COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm

Celestine Iwendi[1]*, Ali Kashif Bashir[2], Atharva Peshkar[3], R. Sujatha[4],
Jyotir Moy Chatterjee[5], Swetha Pasupuleti[6], Rishita Mishra[7], Sofia Pillai[8] and Ohyun Jo[9]*

[1] BCC of Central South University of Forestry and Technology, Changsha, China, [2] Department of Computing and Mathematics, Manchester Metropolitan University, Manchester, United Kingdom, [3] Department of Information Technology, G H Raisoni College of Engineering, Nagpur, India, [4] School of Information Technology and Engineering, VIT University, Vellore, India, [5] Department of Information Technology, Lord Buddha Education Foundation, Kathmandu, Nepal, [6] School of Civil Engineering, Galgotias University, Greater Noida, India, [7] Department of Electronics and Telecommunications Engineering, G H Raisoni College of Engineering, Nagpur, India, [8] School of Civil Engineering, Galgotias University, Greater Noida, India, [9] Department of Computer Science, College of Electrical and Computer Engineering, Chungbuk National University, Cheongju-si, South Korea

Integration of artificial intelligence (AI) techniques in wireless infrastructure, real-time collection, and processing of end-user devices is now in high demand. It is now superlative to use AI to detect and predict pandemics of a colossal nature. The Coronavirus disease 2019 (COVID-19) pandemic, which originated in Wuhan China, has had disastrous effects on the global community and has overburdened advanced healthcare systems throughout the world. Globally; over 4,063,525 confirmed cases and 282,244 deaths have been recorded as of 11th May 2020, according to the European Centre for Disease Prevention and Control agency. However, the current rapid and exponential rise in the number of patients has necessitated efficient and quick prediction of the possible outcome of an infected patient for appropriate treatment using AI techniques. This paper proposes a fine-tuned Random Forest model boosted by the AdaBoost algorithm. The model uses the COVID-19 patient's geographical, travel, health, and demographic data to predict the severity of the case and the possible outcome, recovery, or death. The model has an accuracy of 94% and a F1 Score of 0.86 on the dataset used. The data analysis reveals a positive correlation between patients' gender and deaths, and also indicates that the majority of patients are aged between 20 and 70 years.

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

INTRODUCTION

The healthcare industry is a vast industry that requires real time collection and processing of medical data. Moreover, at the core of this industry lies the problem of data handling which requires real time prediction and dissemination of information to practitioners for quick medical attention. Major actors of this industry, such as physicians, vendors, hospitals, and health-based companies have attempted to collect, manage, and revive data with the aim of using it to enhance medical practices and for technological innovation. However, dealing with healthcare data has, of late, become a complex task due to the massive volume of the data, security issues,

wireless network application incompetence, and the velocity at which it is increasing. Thus, to increase the efficiency, accuracy, and workflow healthcare industries need data analytics tools to manage such complex data.

Coronavirus disease 2019 (COVID-19) is a virus of the Corona virus family and the source of a respiratory illness outbreak throughout the world that originated in Wuhan, China. Studies (1–3) show that Covid-19 has clinical characteristics akin to the SARS-CoV. The dominant symptoms include fever and cough, while gastrointestinal symptoms are uncommon. In COVID-19 infected patients the absence of fever is more frequent than in patients infected by similar viruses, i.e., MERS Corona Virus (2%) and SARS Corona Virus (1%) (4); therefore, there is a possibility of non-febrile patients being missed by a surveillance mechanism with a primary focus on detecting fever (5). The initial patients infected by COVID-19, reportedly indicated an association with a large seafood and animal market in Wuhan that demonstrated an animal-to-person spread. Per contra, a burgeoning number of patients have not displayed any association with the animal markets, revealing the fact of human-to-human transmission of COVID-19. This pandemic has been declared a global health emergency and is spreading at an alarming rate (6). The origin of the virus in Wuhan, China has caused 175,694 deaths globally and has 2,544,792 active patients globally (7). With the stress on medical facilities, it is essential for governments and healthcare facilities to identify and treat cases that are most likely to survive, by so doing, judiciously utilizing the limited stock of medical resources and medications.

Artificial Intelligence (AI) has emerged as the breakthrough technology of the twenty-first century and has found multiple applications in fields from weather prediction, astronomical exploration, to autonomous systems (8). We note a few related works where AI has been applied for detection, prevention, and prediction to combat the COVID-19 pandemic. In Wang and Wong (9) researchers have implemented a Convolutional Neural Network based model to detect COVID-19 patients using CXR images. They used a pre-trained ImageNet and trained the model on an open source dataset of Chest X-Ray images (CXR). While Pal et al. (10) implemented a LSTM model to predict the country-specific risk of COVID-19, that relies on trends and weather data of a particular country to predict the probable spread of COVID-19 in that country. In Liu et al. (11) the AI practitioners applied ML to process internet activity, news reports, health organization reports, and media activity to predict the spread of the outbreak on the providence level in China (12). In Bayes and Valdivieso (13) the authors made use of the Bayesian approach to predict the number of deaths in Peru for 70 days in the future, using the empirical data from China. The authors in Beck et al. (14) applied Artificial intelligence to identify the commercially available drugs that could be used to treat COVID-19 patients. They used Bidirectional Encoder Representations from the Transformers (BERT) framework at the core of their model. In Tang et al. (15) the researchers implemented the random forest algorithm for severity analysis of COVID-19 patients using the Computed Tomography (CT) Scans. In Khalifa et al. (16) the authors proposed a Generative Adversarial Network

based fine-tuned model for detecting pneumonia from Chest X-Ray scans, which is one of the symptoms of COVID-19 infection. In Sujatha et al. (17), authors proposed a method which could be helpful in predicting the stretch of COVID-2019, by performing linear regression, and the Multilayer perceptron and Vector autoregression model which could provide an expectation on the COVID-19 Kaggle information, to anticipate the epidemiological pattern of the disease and rate of COVID-2019 cases in India.

Kutia et al. (18) tried to break down client perspectives to eHealth applications in China and the eHealth framework in the Ukraine, which afterwards provided bits of knowledge and proposals for the improvement of an eHealth application (eZdorovya) for mainly health information benefits. Sultan et al. (19) presented a hybrid method that generates and facilitates Alzheimer patients to recall their memories. This egocentric video summary uses important people, objects, and medicines as tools in the realization of their method. Furthermore, an emerging tactile Internet-based nanonetwork that promises a new range of e-health applications has been proposed by Feng et al. (20). The authors use an information based transmit network that goes to an operator via the terahertz band. Finally, the authors in Jain and Chatterjee (21) presented an assortment of strategies intended to speak to, improve, and enable multi-disciplinary and multi-institutional ML to explore in healthcare informatics (22). Khamparia et al. (23) introduced a unique way of an internet of health things (IoHT)-driven deep learning structure for identification and arrangement of cervical cancer in Pap smear pictures, utilizing ideas of transfer learning. Waheed et al. (24) suggested a technique to produce manufactured chest X-ray (CXR) pictures by building up an Auxiliary Classifier Generative Adversarial Network (ACGAN) utilized model called CovidGAN. Sakarkar et al. (25) suggested a profound learning-based mechanized discovery and characterization model for fundus DR pictures.

This paper aims to fill the void of the traditional healthcare system, using machine learning (ML) algorithms to simultaneously process healthcare and travel data along with other parameters of COVID-19 positive patients, in Wuhan, to predict the most likely outcome of a patient, based on their symptoms, travel history, and the delay in reporting the case by identifying patterns from previous patient data. Our contribution includes:

- Processing of healthcare and travel data using machine learning algorithms in place of the traditional healthcare system to identify COVID infected person.
- This work compared multiple algorithms that are available for processing patient data and identified the Boosted Random Forest as the best method for processing data. Further, it executed a grid search to fine-tune the hyper parameters of the Boosted Random Forest algorithm to improve performance.
- Our work obliterates the need to re-compare existing algorithms for processing COVID-19 patient data.

TABLE 1 | Dataset description.

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

- 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.

The rest of the article is organized as follows: section Materials and Methods discusses the materials and methodology used in detail, along with the dataset description, data pre-processing, and the data analysis of the classification algorithms used. Section Results discusses the result of the experiment followed by further discussion in section Discussion. Section Conclusion and Future Work discusses the results and provides a conclusion and the future direction of the current work.

MATERIALS AND METHODS

The dependencies for the project include the following packages and libraries: Datetime, Numpy, Pandas, SciPy, Scikit Learn, and Matplotlib. The project has been implemented on the Google Colab platform using the CPU runtime. The CPU specifications for Google Colab are; model: 79, CPU Family: 6, model name:
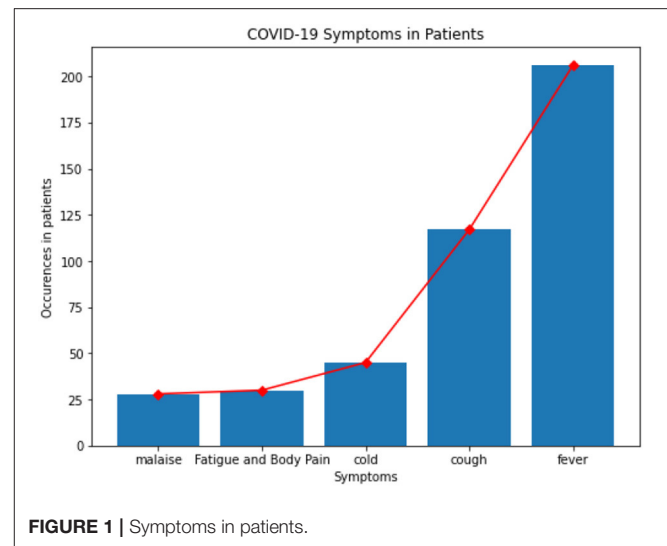
FIGURE 1 | Symptoms in patients.

Intel(R) Xeon(R) CPU @ 2.20 GHz and cache size: 56,320 KB. The storage used is Google Drive.

Dataset

The dataset used in this study was accessed from Kaggle as "Novel Corona Virus 2019 Dataset" (26). The dataset has been compiled from various sources including the World Health Organization and John Hopkins University. However, this dataset has been pre-processed further by us to meet the needs of this study. **Table 1** presents the features of the data.

Data Analysis

Fever, cough, cold, fatigue, body pain, and malaise were the most common symptoms that were noticed in patients whose data is available in this dataset and are shown in **Figure 1**.

Correlation between features of the dataset provides crucial information about the features and the degree of influence they have over the target value. The heat map of Pearson Correlation between the features of the dataset is shown in **Figure 2**, which clearly reveals a relatively stronger positive correlation between age of the patient, whether the patient was native to Wuhan, gap between (in days) when they first felt the symptoms and visited the hospital, and death. However, the country of the patient has a positive correlation with recovery. This implies that foreign patients who visited China had a higher recovery rate. There is also a strong positive correlation between symptom1 and symptom2, and also between symptom2 and symptom3.

Data Pre-processing

The dataset consists of columns with the data being the Date, String, and Numeric type. We also have categorical variables in the dataset. Since the ML model requires all the data that is passed as input to be in the numeric form, we performed label-encoding of the categorical variables. This assigns a number to every unique categorical value in the column.

The dataset consists of multiple missing values which cause an error when passed directly as an input. Thus, we fill the
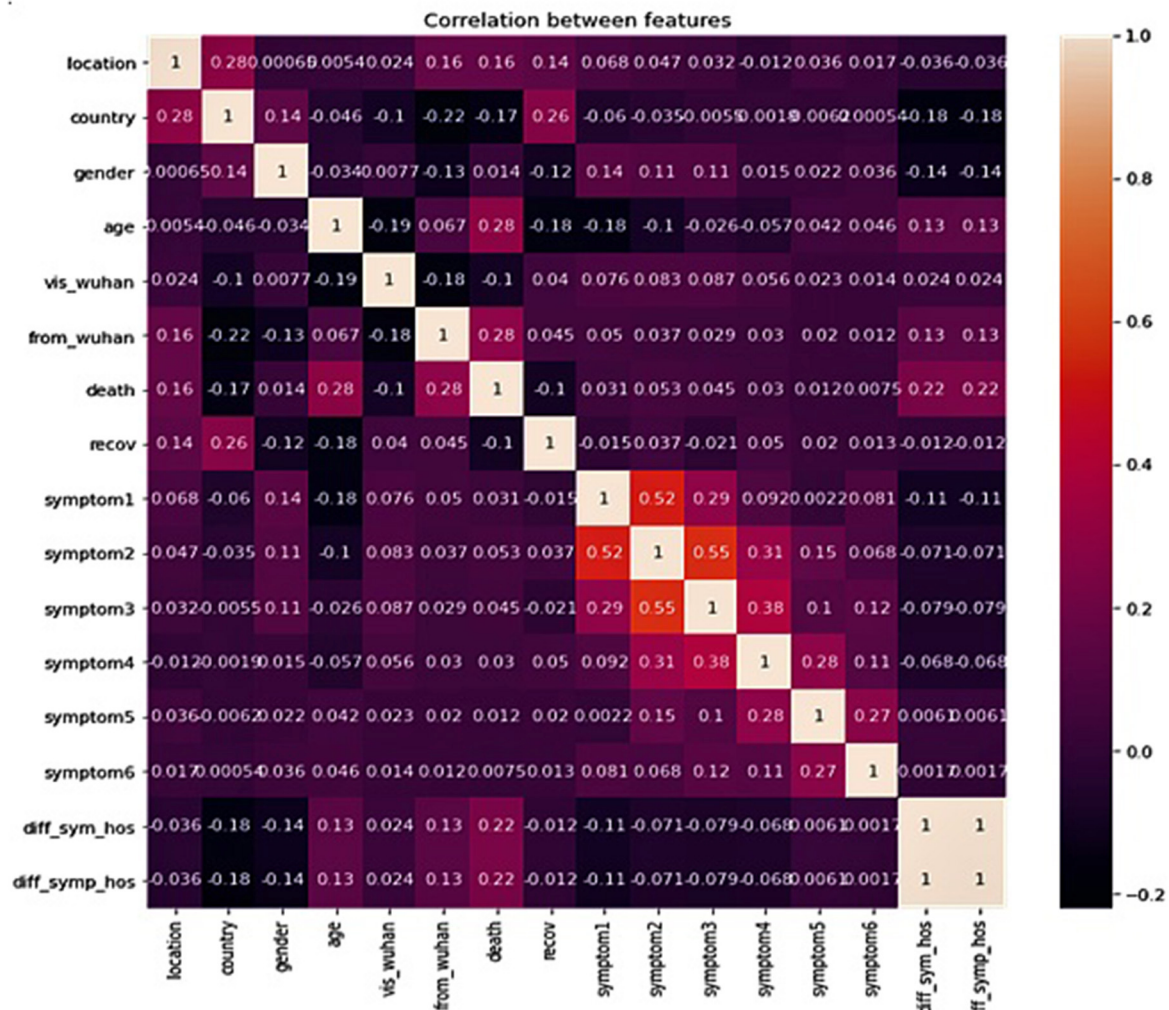
FIGURE 2 | Correlation between data features.

missing values with "NA." Certain patient data records contain missing values for both the "death" and "recov" columns, such patient records have been separated from the main dataset and compiled into the test dataset, while the remaining records have been compiled into the train dataset.

The dataset also consists of columns in the date format. Since the data columns are not directly used, feature engineering has been applied. A new column has been populated with the corresponding (hosp_vis—sym_on) value. This provides us with the number of days that have passed between the symptoms being noticed and the patient visiting the hospital.

Evaluation Metrics

The purpose of the following study is to accurately predict the outcome of a particular patient depending on multiple factors, including but not limited to travel history, demographics etc. Since this is a very crucial prediction, accuracy is very important.
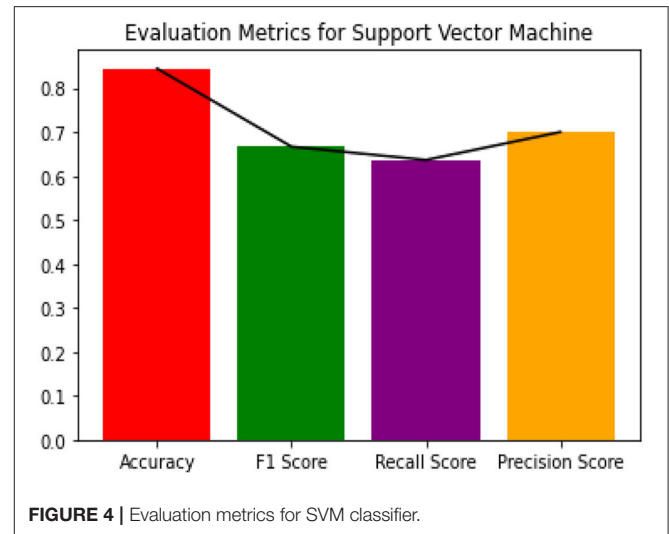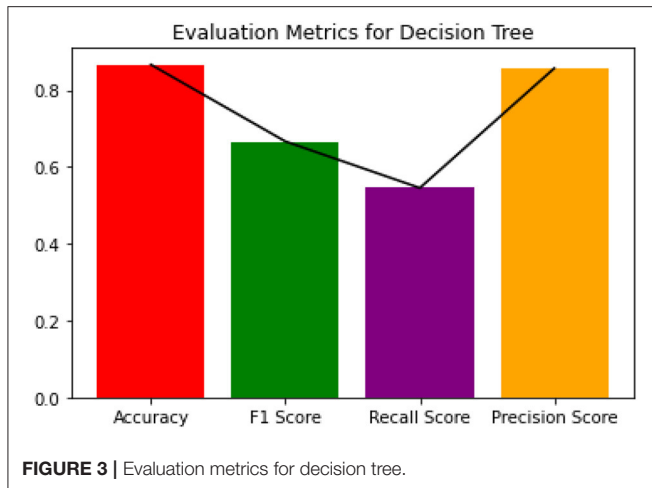
Thus, for the purpose of evaluating the model we considered three evaluation metrics for this study.

The following terms are used in the equations: TP, True Positive; TN, True Negative; FP, False Positive; and FN, False Negative.

Accuracy

Given a dataset consisting of $(TP + TN)$ data points, the accuracy is equal to the ratio of total correct predictions $(TP + TN + FP + FN)$ by the classifier to the total data points. Accuracy is an important measure which is used to assess the performance of the classification model. Accuracy is calculated as shown in Equation (1) as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad 0.0 < \text{Accuracy} < 1.0$$

$$(1)$$

FIGURE 3 | Evaluation metrics for decision tree.

FIGURE 4 | Evaluation metrics for SVM classifier.

FIGURE 5 | Evaluation metrics for Gaussian NB.

Precision

Precision is equal to the ratio of the True Positive (*TP*) samples to the sum of True Positive (*TP*) and False Positive (*FP*) samples. Precision is also a key metric to identify the number of correctly classified patients in an imbalanced class dataset. Precision is calculated as given in Equation (2) as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad (2)$$

Recall

Recall is equal to the ratio of the True Positive (*TP*) samples to the sum of True Positive (*TP*) and False Negative (*FN*) samples. Recall is a significant metric to identify the number of correctly classified patients in an imbalanced class dataset out of all the patients that could have been correctly predicted. Recall is calculated as given in Equation (3) as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad (3)$$

F1 Score

F1 Score is equal to the harmonic mean of Recall and Precision value. The F1 Score strikes the perfect balance between Precision and Recall thereby providing a correct evaluation of the model's performance in classifying COVID-19 patients. This is the most significant measure that we will be using to evaluate the model. F1 Score can be calculated as shown in Equation (4) as follows:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (4)$$

RESULTS

We have used the pre-processed dataset to train multiple ML classification models. The models included in this study include: Decision Tree Classifier, Support Vector Classifier, Gaussian Naïve Bayes Classifier, and Boosted Random Forest Classifier.
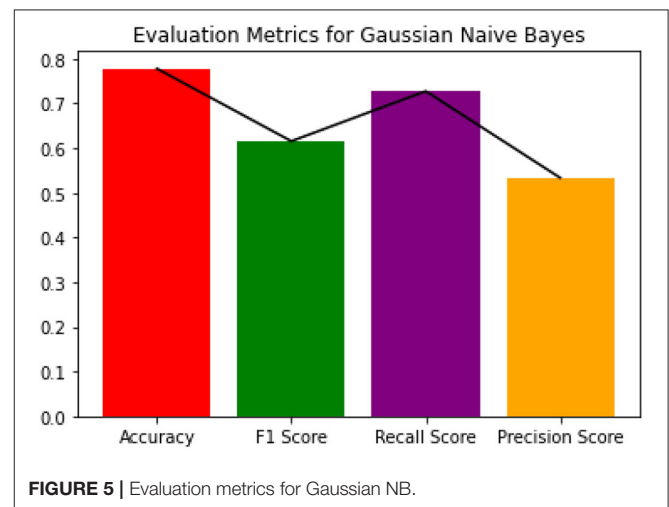
Since the dataset we used can be an imbalanced dataset, we will be using F1 Score as the primary metric for comparison. **Figures 3**–**6** shows the model performances for all the models stated above.

The decision tree constructed for estimating the target variable is visualized in **Figure 7**. The decision tree has a depth of 2 and the Gini index of every node is <0.5, which indicates an imbalance in the training data.

Since Boosted Random Forest algorithm is the best performing model, we will fine tune the model for better performance on the dataset.

DISCUSSION

Boosted Random Forest Classification

A Boosted Random Forest is an algorithm, which consists of two parts; the boosting algorithm: AdaBoost and the Random Forest classifier algorithm (27)—which in turn consists of multiple decision trees. A decision tree builds models that are similar to an

FIGURE 6 | Evaluation metrics for Boosted Random Forest.

FIGURE 7 | Decision tree.

FIGURE 8 | Decision tree 1.
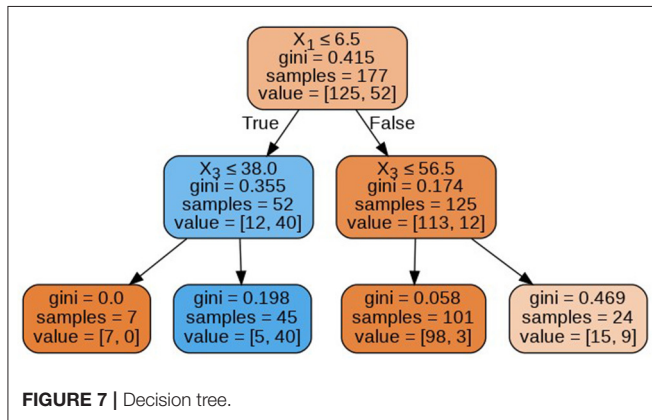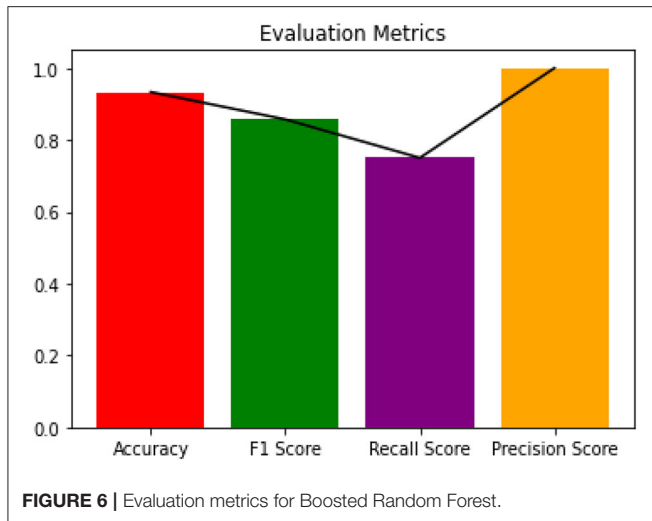
actual tree. The algorithm divides our data into smaller subsets, simultaneously adding branches to the tree. The outcome is a tree consisting of leaf nodes and decision nodes. A decision node has two or more branches representing the value of each feature (like age, symptom1, etc.) tested and the leaf node holds the result value on the patient's prospective condition (target value).

Multiple classifier decision trees (ensemble of classifiers) eliminate the risk of failure of a single decision tree to correctly predict the target value. Thus, the random forest averages the result provided by multiple trees to provide the final result.

The margin function for the random forest is expressed in Equation (5), the generalization error in Equation (6), and confidence in the prediction in Equation (7). Here $h_1(x), h_2(x), \ldots, h_k(x)$ is the ensemble of classifiers (decision trees) and the training data is drawn from the vectors X, Y.

The margin function is expressed as follows:

$$mg(X, Y) = av_k I\left(h_k(X) = Y\right) - max_{j \neq Y} av_k I\left(h_k(X) = j\right) \quad (5)$$

where the indicator function is denoted by $I(.)$. The generalization error is given as follows:

$$PE^* = P_{X,Y}\left(mg(X, Y) < 0\right) \quad (6)$$
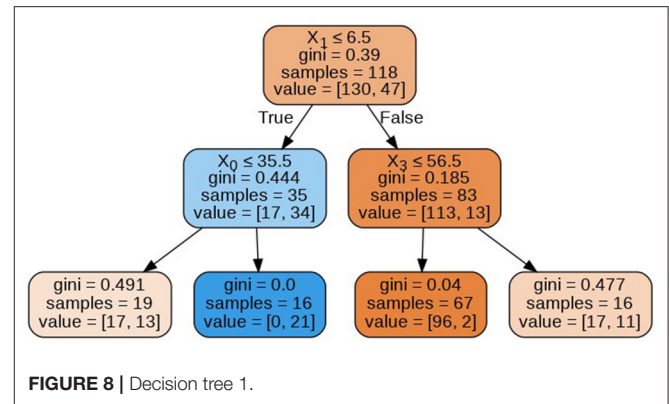
where the probability is expressed over the X, Y space. In random forests, we have $h_k(X) = h(X, \Theta_k)$, therefore the number of classifiers (decision trees) increases, for all the sequences of trees. The probability PE^* converges to Equation (7), from the Strong Law of Large Numbers and tree structure.

$$P_{X,Y}\left(P_\Theta\left(h(X, \Theta) = Y\right) - max_{j \neq Y} P_\Theta\left(h(X, \Theta) = j\right) < 0\right) \quad (7)$$

Applying the boosting algorithm AdaBoost (28) provides a corrective mechanism to improve the model after every prediction of patient state. Eventually, the decision is a result of summing up of all the base models. It is one of the most efficient techniques in ML.

The corrective mechanism can be expressed as follows Equation (8). Given $(x_1, y_1), \ldots, (x_m, y_m)$, where $x_i \in X$, $y_i \in Y = \{-1, +1\}$. For, $t = 1, \ldots, T$. Initialize $D_1(i) = \frac{1}{m}$. After training a weak learner, random forest in our case, using distribution D_t.

Get the hypothesis, $h_t : X \to \{-1, +1\}$,

With the error $e_t = P_{r_i \sim D_t}\left[h_t(x_i) \neq y_i\right]$

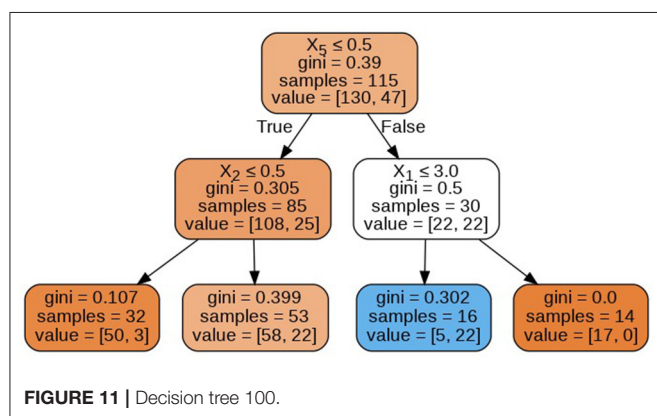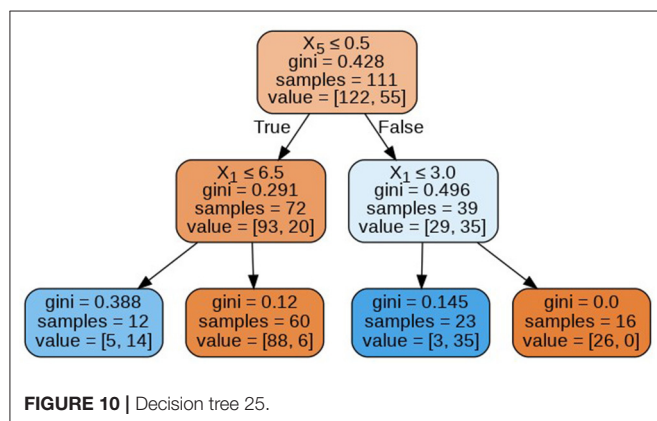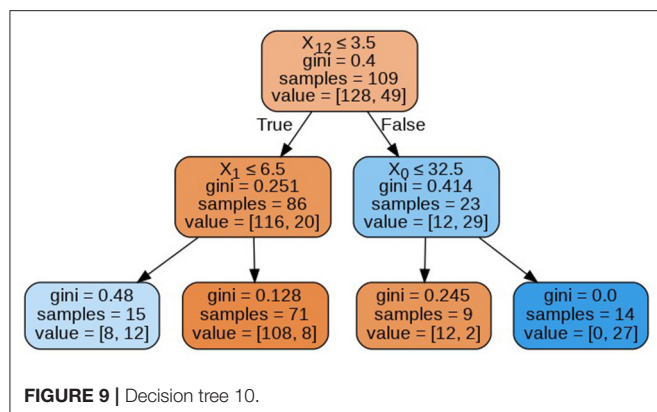After choosing $\alpha_t = \frac{1}{2} ln\, ln\left(\frac{1-e_t}{e_t}\right)$

Update: $D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \{e^{-\alpha t}\ \ if\ \ h_t(x_i) = y_i\ \ e^{\alpha t}\ if\ h_t(x_i) \neq y_i$

$= \frac{D_t(i)\ exp(-\alpha_t y_i h_t(x_i))}{Z_t}$

Here, Z_t is a normalization factor. We get the final hypothesis as follows:

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right) \quad (8)$$

Here the dependent variable was the patient state (dead/recovered) while the independent variables were location, country, vis_wuhan, from_wuhan (hosp_vis—sym_on), age, gender, symptom (1–6). We have used the boosted random forest because of its accurate classification performance on imbalanced datasets (25, 29).

The decision trees visualized in **Figures 8–11** have a depth equal to two. Also, the Gini index in all the leaf nodes of all the trees is <0.5, which indicates the training dataset is imbalanced. Hence, for optimizing the performance of the model

FIGURE 9 | Decision tree 10.

FIGURE 10 | Decision tree 25.

FIGURE 11 | Decision tree 100.

TABLE 2 | Optimal hyperparameters returned by grid search.

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

TABLE 3 | Evaluation results.

Metric	Score
Recall score	0.75
Precision score	1.0
F1 score	0.86
Accuracy	0.94

FIGURE 12 | Comparison of Models' performance.

from Sklearn library. **Table 2** presents the hyperparameters as returned by the grid search algorithm (30).

Table 3 presents the evaluation metrics of the Fine Tuned Boosted Random Forest.

The study shows that Boosted Random Forest performs better while predicting COVID-19 patient deaths. **Figure 12** graph compares the performance of all the models including Boosted Random Forest.

CONCLUSION AND FUTURE WORK

The application of Artificial Intelligence is very crucial to process patient data for efficient treatment strategies. In this paper we presented a model that implements the Random Forest algorithm boosted by the AdaBoost algorithm, with a F1 Score of 0.86 on the COVID-19 patient dataset. We have discovered that the Boosted Random Forest algorithm provides accurate predictions even on imbalanced datasets. The data analyzed in this study has revealed that death rates were higher amongst the Wuhan natives

we have reduced the depth of trees to 2 and increased the number of estimators (decision trees) in the random forest to 100. This prevents high variance in the model and provides accurate predictions.
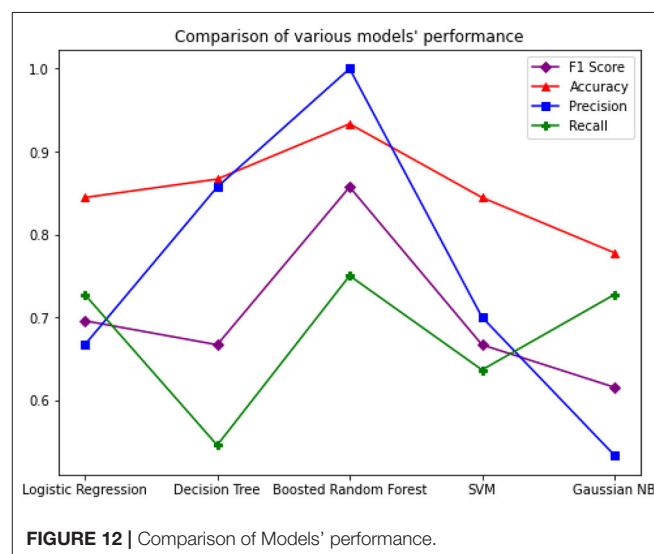
Hyperparameter Optimization

Since the Boosted Random Forest Classifier was implemented using the default parameters, for the optimal performance of the model, we conducted a grid search over a grid of chosen parameters to gain a set of best performing parameters. We implemented the grid search using the GridSearchCV() function

compared to non-natives. Also, male patients had a greater death rate compared to female patients. The majority of affected patients are aged between of 20 and 70 years.

Future work will focus on creating a pipeline that combines CXR scanning computer vision models with these types of demographic and healthcare data processing models. These models will then be integrated into applications that will support the growth of mobile healthcare. This can provide a step toward a semi-autonomous diagnostic system that can provide rapid screening and detection for COVID-19 affected regions and prepare us for future outbreaks.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. For the reproducible code, please check out the GitHub repository: https://github.com/Atharva-Peshkar/Covid-19-Patient-Health-Analytics.

AUTHOR CONTRIBUTIONS

AP, CI, and RM: conceptualization. AP and RM: methodology, investigation, data curation, and writing—original draft preparation. AP, RM, SP, OJ, and NP: software. RS and JC: validation and visualization. CI, RS, and JC: formal analysis. AP, AB, and RM: resources. JC and CI: writing—review and editing, supervision. AB, AP, RM, SP, NP, RS, CI, OJ, and JC: project administration. All authors have read and agreed to the published version of the manuscript.

FUNDING

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)30183-5
2. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. *N Engl J Med.* (2020) 382:1199–207. doi: 10.1056/NEJMoa2001316
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)30211-7
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).
11. Liu D, Clemente L, Poirier C, Ding X, Chinazzi M, Davis JT, et al. A machine learning methodology for real-time forecasting of the 2019–2020 COVID-19 outbreak using Internet searches, news alerts, and estimates from mechanistic models. *arXiv.* (2020) 2004.04019. Available online at: https://arxiv.org/abs/2004.04019 (accessed May 6, 2020).

12. Cai H. Sex difference and smoking predisposition in patients with COVID-19. *Lancet Respir Med.* (2020) 8:e20. doi: 10.1016/S2213-2600(20)30117-X
13. Bayes C, Valdivieso L. Modelling death rates due to COVID-19: a Bayesian approach. *arXiv.* (2020) 2004.02386. Available online at: https://arxiv.org/abs/2004.02386 (accessed May 5, 2020).
14. Beck BR, Shin B, Choi Y, Park S, Kang K. Predicting commercially available antiviral drugs that may act on the novel coronavirus (2019-nCoV), Wuhan, China through a drug-target interaction deep learning model. *bioRxiv.* (2020). Available online at: https://www.biorxiv.org/content/10.1101/2020.01.31.929547v1.abstract (accessed May 5, 2020).
15. Tang Z, Zhao W, Xie X, Zhong Z, Shi F, Liu J, et al. Severity assessment of coronavirus disease 2019 (COVID-19) using quantitative features from chest CT images. *arXiv.* (2020) 2003.11988. Available online at: https://arxiv.org/abs/2003.11988 (accessed May 10, 2020).
16. Khalifa NEM, Taha MHN, Hassanien AE, Elghamrawy S. Detection of coronavirus (COVID-19) associated pneumonia based on generative adversarial networks and a fine-tuned deep transfer learning model using chest X-ray dataset. *arXiv.* (2020) 2004.01184. Available online at: https://arxiv.org/abs/2004.01184 (accessed May 5, 2020).
17. Sujatha R, Chatterjee JM, Hassanien AE. A machine learning forecasting model for COVID-19 pandemic in India. *Stoch Environ Res Risk Assess.* (2020) 34:959–72. doi: 10.1007/s00477-020-01827-8
18. Kutia S, Chauhdary SH, Iwendi C, Liu L, Yong W, Bashir AK. Socio-Technological factors affecting user's adoption of eHealth functionalities: a case study of China and Ukraine eHealth systems. *IEEE Access.* (2019) 7:90777–88. doi: 10.1109/ACCESS.2019.2924584
19. Sultan S, Javed A, Irtaza A, Dawood H, Dawood H, Bashir AK. A hybrid egocentric video summarization method to improve the healthcare for Alzheimer patients. *J Ambient Intell Human Comput.* (2019) 10:4197–206. doi: 10.1007/s12652-019-01444-6
20. Feng L, Ali A, Iqbal M, Bashir AK, Hussain SA, Pack S. Optimal haptic communications over nanonetworks for E-health systems. *IEEE Trans Ind Inform.* (2019) 15:3016–27. doi: 10.1109/TII.2019.2902604
21. Jain V, Chatterjee JM. *Machine Learning with Health Care Perspective.* (2020). Available online at: https://link.springer.com/book/10.1007%2F978-3-030-40850-3 (accessed May 5, 2020).
22. Chatterjee JM. Bioinformatics using machine learning. *Glob J Internet Interv IT Fusion.* (2018) 1:28–35.
23. Khamparia A, Gupta D, de Albuquerque VHC, Sangaiah AK, Jhaveri RH. Internet of health things-driven deep learning system for detection and classification of cervical cells using transfer learning. *J Supercomput.* (2020) 76:1–19. doi: 10.1007/s11227-020-03159-4

24. Waheed A, Goyal M, Gupta D, Khanna A, Al-Turjman F, Pinheiro PR. Covidgan: data augmentation using auxiliary classifier gan for improved covid-19 detection. *IEEE Access.* (2020) 8:91916–23. doi: 10.1109/ACCESS.2020. 2994762

25. Sakarkar G, Pillai S, Rao CV, Peshkar A, Malewar S. Comparative study of ambient air quality prediction system using machine learning to predict air quality in smart city. In *Proceedings of International Conference on IoT Inclusive Life (ICIIL 2019), NITTTR Chandigarh, India.* Singapore: Springer (2020). p. 175–82. doi: 10.1007/978-981-15-3020-3_16

26. *Novel Corona Virus 2019 Dataset.* (2020). Available online at: https://www. kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset/ (accessed April 23, 2020).

27. Breiman L. Random forests. *Mach Learn.* (2001) 45:5–32. doi: 10.1023/A:1010933404324

28. Freund Y, Schapire R, Abe N. A short introduction to boosting. *J Jpn Soc Artif Intell.* (1999) 14:1612.

29. Khalilia M, Chakraborty S, Popescu M. Predicting disease risks from highly imbalanced data using random forest. *BMC Med Inform Decis Mak.* (2011) 11:51. doi: 10.1186/1472-6947- 11-51

30. Pillai SK, Raghuwanshi MM, Gaikwad M. Hyperparameter tuning and optimization in machine learning for species identification system. In: *Proceedings of International Conference on IoT Inclusive Life (ICIIL 2019), NITTTR Chandigarh, India.* Singapore: Springer (2020). p. 235–41. doi: 10.1007/978-981-15-3020-3_22