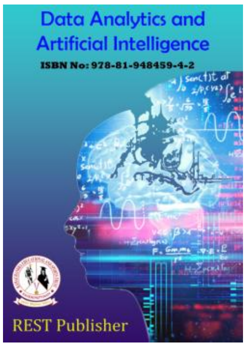
Data Analytics and Artificial Intelligence

Vol: 3(2), 2023

REST Publisher; ISBN: 978-81-948459-4-2

Website: http://restpublisher.com/book-series/daai/

DOI: https://doi.org/10.46632/daai/3/2/13


Data Analytics and Artificial Intelligence
ISBN No: 978-81-948459-4-2
REST Publisher

Covid-19 Patient Health Prediction Using Boosted Random Forest Algorithm

***S. Saranya, S. Bobby**

St. Joseph's College of Arts and Science for Women, Hosur, Tamilnadu, India

*Corresponding Author Email: saranyas0122@gmail.com

Abstract. *COVID-19, also known as 2019-nCoV, is no longer a pandemic but an endemic disease that has killed many people worldwide. COVID-19 has no precise treatment or remedy at this time, but it is unavoidable to live with the disease and its implications. By quickly and efficiently screening for covid, one may determine whether or not one has COVID-19 and thus limit the financial and administrative burdens on healthcare systems. This reality puts a huge demand on these countries' healthcare systems, especially in emerging nations, due to the poor healthcare systems around the world. Although the COVID-19 pandemic cannot be stopped by any licenced vaccine or antiviral medicine, there are other possible solutions that could lighten the burden of the virus on healthcare systems and the economy. The most promising approaches for usage outside of a clinical environment include non-clinical approaches like machine learning, data mining, deep learning, and other artificial intelligence technologies. Artificial intelligence (AI) approaches are increasingly being integrated into wireless infrastructure, real-time data collection, and end-user device processing. A positive and negative COVID-19 case dataset is used to validate artificial intelligence (AI) systems such decision trees, support vector machines, artificial neural networks, and naive Bayesian models. The correlation coefficients between various dependent and independent variables were examined to determine the strength of the relationship between the dependent features. The model was tested 20% of the time while being trained 80% of the time during the preparation phase. The Random Forest had the highest precision (94.99%), according to the evaluation of success.*

Keywords: *COVID-19, healthcare analytics, patient data, infection boosting, random forest classification.*

1. INTRODUCTION

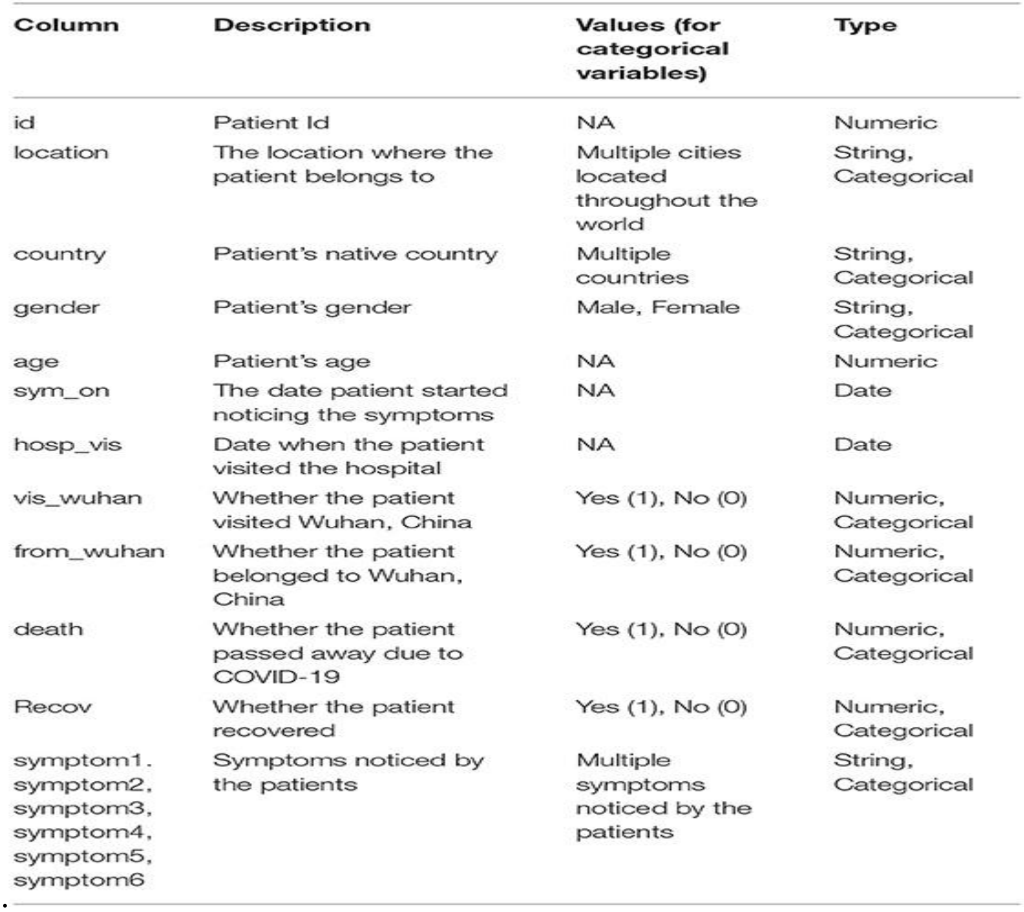
Real-time medical data collection and processing are necessary due to the size and scope of the healthcare sector. Additionally, the challenge of data management, which necessitates real-time prediction and information dissemination to practitioners. For prompt medical intervention, is at the heart of this industry. Major players in this sector, including doctors, vendors, hospitals, and health-related businesses, have made an effort to gather, manage, and resurrect data with the goal of using it to further technology innovation and medical procedures. The world-wide respiratory sickness outbreak that started in Wuhan, China, is being caused by the Corona virus disease 2019 (COVID-19), a member of the family of Corona viruses. According to studies (1-3), Covid-19 exhibits SARS-Covid-like clinical traits. Fever and cough are the main symptoms, although gastrointestinal problems are rare. Since the absence of fever in COVID-19 infected patients is more common than in patients infected by closely related viruses, such as MERS Corona Virus (2%) and SARS Corona Virus (1%)(4), there is a chance that non-febrile patients will be overlooked by a surveillance system with a primary focus on detecting fever.

2. RELATED WORKS

The following packages and libraries are dependencies for the project: Datetime, NumPy, Pandas, SciPy, Scikit Learn, and Matplotlib. The Google Colab platform's CPU runtime was used to implement the project. Model: 79, CPU Family: 6, Model Name: Intel(R) Xeon(R) CPU @ 2.20 GHz, and Cache Size: 56,320 KB are the CPU specs for Google Colab. Google Drive is the storage system. Artificial intelligence systems support vector machines, decision trees,

Data set: The "Novel Corona Virus 2019 Dataset" from Kaggle was utilised to access the dataset for this investigation. The World Health Organization and John Hopkins University are only two of the many sources that went into compiling the information. To suit the requirements of this investigation, we have additionally pre-processed this dataset. This work will enable researchers to further work on developing a solution that combines the processing of patient demographics, travel, and subjective health data with image data (scans) for better prediction of COVID-19 patient health outcomes.

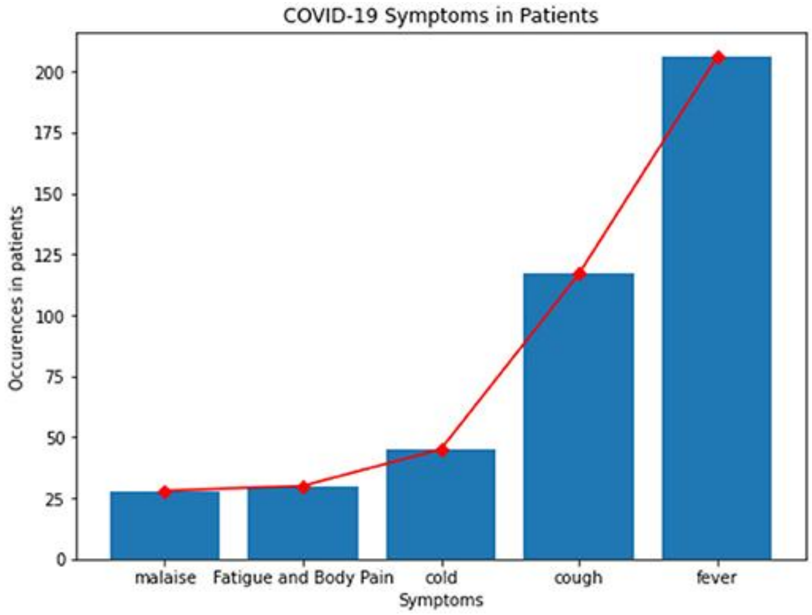
TABLE 1. Characteristics are displayed



Column	Description	Values (for categorical variables)	Type
id	Patient Id	NA	Numeric
location	The location where the patient belongs to	Multiple cities located throughout the world	String, Categorical
country	Patient's native country	Multiple countries	String, Categorical
gender	Patient's gender	Male, Female	String, Categorical
age	Patient's age	NA	Numeric
sym_on	The date patient started noticing the symptoms	NA	Date
hosp_vis	Date when the patient visited the hospital	NA	Date
vis_wuhan	Whether the patient visited Wuhan, China	Yes (1), No (0)	Numeric, Categorical
from_wuhan	Whether the patient belonged to Wuhan, China	Yes (1), No (0)	Numeric, Categorical
death	Whether the patient passed away due to COVID-19	Yes (1), No (0)	Numeric, Categorical
Recov	Whether the patient recovered	Yes (1), No (0)	Numeric, Categorical
symptom1. symptom2, symptom3, symptom4, symptom5, symptom6	Symptoms noticed by the patients	Multiple symptoms noticed by the patients	String, Categorical

3. METHODOLOGY

Data analysis: The most prevalent symptoms that patients whose data is included in this dataset reported experiencing were fever, cough, cold, exhaustion, bodily ache, and malaise,



Data Preprocessing: The data in the columns of the dataset are of the types Date, String, and Numeric. The dataset also includes category variables. We did label-encoding of the categorical variables because the ML model demands that all data supplied as input be in the numeric form. This gives each distinct categorical value in the column a number. The dataset has numerous missing values, which results in an error when they are supplied as input. So, we enter "NA" in place of the missing values. Some patient data entries have been separated from the main dataset and combined into the test dataset because they lack values in both the "death" and "recov" columns. The remaining records have been combined into the main dataset. The disease COVID-19, also known as 2019-nCoV, is no longer a pandemic but an endemic condition that has claimed countless lives all over the world. Living with COVID-19 and its effects is inescapable at this time, but there is no specific treatment or cure. By rapidly and effectively checking for COVID, one can find out if they have COVID-19 and lessen the administrative and cost constraints on healthcare systems. Correlation between features of the dataset provides crucial information about the features and the degree of influence they have over the target value.

Boosted random forest algorithm: The Random Forest classifier algorithm (27)—which in turn is made up of several decision trees—and the boosting method: Ada Boost—combines to form the algorithm known as Boosted Random Forest. A decision tree creates models that resemble real trees. As the algorithm splits our data into smaller subsets, it also grows the tree's branches. The result is a tree with decision nodes and leaf nodes. A decision node has two or more branches that each indicate the value of a tested feature (such as age, symptom1, etc.), and the leaf node has the value of the patient's potential state as a result of the decision (target value).

4. EVALUATION AND FINDINGS

Accuracy: The accuracy of a classifier given a dataset of (TP + TN) data points is equal to the ratio of the classifier's total correct predictions (TP + TN + FP + FN) to the total number of data points. An essential metric for evaluating the effectiveness of the classification model is accuracy. Accuracy is determined by

Precision: The ratio of True Positive (TP) samples to the total of True Positive (TP) and False Positive (FP) samples is known as precision. In an imbalanced class dataset, the precision metric counts the number of patients who were properly categorised.

$$PRECISION = TP/(TP+FP)$$

Recall: The ratio of True Positive (TP) samples to the total of True Positive (TP) and False Negative (FN) samples is known as the recall rate. Recall is a crucial metric to determine the proportion of patients in an imbalanced class dataset that were correctly classified out of all the patients who may have been predicted correctly.

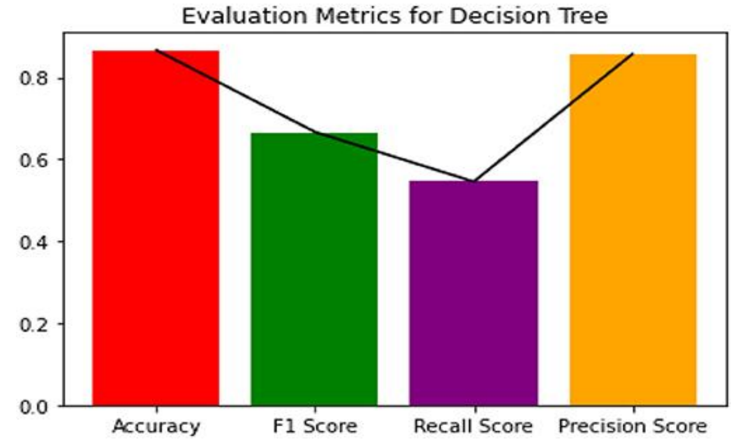
$$RECALL=TP/(TP+FN)$$

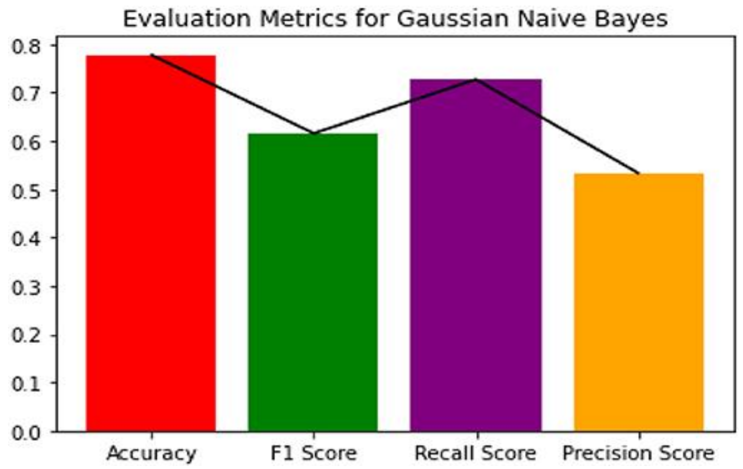
F1 Score: The harmonic mean of the Recall and Precision values is the F1 Score. The F1 Score accurately assesses the model's performance in categorising COVID-19 patients by striking the ideal balance between Precision and Recall. The most important metric we'll use to assess the model is this one.

$$F1\ SCORE= 2 \times PRECISION \times RECALL/ PRECISION + RECALL$$

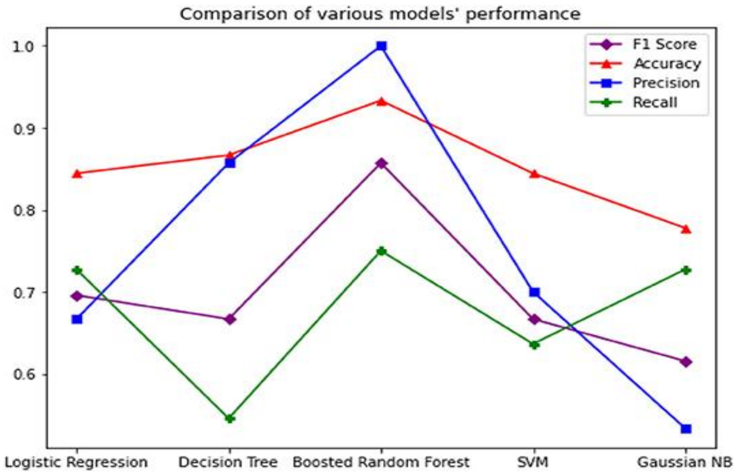
5. RESULT

The pre-processed dataset was utilised to train a number of ML classification models. The decision tree classifier, support vector classifier, gaussian naive bayes classifier, and boosted random forest classifier are among the models used in this work.

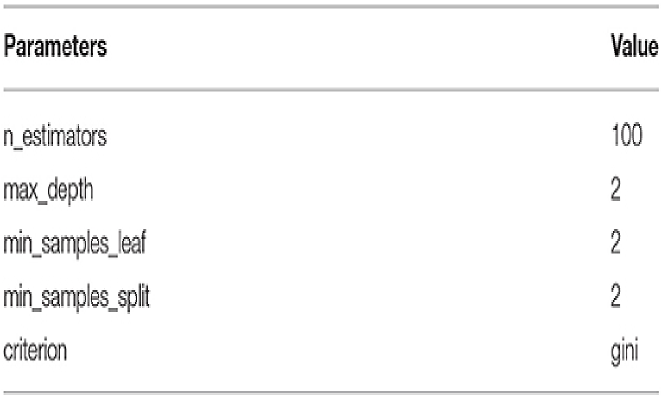




Hyper parameter Optimization: We did a grid search through a grid of selected parameters to identify a set of the best performing parameters because the Boosted Random Forest Classifier was developed using the default settings. Using the GridSearchCV () method from the Sklearn library, we created the grid search.



The study shows that Boosted Random Forest performs better while predicting COVID-19 patient deaths.



Parameters	Value
n_estimators	100
max_depth	2
min_samples_leaf	2
min_samples_split	2
criterion	gini

6. CONCLUSION

Processing patient data for effective treatment plans necessitates the use of artificial intelligence. With an F1 Score of 0.86 on the COVID-19 patient dataset, the model we presented in this research employs the Random Forest approach and is enhanced by the Gadabouts algorithm. Even on unbalanced datasets, the Boosted Random Forest method produces precise predictions, as we have seen. According to the statistics collected for this study, Wuhan locals had greater death rates than non-natives. Additionally, Compared to female patients, male patients had a higher fatality rate. Most of the affected patients range in age from 20 to 70. Future work will concentrate on developing a pipeline that combines these kinds of models for processing demographic and healthcare data with computer vision models for CXR scanning. Then, these models will be included into programmers that will aid in the expansion of mobile healthcare. This could be a step toward the development of a semi-autonomous diagnostic system that would allow for quick screening and diagnosis of COVID-19 affected areas and would also get us ready for potential future outbreaks.

ACKNOWLEDGMENT

I am delighted to express my heartfelt appreciation to our department's head and staff, as well as family and friends. This paper is made possible by their encouragement, assistance, and support.

REFERENCES

[1]. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet. (2020) 395:497–506. Doi: 10.1016/S0140-6736(20)30183-5PubMed Abstract | Cross Ref Full Text |

[2]. . Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet. (2020) 395:497–506. Doi: 10.1016/S0140-6736(20)30183-5PubMed Abstract | Cross Ref Full Text |

[3]. Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet. (2020) 395:507–13. Doi: 10.1016/S0140-6736(20)30211-7PubMed Abstract | Cross Ref Full Text |

[4]. Clinical Management of Severe Acute Respiratory Infection When Novel Coronavirus (2019-nCoV). Infection Is Suspected: Interim Guidance. (2020). Available online at: https://apps.who.int/iris/handle/10665/330893 (accessed April 31, 2020).

[5]. Zumla A, Hui DS, Perlman S. Middle East respiratory syndrome. Lancet. (2015) 386:995–1007. Doi: 10.1016/S0140-6736(15)60454-8

[6]. Pham QV, Nguyen DC, Hwang WJ, Pathirana PN. Artificial intelligence (AI) and big data for coronavirus (COVID-19) pandemic: a survey on the state-of-the-arts. Preprints. (2020) 2020:2020040383. Doi: 10.20944/preprints202004. 0383. V1

[7]. WHO Situation Report-94 Coronavirus disease 2019 (COVID-19). (2020). Available online at: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200423-sitrep-94-covid-19.pdf?sfvrsn=b8304bf0_4 (accessed March 10, 2020).

[8]. Kathiresan S, Sait ARW, Gupta D, Lakshmanaprabu SK, Khanna A, Pandey HM. Automated detection and classification of fundus diabetic retinopathy images using synergic deep learning model. Pattern Recogn Lett. (2020) 133:210–6. Doi: 10.1016/j.patrec.2020.02.026

[9]. Wang L, Wong A. COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest radiography images. ArXiv. (2020) 2003.09871. Available online at: https://arxiv.org/abs/2003.09871 (accessed May 5, 2020).

[10]. Pal R, Sekh AA, Kar S, Prasad DK. Neural network-based country wise risk prediction of COVID-19. ArXiv. (2020) 2004.00959. Available online at: https://arxiv.org/abs/2004.00959 (accessed May 7, 2020).