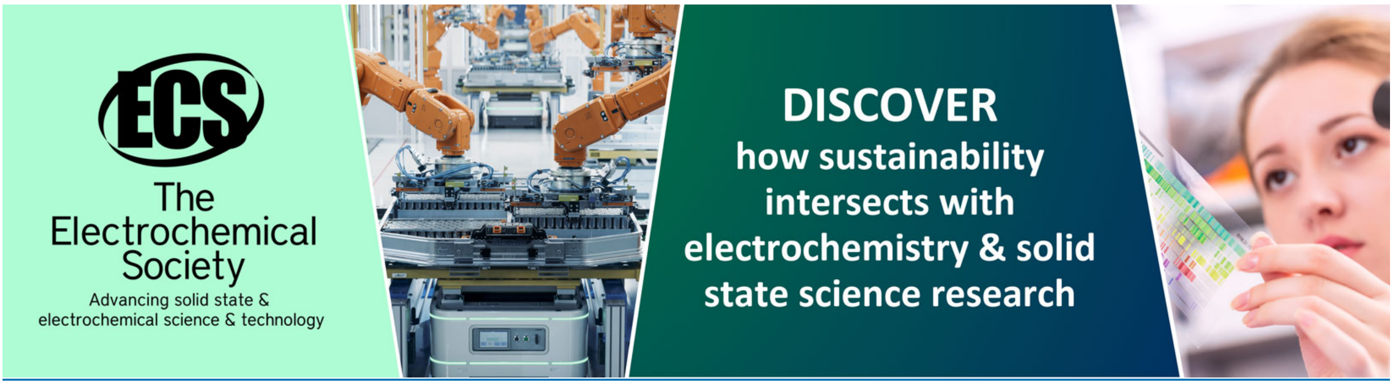
Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Medical diagnosis of COVID-19 using blood tests and machine learning

To cite this article: Krishnaraj Chadaga *et al* 2022 *J. Phys.: Conf. Ser.* **2161** 012017



Medical diagnosis of COVID-19 using blood tests and machine learning

Krishnaraj Chadaga[1], Srikanth Prabhu[1]*, Vivekananda Bhat K[1], Shashikiran Umakanth[2] and Niranjana Sampathila[3]

[1]Department of Computer Science and Engineering, Manipal Institute of Technology, Manipal, India
[2]Department of Medicine, Dr. TMA Pai Hospital, Udupi India
[3]Department of Biomedical Engineering, Manipal Institute of Technology, Manipal India
* Corresponding author

Abstract. Severe Acute Respiratory Syndrome Coronavirus 2(SARS-CoV-2), colloquially known as Coronavirus surfaced in late 2019 and is an extremely dangerous disease. RT-PCR (Reverse transcription Polymerase Chain Reaction) tests are extensively used in COVID-19 diagnosis. However, they are prone to a lot of false negatives and erroneous results. Hence, alternate methods are being researched and discovered for the detection of this infectious disease. We diagnose and forecast COVID-19 with the help of routine blood tests and Artificial Intelligence in this paper. The COVID-19 patient dataset was obtained from Israelita Albert Einstein Hospital, Brazil. Logistic regression, random forest, k nearest neighbours and Xgboost were the classifiers used for prediction. Since the dataset was extremely unbalanced, a technique called SMOTE was used to perform oversampling. Random forest obtained optimal results with an accuracy of 92%. The most important parameters according to the study were leukocytes, eosinophils, platelets and monocytes. This preliminary COVID-19 detection can be utilised in conjunction with RT-PCR testing to improve sensitivity, as well as in further pandemic outbreaks.

Keywords: COVID-19, RT-PCR, machine learning, blood tests, logistic regression, K nearest neigbhours, Xgboost, SMOTE, random forest

1. Introduction

Coronavirus is a very contagious disease that has spread all over the world [1]. The infectious virus was first discovered in bats and was then transferred to humans in Wuhan, China [2]. The SARS-CoV-2 virus is extremely dangerous since it is spreads faster than its close relative, SARS. According to simulation and computer modelling approaches, every new COVID-19 case infects an average of 2.67 people worldwide [3]. Social isolation, rapid and early identification and isolation are the most effective ways to combat this new deadly infection. The incubation period can vary from 3 to 15 days according to WHO (World Health Organisation). The effect of the viral infection is often asymptomatic, or people develop influenza-like symptoms such as fever, cough and shortness of breath. Myocardial infarction or chest pressure are more serious signs and happen to only a small part of the population. Early detection is a challenge since it resembles other respiratory diseases like influenza. For COVID-19 diagnosis, the RT-PCR (Reverse Transcription Polymerase Chain Reaction) test is presently the gold standard [4]. However, it is prone to false negatives and incorrect results. It's also possible that it won't be able to detect newer COVID-19 strains in the future. As a result, CT scans, X-rays, blood testing, and sound analysis can all be utilised to accurately diagnose COVID-19 as an alternate technique. Because the RT-PCR test takes a long time to give findings, the above approaches can be used in scenarios like pandemic peaks. To boost sensitivity, these techniques can be utilised in conjunction with normal RT-PCR testing.

In the fight against this highly infectious virus, machine learning is already playing a critical role. It also makes a substantial contribution to academic and clinical research [5]. Machine learning has a lot of promise in engineering, multidisciplinary science, psychology, analytical practice, earth sciences, hazard mitigation strategies, urbanised environments, universal healthcare, and other fields. In this study, we present an early filtering technique for diagnosing COVID-19 using regular blood tests. This algorithm can assist clinicians in determining who should be tested, but it is not intended for exact diagnosis. Before using the standard COVID-19 detection tests, our technique can be utilised as a preliminary screening measure. The COVID-19 patient can be discharged if it reveals a low or zero percent chance of infection, and no further testing may be required. Otherwise, additional testing may be required to validate the findings. The sensitivity of COVID-19 detection can also be improved if the training dataset contains additional examples of patients who have been infected with other viruses. The structure of this paper is described as follows: In Section 2 we look at the existing literature that uses machine learning to tackle COVID-19. In section 3, we perform exploratory data analysis to pre-process our data and also find out various co-relations between the attributes. In Section 4, description of the various classification algorithms used for COVID-19 preliminary filtering is presented. In section 5, we discuss the experimental results. The paper concludes in section 6.

2. Related literature

Machine learning has already been utilized by a number of academics to identify and predict this dreadful disease. Many researchers attempt to diagnose COVID-19 infection using typical evaluation techniques (analysis), such as X-rays, CT scans and antigen tests. However, COVID-19 can be diagnosed using blood tests and sound analysis too. In this section, we perform a literature survey of some of above diagnostic models. Khuzani et al., [6] suggested a fully automated machine learning solution to assist health care providers in accurately diagnosing COVID-19 utilising chest X-rays (XSR) pictures. A dimensionality reduction strategy was used to create an effective classifier that was able to detect COVID-19 instances with excellent accuracy and sensitivity. This model had a sensitivity of 100 percent and a precision of 96 percent. Rasheed et al. [7] developed two models for successfully diagnosing and predicting COVID-19 from XSR images: convolutional neural networks (CNN) and logistic regression. In comparison to traditional approaches, the deep learning-based algorithms require a high number of training examples. However, for COVID-19 XSR images, an adequate number of labelled training data was not available. To expand the training data and alleviate the problem of overfitting, a dataset augmentation strategy using generative adversarial networks was used. The LR and CNN models achieved an overall accuracy of 95-97%. Sharma et al., [8] proposed a model which diagnosed the deadly virus from CT Scans. Lung CT scans of infected patients from Italy, China, Russia and India were chosen and custom Microsoft Azure vision machine learning techniques were used for training, testing and deployment. An overall accuracy of 91% was obtained. Serte et al., [9] used deep learning models that used 3D CT scans to diagnose COVID-19. The system used ResNet-18 model along with the conjunction of majority voting algorithm. The proposed model achieved an AUC of 96% for diagnosing COVID-19. Pahar et al., [10] used a classification model to classify COVID-19 patients using smart phone recordings. The public dataset, Coswara has the voice recordings of 92 COVID-19 suspected patients and 1079 healthy people. COVID-19 positive coughs were 15%-20% shorter in wavelength, according to the study. Seven machine learning classifiers were used for this purpose. However, the best results were obtained by Resnet50 which discriminated COVID-19 patients with an AUC of 0.98. A thorough examination of breathing sounds and their significance in identifying respiratory difficulties was made by Faezipour and Abuzneid., [11]. COVID-19 patients' breathing sounds may indicate a certain acoustic signal pattern that are worth studying, claimed the paper. Obtaining respiratory data from breathing sounds using a smart phone's microphone appears to be a very interesting approach in this regard, according to the preliminary study. According to the article, advanced signal processing, as well as latest deep/ machine learning and pattern recognition algorithms, can be used to segregate oxygenation, measure lung volume and breathing intervals and further categorize breathing data into healthy or unhealthy groups. Brinati et al., [12] designed ML algorithms that diagnosed COVID-19 based on common blood tests. EHR reports of 279 patients (177 COVID-19 positive, 102 COVID-19 negative) were collected from San Raffaele Hospital (Milan, Italy). Many classification algorithms were used, however random forest achieved the highest results with an accuracy of 86%. The paper concluded that leukocytes, C- reactive protein (CRP), platelets, GGT, ALT, AST, neutrophils, monocytes, LDH, lymphocytes, basophils and eosinophils were extremely important parameters. Heldt et al. [13] established a machine learning-based algorithm that used laboratory characteristics to predict the severity of COVID-19. The dataset for the aforementioned model was made up of the details of 879 COVID-19 confirmed patients. The techniques employed were multivariable logistic regression, extreme gradient boosting trees, and random forest. Age, oxygenation rate, creatinine and blood lactate levels were the most predictive parameters according to the study. AUC of 0.76-0.87

were obtained by these models. The paper concluded that artificial intelligence can help early identification of patients with poor prognosis. Soares et al., [14] presented a ML classifier that accurately diagnosed COVID-19 using haematological and demographic parameters. Details of 5644 patients from Albert Einstein Hospital, Brazil along with 16 parameters were considered for the final model. The model developed was called "ER-CoV". The specificity, sensitivity and AUC obtained were 85.98%, 70.25% and 86.78% respectively. Red blood cells, leukocytes, basophils, monocytes, lymphocytes and platelets were deterministic parameters, claimed the study. Czako et al., [15] used an adaptive AI platform to perform preliminary patient filtering. The hybrid algorithm was called PSO-SA that diagnosed COVID-19 using a generic machine learning solution along with hyper parameter tuning techniques.

3. Exploratory data analysis

3.1. Dataset description

The anonymous data we used in our study came from patients who visited the Albert Einstein Hospital, Brazil. The COVID-19 and other routine tests, were performed on the patients. All data was anonymized in accordance with industry best practises and requirements. The clinically obtained data was normalised with zero mean to obtain a perfect normal distribution. The hospital provided the data to Kaggle, and it covers the days of March 28th and April 3rd, 2020 [16]. It includes the COVID-19 labels (both positive and negative) and has 5644 instances and 111 variables. 558 (about 10%) of the 5644 tests were positive for COVID-19. The dataset was very unbalanced, with a substantially fewer number of positive cases than negative cases. In the current scenario, this data set is accurate (Generally the number of healthy people is far greater than infected patients). It includes the results of a routine blood test, including haematocrit, platelets, haemoglobin, red blood cells, mean platelet volume, lymphocytes, leukocytes, MCHC (Mean corpuscular haemoglobin concentration), basophils and so on. The findings of many tests for viruses such as Influenza A, Respiratory Syncytial Virus, Influenza B, Rhinovirus/Enterovirus, Adenovirus, and others were also included. The label column is a binary column with "positive" for COVID-19-infected patients and "negative" for COVID-19 uninfected patients, based on the results of the RT-PCR test.

3.2. Dataset preprocessing

The dataset contains a lot of null values. Most of the features have null values of at least 80% with a high amount of them above 90% as showed in figure 1. Filling those missing values would render the model useless, so we dropped all the columns containing at least 90% of null values. After this procedure, the remaining attributes were 39 out of 101. Variances of the variables were also compared to check whether there existed only a lone value. The variable "Parainfluenza 2" had a variance of 0 (has only 1 value) and was dropped.

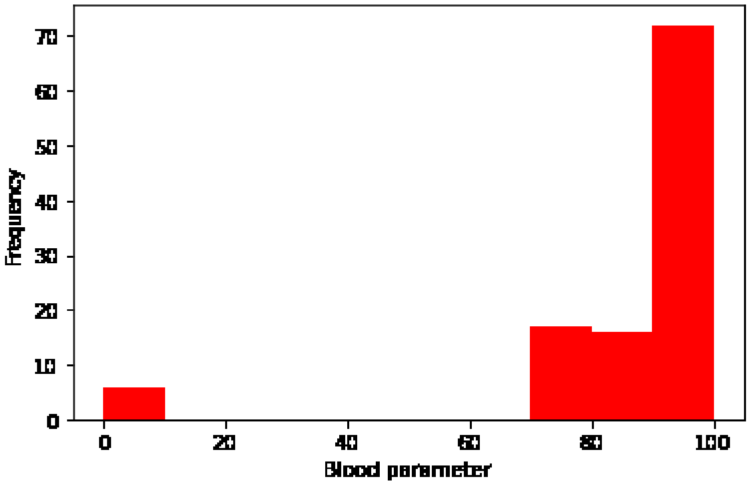


Figure 1. Frequency of null values vs blood parameters

Out of 5644 rows, 3596 had 32 columns with missing values. The dataset was trimmed again and all those rows were removed which had more than 26 null values. There were at least 19 columns in the dataset that mentioned the presence of various antigens. All the above variables were binary. When we analysed each of the variable, we realized the number of nulls was very high. Therefore, we created a new variable called 'has-disease' which contained the binary values of the corresponding 19 columns. This variable indicated, for each patient, if atleast one of those variables is positive. After thorough analysis, it was seen that 13% of the patients tested positive for atleast one of the antigens. The other blood variables were all continuous values. The author of the dataset had already normalised all of the columns, and each attribute had extremely very small values in the range of [-3, 3]. The lone exception is the" Patient age quantile" attribute, which has values ranging from [1,19]. As the magnitude of features is crucial in some AI algorithms (and might impact the final outcome), we changed the age column to have values in the range of [-3, 3], just like the attributes, to avoid the detrimental impact of attributes with various scales. After pre-processing of data, we were left with 20 columns and the missing data was imputed with the mean value of the corresponding columns.

3.3. Co-relation analysis and feature importance

We used Pearson's co-relation co-efficient (PCC) to evaluate the correlation between the attributes and the class label (COVID-19 positive/negative) as shown in figure 2. The PCC value ranges between -1 and 1. This study tries to focus whether a feature was mapped to the target variable. Some features showed strong co-relations that might indicate COVID-19. Leukocytes, eosinophils, platelets and has_disease attributes attained a co-relation of -0.3, indicating a negative co-relation. Monocytes, age, hemoglobin and red blood cells showed a slight positive co-relation (2nd column in the correlation heat map).

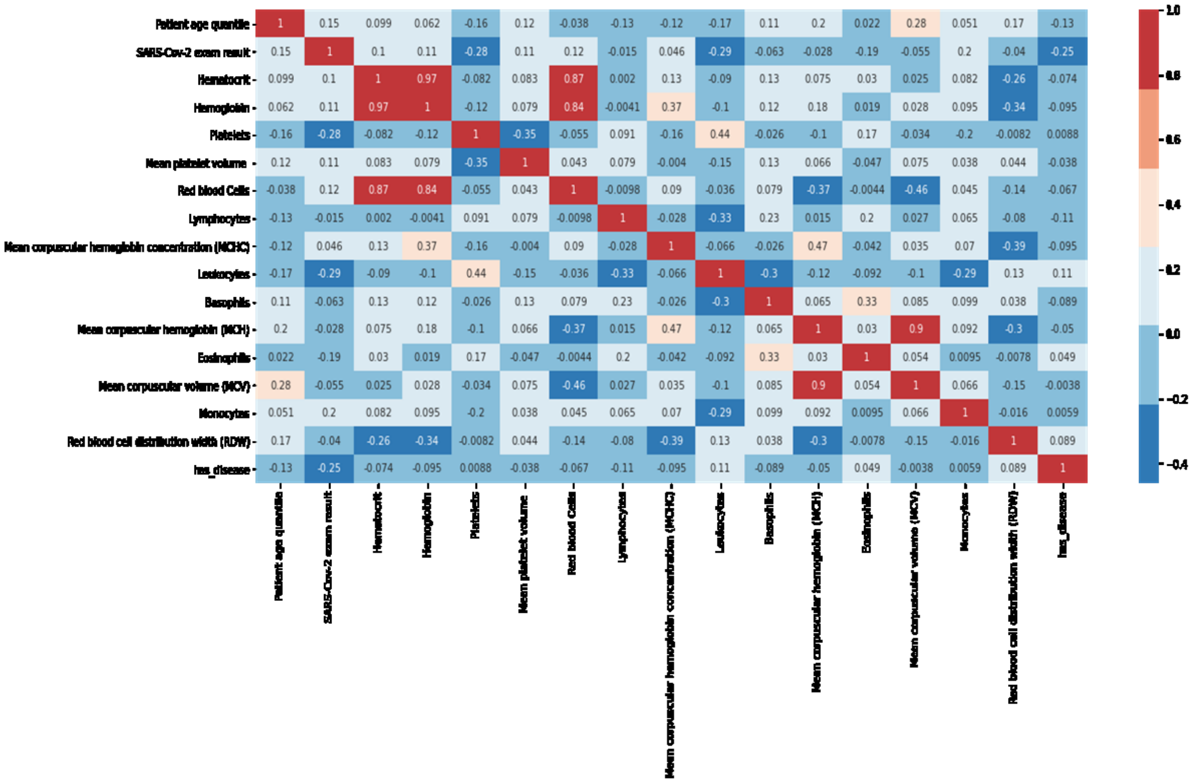


Figure 2. Pearson co-relation matrix

Some features had a high co-relation among themselves. In order to reduce noise, we need to remove the attributes which have a high collinearity among themselves. Haematocrit and haemoglobin had a correlation of 0.97. The above parameters had a high correlation with red blood cells as well (0.87 and 0.84). Their co-relations with the target are very similar as well, so we kept red blood cells which had the highest correlation (0.12) and removed haematocrit and haemoglobin. The other two variables which were highly correlated were MCV (Mean corpuscular volume) and MCH (Mean Corpuscular haemoglobin). MCV was more correlated to the label (-0.055 vs -0.,028) and was retained. In the medical domain, high levels of accountability and openness are required, which means machine judgments should be trustworthy and reliable. Feature importance was calculated using random forest and SHAP (Shapley Additive Explanations) method. These methods determine the relevance of each feature by monitoring the influence on model accuracy when each predictor feature is randomly shuffled. Figure 3 shows the value of features using random forest.

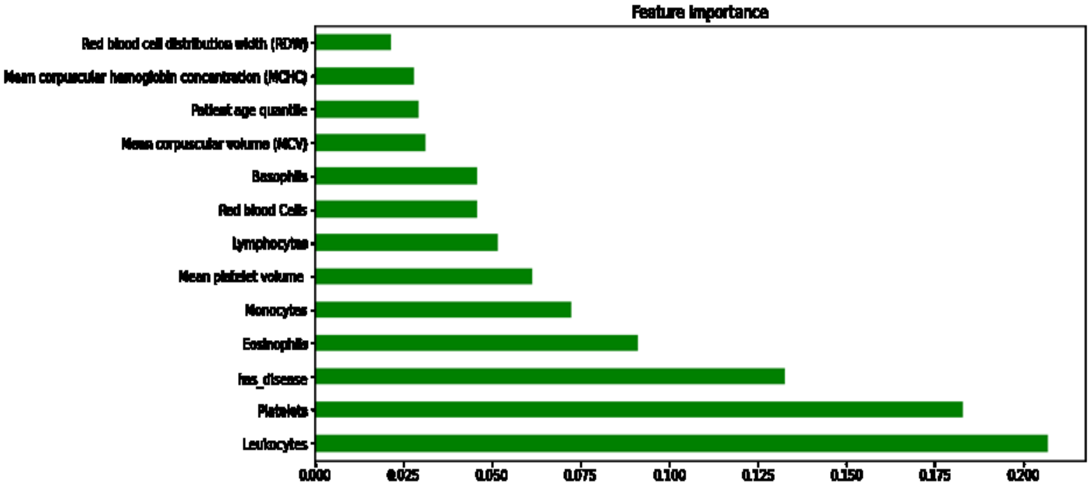


Figure 3. Feature importance using random forest

The most important features are leukocytes, platelets, has_disease, eosinophils, MPV and monocytes. These features were similar to the ones we saw on the Pearson co-relation heat map. SHAP was introduced by Lundberg and Lee [17]. It is a feature selection method that explains the dependence of each feature on the final outcome of the model. Shapley value for one feature is the average marginal contribution of a particular feature across all combination of features. We use the summary plot as shown in figure 4 to find out which features are more important for the model. This plot confirms and summarizes all the previous findings. Low values for eosinophils, platelets and leukocytes are a strong sign of the presence of COVID-19. Monocyte counts above a certain threshold can be a reliable predictor of COVID-19 infection. For the other parameters, there is a high skewness towards “not infected” with a muddy terrain around 0. For the final model prediction our data consisted of 18 columns.

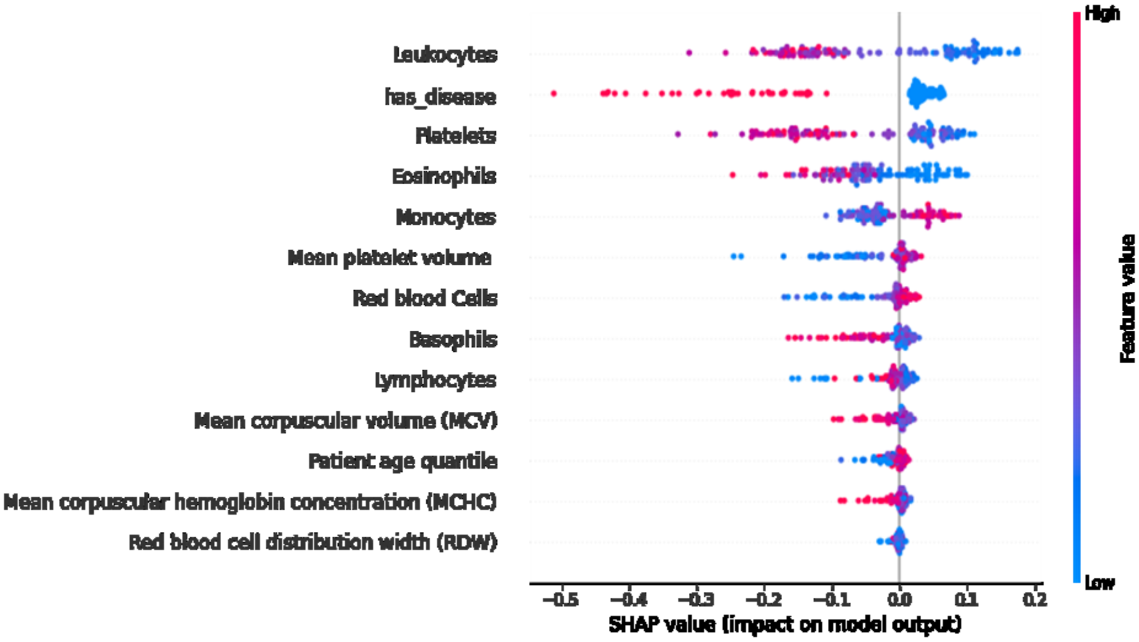


Figure 4. Feature importance using SHAP

4. Methodology

This section discusses about the various classification algorithms used to diagnose COVID-19. A sampling technique called SMOTE is also discussed. It is used to oversample the unbalanced dataset. The evaluation metrics are also explained and the solution pipeline is given in figure 5.

4.1. Classification algorithms

4.1.1 Random forest

It is a technique used to perform classification as well as regression in machine learning. Ade Culter and Leo Breiman developed this method in 1999, and it comprises of many distinct decision trees. Random Forest is made up of decision trees that are simple to deploy, develop, and analyse, but when it comes to classifying fresh samples, it is more complex and inflexible. As a result, Random Forest improves the classification accuracy by using the simplicity and flexibility of the decision trees. It offers advantages that motivate us to use it to accurately identify our dataset.

4.1.2 KNN

The KNN algorithm presumes that similar data points exist nearby. We use whole training instances to predict output for unknown data in this example, rather than weights from training data to predict output. The model does not learn from previous training data, and instead waits until a prediction on a fresh instance is asked before continuing. There is no predefined mapping function form in KNN. Choosing the value of K is critical because it plays a crucial role in classification and avoiding data overfitting.

4.1.3 Logistic regression

It uses a sigmoid curve as a cost function to predict a categorical variable (dependent variable) based on one or more independent factors. The Sigmoid function (Logistic function) is an S-shaped curve that divides data into classes. Binomial, multinomial, and ordinal classifications can all benefit from logistic regression.

4.1.4 Extreme gradient boost (XGboost)

XGBoost is an abbreviation for "Extreme Gradient Boosting." XGBoost is a portable, versatile, and efficient decentralised gradient boosting toolbox. It develops machine learning methods using the gradient boosting architecture. It uses parallel tree boosting to solve a wide range of data science problems rapidly and accurately. XGBoost is a modified gradient boosted decision tree (GBM) that improves speed and performance.

- **Regularized Learning:** This helps to smooth the final learned weights in order to minimise over-fitting. The regularised weights will favour models with simple and predictive functions.

- **Gradient Tree Boosting:** Traditional Euclidean space optimization methods cannot be used to optimise the tree ensemble model. Rather, the model is trained in an additive fashion.

- **Shrinkage Subsampling:** To avoid overfitting, two additional tactics are used in addition to the regularised goal. Shrinkage is the first approach introduced by Friedman. After each step of tree boosting, shrinkage scales newly add weights by a factor. Shrinkage minimises each tree's influence while allowing future trees to enhance the model, similar to a learning rate in stochastic optimization.

4.2. SMOTE (Synthetic Minority Over-sampling Technique)

Imbalanced classification entails building prediction models on classification datasets with a high degree of class imbalance. Working with imbalanced datasets presents the difficulty that most machine learning algorithms will overlook, and so perform poorly on, the minority class, despite the fact that performance on the minority class is often the most essential. Oversampling the minority class is one method for dealing with imbalanced datasets. The most basic method includes copying instances from the minority class, even if these examples offer no new data to the model. Instead, new instances can be created by combining old ones. SMOTE, is a type of data augmentation for the minority class. SMOTE generates synthetic data using a k-nearest neighbour method. SMOTE starts by picking random data from the minority class, then determining the k-nearest neighbours. The synthetic data would next be created by combining the random data with the randomly chosen k-nearest neighbour. To improve our unbalanced dataset, we used this strategy.

4.3 Grid search

After using the SMOTE sampling technique to our model, we optimized it using the grid search hyper parameter tuning technique. All the multiple parameters in a machine learning model are not analyzed by the trained data. These parameters govern the model's accuracy. As a result, hyperparameters are very significant in a data science project. The hyperparameters are set up front and supplied by the model's caller before the model is trained. The learning rate of a model, for example, is a hyper parameter since it is set by the caller before the model receives the training data. Grid search is a tuning technique that aims to find the optimal hyper parameter values. It's a comprehensive search of a model's given parameter values. The model is also known as an estimator.

4.4 Metrics used for model evaluation

- Accuracy: The proportion of valid and correct classifications among all values in the dataset

- Recall: The percentage of successfully detected positives in the dataset compared to the total number of positives.
- F1-Score: Precision and Recall's harmonic mean.
- Confusion Matrix: The results of classification issue prediction are summarised in a confusion matrix. Count values are used to sum and divide the number of correct and wrong predictions per class. The confusion matrix displays the correct, false positive, and false negative findings while making predictions.
- Area Under the Curve (AUC): The degree of distinction, or the measure of it, shows how well the model can differentiate between categories. The higher the AUC, the greater the model distinguishes between diseased and unaffected patients in general.

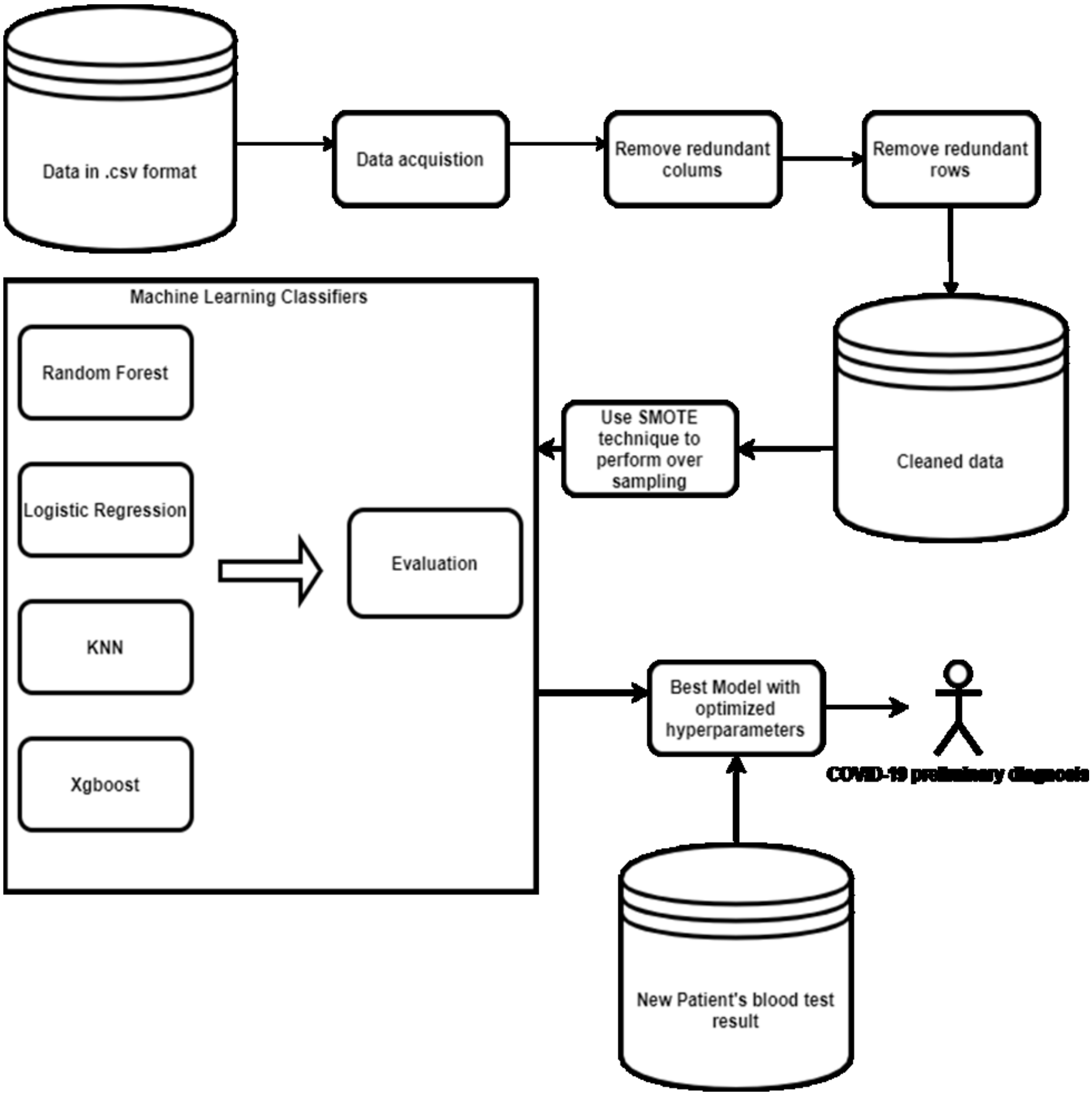

Data in .csv format
Data acquistion
Remove redundant colums
Remove redundant rows
Cleaned data
Use SMOTE technique to perform over sampling
Machine Learning Classifiers
Random Forest
Logistic Regression
KNN
Xgboost
Evaluation
Best Model with optimized hyperparameters
COVID-19 preliminary diagnosis
New Patient's blood test result

Figure 5. Solution pipeline

5. Results

Four different classification machine learning models were designed along with the grid search hyper parameter tuning technique to test different parameters to achieve the best COVID-19 prediction. Python libraries such as pandas, matplotlib, numpy, seaborn and scikit learn were used. The code was written and run on the Jupyter notebook. SMOTE technique was used priorly to oversample the data and achieve optimal results. Table 1 summarizes the overall results.

Random forest obtained optimal results with an accuracy of 92%. The recall, f1-score and AUC were 71%, 83% and 0.80 respectively. The recall was not very high, but it is still acceptable because of the complications present in the dataset. The model is still able to classify correctly (92% accuracy). The confusion matrix obtained is given in figure 6. The best parameters were as follows: ‘n_estimators’:50, ‘min_samples_split’:2, ‘min_samples_leaf’: 1, ‘max_features’: ’ sqrt’, ‘max_depth’: 32.

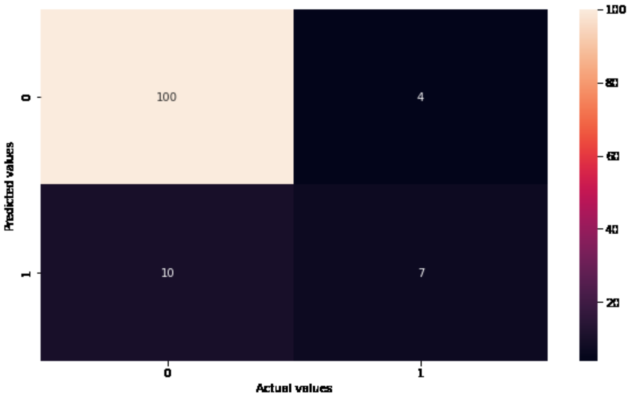


Figure 6. Confusion matrix obtained from random forest

Logistic regression obtained an accuracy of 84%. The recall, f1-score and AUC were 71%, 73% and and 0.78 respectively. The best parameters obtained according to grid search were as follows: ‘penalty’:12, ‘C’: 100. This model was similar to random forest, but slightly less accuracy was obtained. The confusion matrix is given in figure 7.

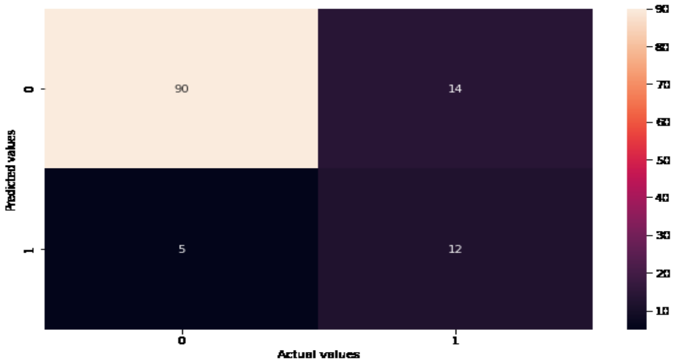


Figure 7. Confusion matrix obtained from logistic regression

KNN obtained an accuracy of 74%. The recall, f1-score and AUC were 59%, 62% and 0.67 respectively. The best parameters were as follows: ‘weights’: ’distance’, ‘p’:1, ‘n_neigbhours’:2 This model was not efficient to classify the COVID-19 patients. The confusion matrix is given in figure 8.

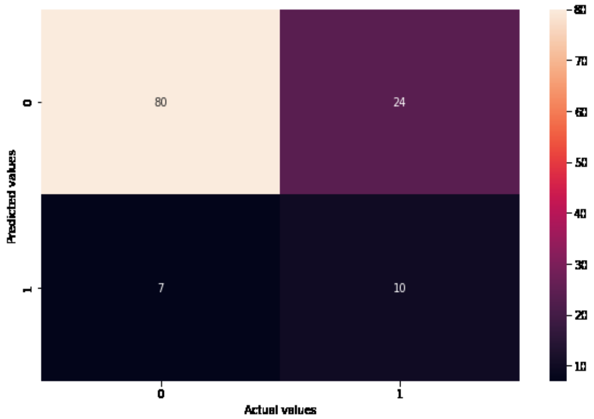


Figure 8. Confusion matrix obtained using KNN

XGBoost model obtained an accuracy of 90%. The recall, f1-score and AUC were 70%, 83% and 0.79. This model was similar to random forest, but not better. ‘n_estimatord’:100, ‘max_depth’:8, ’gamma’ : 0, ‘colsample_bytree’: 0.8 were the best parameters obtained by the grid search technique. The confusion matrix obtained by the model is given in figure 9.

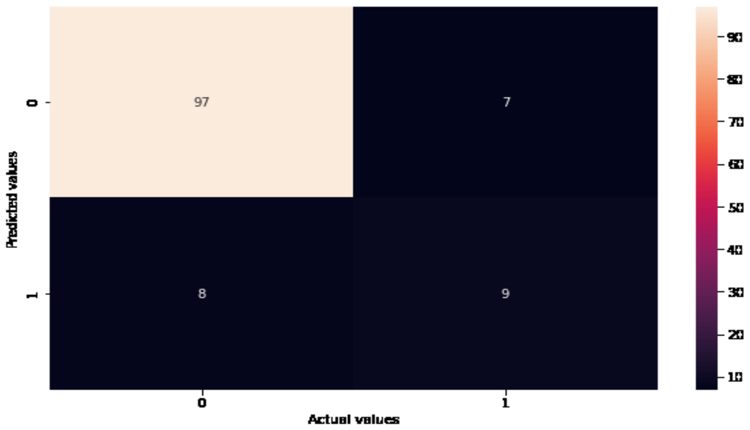


Figure 9. Confusion matrix obtained using XGboost

Table 1: Performance metrics obtained using various ML algorithms

Classifier	Accuracy	Recall	F1--score	AUC
Random Forest	92%	71%	83%	0.80
Logistic Regression	84%	71%	73%	0.78
KNN	74%	59%	62%	0.67
Xgboost	90%	70%	83%	0.79

Though COVID-19 is usually diagnosed using RT-PCR tests, studies have proved that it can also be diagnosed using CT-scans, X-ray images, sound analysis and blood tests [18]. CT-Scans are not easily available everywhere and can also cause unnecessary radiation. X-rays are also prone to false negative results. According to numerous medical studies, the blood and laboratory markers of COVID-19 patients can change dramatically, and these parameters can be used in the preliminary screening for COVID-19. In this study, the value of leukocytes, eosinophils, monocytes and platelets tend to decrease for COVID-

19 patients and monocytes count increase. This paper agrees with the researches that have been already conducted on other datasets that are discussed in table 2. This table gives a comparison of various researches that diagnose and predict COVID-19 from blood tests and machine learning.

Table 2: Comparison of various literatures that diagnose COVID-19 using ML.

Reference	**Dataset Source**	**Total features**	**Model**	**Accuracy**	**AUC**
[12]	279 patients, San Raffaele Hospital (Italy)	Eight features	Seven ML models	86%	84%
[13]	879 patients, NHS Trust Hospital, England	Several features	Multivariate logistic regression, random forest, Extreme gradient boosted trees	-	87%
[14]	5644 patients, Albert Einstein Hospital, Brazil	Several features	A combined ensemble model	99%	95%
[15]	5644 patients, Albert Einstein Hospital, Brazil	12 features	Random forest and extra tree classifiers	98%	97%

6. Conclusion

We investigated different machine learning algorithms to understand the relationship between various blood tests and COVID-19 in this paper. Dataset of patients obtained from Israelta Albert Einstein Hospital, Brazil was used. The SMOTE technique was used to perform oversampling since the dataset was very unbalanced. Random forest, logistic regression, KNN and Xgboost were the four machine learning models used as classifiers. Among all these, random forest achieved the best results. Leukocytes, platelets, eosinophils, mean platelet volume and monocytes were the most important features to diagnose COVID-19.

For future research, collection of a more reliable and balanced dataset can be done. Deep learning models such as ANN can also be used since they provide better results. Parameters like D-Dimer, CRP, LDH and ferritin should be used as parameters since studies have proved that they are very important in COVID-19 diagnosis and prognosis prediction. More clinical studies are also needed to verify if the theoretically positive results are confirmed in real practice. The existing doctors' input would also have an impact on the direction of our research.

References

[1] Wang W, Zhang W, Zhang J, He J, Zhu F. “Distribution of HLA allele frequencies in 82 Chinese individuals with coronavirus disease-2019 (COVID-19)”. *Hla,* 2020 2020 Aug;96(2):194-6.

[2] Singhal T. A review of coronavirus disease-2019 (COVID-19). The indian journal of pediatrics. 2020 Apr;87(4):281-6.

[3] Liu Y, Gayle AA, Wilder-Smith A, Rocklöv J. “The reproductive number of COVID-19 is higher compared to SARS coronavirus”. *Journal of travel medicine.* 2020 Mar 13.

[4] Udugama B, Kadhiresan P, Kozlowski HN, Malekjahani A, Osborne M, Li VY, Chen H, Mubareka S, Gubbay JB, Chan WC. “Diagnosing COVID-19: the disease and tools for Detection”. *ACS nano*. 2020 Mar 30;14(4):3822-35.

[5] Browning L, Colling R, Rakha E, Rajpoot N, Rittscher J, James JA, Salto-Tellez M, Snead DR, Verrill C. Digital pathology and artificial intelligence will be key to supporting clinical and academic cellular pathology through COVID-19 and future crises: the PathLAKE consortium perspective. Journal of clinical pathology. 2021 Jul 1;74(7):443-

[6] Khuzani AZ, Heidari M, Shariati SA. "COVID-Classifier: An automated machine learning model to assist in the diagnosis of COVID-19 infection in chest x-ray images". *Scientific Reports*. 2021 May 10;11(1):1-6.

[7] Rasheed J, Hameed AA, Djeddi C, Jamil A, Al-Turjman F." A machine learning-based framework for diagnosis of COVID-19 from chest X-ray images". *Interdisciplinary Sciences: Computational Life Sciences*. 2021 Mar;13(1):103-17.

[8] Sharma S. "Drawing insights from COVID-19-infected patients using CT scan images and machine learning techniques: a study on 200 patients". *Environmental Science and Pollution Research*. 2020 Oct;27(29):37155-63.

[9] Serte S, Demirel H. "Deep learning for diagnosis of COVID-19 using 3D CT scans". *Computers in biology and medicine.* 2021 May 1;132:104306.

[10] Pahar M, Klopper M, Warren R, Niesler T. COVID-19 "Cough Classification using Machine Learning and Global Smartphone Recordings". *Computers in Biology and Medicine*. 2021 Jun 17:104572.

[11] Faezipour M, Abuzneid A. "Smartphone-based self-testing of COVID-19 using breathing Sounds". *Telemedicine and e-Health.* 2020 Oct 1;26(10):1202-5.

[12] Brinati D, Campagner A, Ferrari D, Locatelli M, Banfi G, Cabitza F. "Detection of COVID-19 infection from routine blood exams with machine learning: a feasibility study". *Journal of medical systems*. 2020 Aug;44(8):1-2.

[13] Heldt FS, Vizcaychipi MP, Peacock S, Cinelli M, McLachlan L, Andreotti F, Jovanović S, Dürichen R, Lipunova N, Fletcher RA, Hancock A. "Early risk assessment for COVID-19 patients from emergency department data using machine learning". *Scientific reports*. 2021 Feb 18;11(1):1-3.

[14] AlJame M, Ahmad I, Imtiaz A, Mohammed A. "Ensemble learning model for diagnosing COVID-19 from routine blood tests". *Informatics in Medicine Unlocked*. 2020 Jan 1;21:100449.

[15] Czako Z, Sebestyen G, Hangan A. COVID-19 Preliminary Patient Filtering based on Regular Blood Tests using Auto-Adaptive Artificial Intelligence Platform. In2020 IEEE 16th International Conference on Intelligent Computer Communication and Processing (ICCP) 2020 Sep 3 (pp. 109-116). IEEE.

[16] Kaggle, 2020, Einstein Data4u, accessed 22th of June, 2021, https://www.kaggle.com/einsteindata4u/covid19/version/4

[17] Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." *Advances in Neural Information Processing Systems.* 2017.

[18] Chadaga K, Prabhu S, Vivekananda BK, Niranjana S, Umakanth S. "Battling COVID-19 using machine learning: A review". *Cogent Engineering*. 2021 Jan 1;8(1):1958666.