study

Some authors have already used machine learning to diagnose and predict this devastating disease. Several studies use standard evaluation techniques in an effort to detect COVID-19 infection such as antigen tests, CT scans and X-rays. Other reliable techniques for diagnosing COVID-19 include blood tests and sound analysis. The literature for some of the diagnostic models has been reviewed in this section.

The authors employed Random Forest and SVM on a dataset comprising of 294 blood samples taken from the Chinese hospital, Kunshan People's Hospital and Wuhan Union Hospital. SVM outperforms random forest classifier with an accuracy of 84%, according to experimental results on the fifteen features that were chosen for analysis (Bao, Sheng, et al.).

The diagnosis of COVID-19 was suggested by authors in (De Moraes, Batista, Filipe, et al.) using five different machine learning algorithms such as neural networks, random forest, gradient boost trees, logistic regression, SVM. A dataset containing 235 blood samples and 102 confirmed cases of COVID-19 was provided by the Albert Einstein Hospital in Brazil. 15 pertinent characteristics were chosen from this dataset for the analysis, with an AUC score of 85 %, sensitivity of 68 %, and specificity of 85% respectively.

Aljame et al. (Aljame et al.) developed an ensemble learning approach for the initial screening of COVID-19 from standard blood tests. Using data from 564 patients at the Albert Einstein Israelita Hospital which is located in Sao Paulo in Brazil, the model successfully distinguished COVID-19 positive cases with a 99.88% accuracy.

In order to detect the presence of COVID-19 from ordinary blood samples, the authors took into

account k-nearest neighbors, SVM, naive bayes (NB), LR and RF as five machine learning models. Patients admitted to San Raphael Hospital in Italy (52% COVID-19 positive), were asked to provide 1,624 routine blood samples. The accuracy, area under the curve (AUC), sensitivity and specificity of the models were in the ranges of 74–88%, 70–89%, 79–92%, and 74–90%, respectively (Cabitza et al.).

Deep learning models are being used to diagnose COVID-19 from routine blood testing was first reported by the authors of a recent study. For evaluation, the deep learning models chosen by the authors were ANN, CNNLSTM, CNNRNN, recurrent neural networks (RNNs), CNNs and long-short term memory (LSTM). The dataset comprises of samples of 600 patients examined at the Albert Einstein Hospital in Brazil, 18 blood characteristics were used to train and evaluate the models. The CNNLSTM with the train-test split strategy was the best performing algorithm, with an accuracy of 92.3% (Alakus, Turkoglu, et al.).

The possibility of machine learning algorithms to forecast COVID-19 based on clinical blood test data has been examined in a number of publications. The majority of these research have shown promising findings, indicating that machine learning models may correctly identify the severity level of COVID-19 affected individuals based on the outcome of blood tests. With the help of detailed literature review, we could able to figure out which type of algorithm could be suitable for blood test data as the dataset comprises data that are mostly numeric in type. The results of the literature study may serve to direct the creation of more precise and reliable machine learning models for COVID-19 prediction based on clinical blood test data, which could have important consequences for reducing the propagation of the disease and boosting patient outcomes.

3. Materials and Methods

In this section, the working process of the proposed methodology is depicted in Figure 1. The process flow starts with Covid-19 data collection followed by pre-processing of the data. Then the processed data is applied with feature selection technique before training the model. The obtained features are used by machine learning algorithm such as Random Forest, SVM and XGBoost and the evaluation metrics of these classifiers were calculated.

As the dataset included in the study comprises numeric data, several machine learning algorithms were taken into consideration. Based on the characteristics of the machine learning algorithm and the analysis from the existing study, the algorithms were chosen. In the existing study, authors have chosen algorithms such as SVM, RF and Naïve Bayes. In order to show the enhanced performance of proposed work, the algorithms which was considered in the base paper has been chosen. XGBOOST has been used instead of Naïve Bayes, as NB can't handle non-linear relationships that exists between the dependent and independent variable.

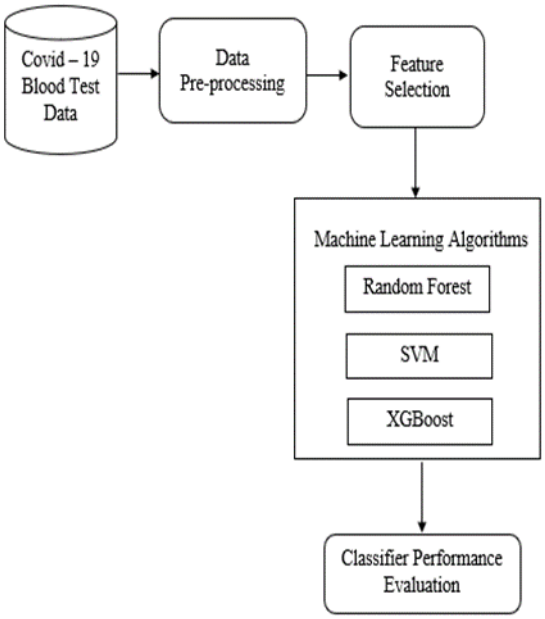

FIGURE 1. Proposed Workflow

3.1. Dataset Description

The dataset comprises of routine blood test results from 1736 patients admitted to the San Raffaele Hospital. The blood-test data contains 1736 samples out of which 816 samples were reported as positive to COVID-19 and 920 samples were reported as negative. The dataset comprises of a total of 35 features including basic demographic information of the patient and the clinical information. The target variable specifies whether the patient is affected with Covid-19 or not. The total count of gender category pertaining to positive and negative cases of patients is depicted in Figure 2. The list of all features used in this study is shown in Table 1 along with its abbreviation.

3.2. Data Pre-Processing

The fineness of the input data determines whether pre-processing is required. The statistical mean-based imputation techniques are used to fill in the missing values that frequently occur in results of routine blood tests. Outliers should be identified and removed as well because they present data that is noticeably different from the majority of the sample (Kistenev et al.). The missing values in the dataset has been treated with an imputation technique known as Multiple imputation which is typically a more reliable method for imputed missing values, especially when the missing data is non-random or when the percentage of missing data is substantial. In order to obtain reliable statistical conclusions and account for the uncertainty in the imputed values, multiple imputation entails producing numerous imputed datasets based on a statistical model.

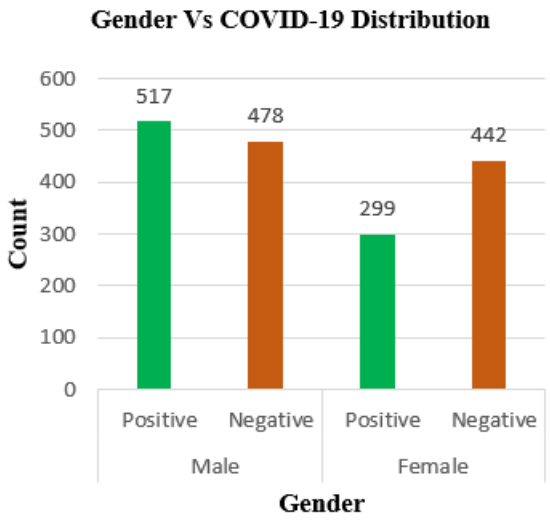

FIGURE 2. Gender Vs COVID-19 Distribution

3.3. Feature Selection

A machine learning model is built through the process of feature selection, which involves choosing a subset of pertinent features. It is a crucial step after pre-processing of the data, since it can enhance model performance, lessen overfitting, and lower the computational expense of training the model.

In this work, we have included wrapper feature selection method known as Recursive Feature Elimination (RFE) for getting the optimal set of features. Initially, we had 34 features; after removing strongly correlated features (greater than 90%) we had left with 29 features and after applying recursive feature elimination technique, finally we had 15 features including the target variable as the result of feature selection technique. The visualization of the features obtained from RFE using matplotlib library

TABLE 1. Features (blood test result parameters)

S. No	Acronym	Abbreviation
1	WBC	White blood cells
2	AST	Aspartate aminotransferase
3	CRP	C-reactive protein
4	LYT	Lymphocytes count
5	MOT	Monocytes count
6	ALP	Alkaline phosphatase
7	RBC	Red blood cells
8	MCV	Mean corpuscular volume
9	LDH	Lactate dehydrogenase
10	CA	Calcium
11	EO	Eosinophils count (%)
12	BA	Basophils count (%)
13	PLT	Platelets
14	CREA	Creatinine
15	GGT	Gamma glutamyl transferase
16	MO	Monocytes count (%)
17	ALT	Alanine aminotransferase
18	MOT	Monocytes count
19	NAT	Sodium
20	NE	Neutrophils count (%)
21	HGB	Hemoglobin
22	EOT	Eosinophils count
23	CK	Creatine Kinase
24	MCHC	Mean corpuscolar hemoglobin
25	UREA	Urea
26	LY	Lymphocytes count (%)
27	GLU	Glucose
28	NET	Neutrophils count
29	KAT	Potassium
30	HCT	Hematocrit
31	BAT	Basophils count
32	MCH	Mean corpuscolar hemoglobin
33	SEX	Sex
34	AGE	Age
35	TARGET	Target (Covid-19 Positive/Negative)

of python is depicted in Figure 3 along with its rankings.

After applying feature selection on the dataset, it has been split into 70% of training data and 30% of test data to be applied on the machine learning model to predict its performance.

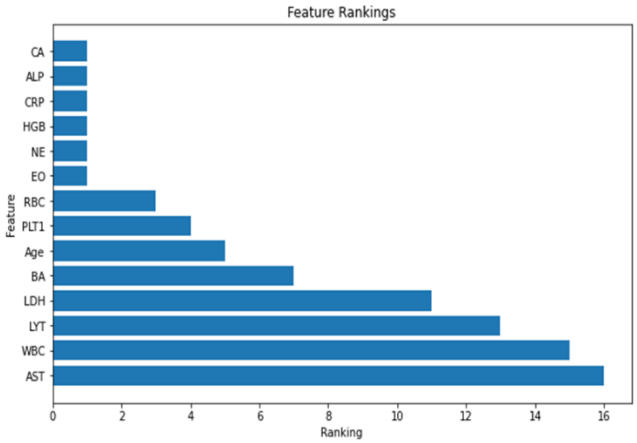

FIGURE 3. Visualization of Feature Ranking from RFE

3.4. Machine Learning Algorithms

3.4.1. Support Vector Machine (SVM)

SVM is a machine learning technique used for classification, regression, and outlier detection. In order to distinguish between the various classes in the dataset, the SVM algorithm seeks out the optimum hyperplane. SVM aims to maximize the margin between classes while minimizing classification error (Chauhan, Dahiya, Sharma, et al.). The SVM determines the best super plane and divides the data points based on their distance from it. It performs well in high-dimensional spaces, making it capable of handling datasets with a large number of features (Mazloumi et al.).

3.4.2. Random Forest

The random forest (RF) is a hierarchical grouping of base classifiers with a tree topology. The most significant and relevant feature is chosen using a predefined probability by the RF algorithm. Prominent machine learning algorithms like random forest are utilized for classification and regression tasks. It is an ensemble learning method that combines multiple decision trees to create a more accurate and stable model (Brinati et al.). This procedure assists in lowering prediction error and enhancing accuracy of the prediction.

3.4.3. XGBoost

XGBoost is a popular open-source machine learning library for developing and training gradient boosting models. It is a development of gradient boosting and renowned for its speed, accuracy and scalability (Rahman, Chowdhury, Amrin, et al.). The underlying idea of XGBoost is to train a series of decision trees in an iterative manner so that each new tree corrects the mistakes of the preceding ones. The algorithm modifies the weights of the data points based on their prior misclassifications and adds a new decision tree to the ensemble with each iteration (Mehta et al.).

4. Results and Discussion

In order to evaluate various parameters and produce the best machine learning model for COVID-19 prediction, three distinct machine learning models based on classification were used along with the feature selection technique. The work is implemented using Python and we used Python libraries used in this work includes pandas, numpy, scikit-learn, matplotlib and sea born.

Based on the results from the feature selection technique which is employed in this study i.e., Recursive Feature Elimination (RFE), the important blood test parameters which are highly related to the presence of COVID-19 has been identified. The parameters such as LDH and CRP whose values exceeding the normal range can lead to the presence of COVID-19. Leukocyte, platelet, and eosinophil counts that are low are a clear indication of existence of COVID-19. Important blood components that can reveal details about a patient's general health are RBC (Red Blood Cell Count) and HGB (Hemoglobin), the disease's severity may be made worse by low RBC and HGB levels. In this study, most of the patients whose LDH and CRP values greater than the normal range and CA, RBC, HGB, WBC values that are lower than the normal range were tested as COVID-19 positive. Also, the increased levels of the two liver enzymes AST and ALP in patients led to the severe presence of COVID-19.

The algorithms considered for the study were Random Forest, SVM and XGBoost. Out of the three algorithms, accuracy of the outcomes from Random Forest was 89%. Comparing the performance metrics of the proposed work with exist-

ing methodology, accuracy of the random forest algorithm has been increased from 76% to 89%. The performance metrics of these algorithms were shown in Table 2.

TABLE 2. Performance Metrics

Algorithm/ Measures	Accuracy	F1 score	Precision	Recall
Random Forest	89%	88%	89%	86%
SVM	81%	79%	82%	75%
XGBOOST	83%	81%	83%	79%

5. Conclusion and Future Work

The ability to identify individuals at risk of infection and disease progression can help guide treatment and prevention strategies, which is especially important in times of pandemic. The study uses blood test data for predicting the criticality and health condition of the patients. The most vital features that are highly related to the presence of COVID-19 were Lactate dehydrogenase (LDH), Alkaline phosphatase (ALP), C-reactive protein (CRP), Platelets, Aspartate aminotransferase (AST) and the different types of WBC including Neutrophils, Lymphocytes, Eosinophils and Basophils. According to experimental findings, the random forest classifier performs better comparing with other classifiers such as SVM and XGBoost that has been used in this study for COVID-19 diagnosis.

For the future work, larger and more diverse datasets that include a variety of patient populations, demographic factors, and disease stages can be used that can improve the accuracy and generalizability of COVID-19 predictions. Another suggestion will be to develop machine learning models that can provide real-time predictions and decision support system to healthcare providers.

References

Alakus, Talha Burak, Ibrahim Turkoglu, et al. "Comparison of deep learning approaches to predict COVID-19 infection". *Chaos, Solitons & Fractals* 140 (2020): 110120–110120.

Aljame, Maryam, et al. "Deep Forest Model for Diagnosing COVID-19 From Routine Blood Tests". *Scientific reports* 11 (2021): 16682–16682.

Bao, Forrest, Sheng, et al. "Triaging moderate COVID-19 and other viral pneumonias from routine blood tests". (2020).

Brinati, Davide, et al. "Detection of COVID-19 Infection from Routine Blood Exams with Machine Learning: a Feasibility Study". *Journal of medical systems* 44 (2020): 1–12.

Cabitza, Federico, 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)* 59.2 (2021): 421–431.

Chadaga, Krishnaraj, et al. "Medical diagnosis of COVID-19 using blood tests and machine learning". *Journal of Physics: Conference Series* 2161.1 (2022): 012017–012017.

Chauhan, Vinod Kumar, Kalpana Dahiya, Anuj Sharma, et al. "Problem formulations and solvers in linear SVM: a review". *Artificial Intelligence Review* 52.2 (2019): 803–855.

De Moraes, Andre Batista, Filipe, et al. "COVID-19 diagnosis prediction in emergency care patients: a machine learning approach". *MedRxiv* (2020): 2020–2024.

Hany, Noran, et al. "Detection COVID-19 using Machine Learning from Blood Tests". *2021 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)* (2021).

Kistenev, Yury V, et al. "Predictive models for COVID-19 detection using routine blood tests and machine learning". *Heliyon* 8.10 (2022): e11185–e11185.

Mazloumi, Rahil, et al. "Statistical analysis of blood characteristics of COVID-19 patients and their survival or death prediction using machine learning algorithms". *Neural Computing and Applications* 34.17 (2022): 14729–14743.

Mehta, Mihir, et al. "Early Stage Machine Learning–Based Prediction of US County Vulnerability to the COVID-19 Pandemic: Machine Learning Approach". *JMIR Public Health and Surveillance* 6.3 (2020): e19446–e19446.

O Abayomi-Alli, Olusola, et al. "An Ensemble Learning Model for COVID-19 Detection from Blood Test Samples". *Sensors* 22.6 (2022): 2224–2224.

Rahman, Md. Siddikur, Arman Hossain Chowdhury, Miftahuzzannat Amrin, et al. "Accuracy comparison of ARIMA and XGBoost forecasting models in predicting the incidence of COVID-19 in Bangladesh". *PLOS Global Public Health* 2.5 (2022): e0000495–e0000495.

Embargo period: The article has no embargo period.

To cite this Article: , Hari Priya N, and Rajeswari S . "**Prediction of COVID-19 using Machine Learning Models based on Clinical Blood Test Data**." International Research Journal on Advanced Science Hub 05.05S May (2023): 338–344. http://dx.doi.org/10.47392/irjash.2023.S046