PREDICTIVE HEALTH MODELING FOR COVID-19 PATIENTS USING ENHANCED RANDOM FOREST ENSEMBLES

Ravikant Kholwal
PDPM IIITDMJ, Jabalpur, India

ABSTRACT

The integration of artificial intelligence (AI) into wireless infrastructure and real-time data processing on end-user devices is pivotal in today's technologically advanced era, especially for detecting and predicting large-scale pandemics. The unprecedented outbreak of COVID-19, originating in Wuhan, China, has underscored the urgency of leveraging advanced technologies to combat the widespread consequences of such pandemics. The virus has had a profound impact globally, pushing even the most advanced healthcare systems to their limits. As of 11 May 2020, the European Centre for Disease Prevention and Control agency reported over 4,063,525 confirmed cases and 282,244 deaths globally. The rapid surge in cases accentuates the need for swift and efficient utilization of AI techniques to predict the trajectories and outcomes of infected individuals, enabling the allocation of appropriate treatments. In response to this, the research proposed a refined Random Forest model, augmented with the AdaBoost algorithm, as a potential solution to predict the severity and outcomes of COVID-19 cases. The model leverages a diverse range of patient data, encompassing geographical, travel, health, and demographic details, to make predictions about the severity of cases and the likelihood of recovery or death. The model demonstrated a high level of accuracy, achieving a 94% accuracy rate on the utilized dataset, with an F1 Score of 0.86. The analysis of the data revealed a notable correlation between gender and mortality rates and indicated that the majority of patients were within the 20 to 70 years age range. This research underscores the potential of AI in enhancing our ability to respond to pandemics by providing insights into the severity and possible outcomes of individual cases based on a variety of factors. The integration of such advanced technologies into our healthcare and data collection systems is crucial for developing more effective strategies for managing and mitigating the impacts of future pandemics.

KEYWORDS: *COVID-19, Healthcare analysis, Random Forest, Boosting*

I. INTRODUCTION

Healthcare is an extensive and complex field that depends on the real-time collection and processing of medical data. Proper data handling practices and timely dissemination to practitioners are integral for providing quick and effective medical assistance. As part of their efforts to improve medical practices and drive technological innovation, various stakeholders in the industry, such as physicians, vendors, hospitals and health-based companies, have taken measures to collect, organize and use data efficiently in order to increase healthcare practices. Healthcare data management has become increasingly challenging due to its sheer volume, security concerns, limitations of wireless network applications and rapid growth. To address these challenges and optimise efficiency, accuracy, and workflow in healthcare industries worldwide, advanced data analytics tools are required in order to handle and analyse complex healthcare data sets efficiently.

II. RELATED WORKS

In one such study [1], the authors developed a method based on Support Vector Machine (SVM) that leverages patient X-ray data. This method was designed to distinguish normal lung images from those exhibiting indications of COVID-19 infection. The results were promising, with the method exhibiting 95.76% sensitivity, 99.7% specificity, and an overall accuracy rate of 97.28%. This high level of efficiency underscores the potential of this approach as a diagnostic tool for COVID-19.

Decision Tree algorithms have also been recognized for their applicability in medical settings. A specific instance [2] involved the creation of a model-based decision tree aimed at assessing the severity of COVID-19 in pediatric patients. The model, which was developed using data from 105 infected children, showcased promising performance in accurately determining the severity of the infection in this demographic.

Too et al. [3] introduced a distinctive approach that incorporated the Hyper Learning Binary Dragonfly Algorithm for feature selection in COVID-19 patients. This method focused on utilizing selected features to predict the conditions of the patients with high precision.

Another study [4] employed time-dependent parameters to model the dynamic propagation of COVID-19. The authors developed a nonlinear approximation of the virus's prevalence using an epidemiological model, providing insights into the evolving nature of the pandemic.

In reference [5], an enhanced fuzzy clustering algorithm was utilized to devise a novel time series forecasting method. This method aimed at estimating the number of COVID-19 patients and related deaths in India. The results of this approach were superior in terms of mean square error, root mean square error, and average forecasting error rate compared to previous methods.

A comprehensive investigation [6] into the application of artificial intelligence-based techniques for predicting COVID-19 positivity and severity revealed that various algorithms, including K nearest neighbor classifier, Neural Networks, Decision Trees, and Partial Least Squares Discriminant Analysis, could yield satisfactory accuracy levels in diagnosing the severity of COVID-19.

The prediction of future intubation requirements among COVID-19 positive cases was explored in [7]. In this study, a machine learning-based model was developed to estimate the probability of intubation, utilizing historical patient information to make projections.

Pahar et al. [8] utilized smartphone audio recordings to classify COVID-19 cough with high accuracy using various machine learning models, specifically comparing several residual neural network classification models before showing that one, in particular, can successfully discriminate between healthy and unhealthy coughs with great accuracy.

Length of stay prediction models has been proposed in [9] to forecast the likelihood of prolonged hospital stays among COVID-19 patients using electronic health record data.

In the realm of research focused on COVID-19, several studies have delved into various aspects of the disease, employing diverse methodologies and analytical techniques. Zhang et al. [10] embarked on a comprehensive exploration of the clinical characteristics and eventual outcomes in distinct cases of COVID-19 that tested positive. The study meticulously identified nine factors associated with mortality, utilizing the least absolute shrinkage and selection operator (LASSO) regression, a method known for its efficiency in handling high dimensional data. These identified factors were subsequently subjected to testing through an artificial neural network algorithm, a form of machine learning designed to recognize patterns.

Another study [11] directed its attention towards sentiment analysis of tweets related to COVID-19. This research employed classical machine learning classification models to scrutinize the sentiments expressed in the tweets. Several models, including SVC, Perceptron, Passive Aggressive Classifier, and Logistic Regression, were utilized, with some achieving prediction rates surpassing the 98% mark, showcasing the efficacy of machine learning in analyzing public sentiment during pandemics.

Singh et al. [12] delved into the exploration of transfer learning techniques specifically for the intelligent screening of COVID-19 through CT images. The study employed models like VGG16 for feature extraction and Principal Component Analysis (PCA) for feature selection, both crucial steps in machine learning model development. The research evaluated various classification models, with bagging ensemble and SVM emerging as the ones demonstrating the highest prediction accuracy, highlighting the potential of transfer learning in medical image analysis.

Lastly, a study [13] introduced a Joint Classification and Segmentation (JCS) model aimed at providing real-time and comprehensible diagnoses of COVID-19 utilizing chest CT images. This model stands as

a testament to the advancements in medical imaging and diagnosis, offering real-time solutions that are crucial in managing and controlling the spread of infectious diseases like COVID-19.

The myriad of studies highlighted herein illustrates the multifaceted approaches and innovative methodologies employed in leveraging artificial intelligence and machine learning to combat COVID-19. These studies have explored diverse aspects, ranging from diagnostic tools utilizing patient X-ray data, decision tree algorithms assessing the severity of infections in pediatric patients, to sophisticated models predicting the dynamic propagation and future intubation requirements of COVID-19 cases. The incorporation of various algorithms, feature selection techniques, and analytical models has not only demonstrated promising results in terms of accuracy and precision but has also provided invaluable insights into the evolving nature of the pandemic. The advancements in medical imaging, sentiment analysis of public reactions, and the exploration of transfer learning techniques underscore the boundless potential of these technologies in enhancing our understanding and management of infectious diseases. These comprehensive investigations into the applications of artificial intelligence-based techniques reveal the pivotal role they play in predicting and diagnosing the severity of COVID-19, ultimately contributing to the global efforts in mitigating the impacts of this unprecedented health crisis.

III. METHODOLOGY

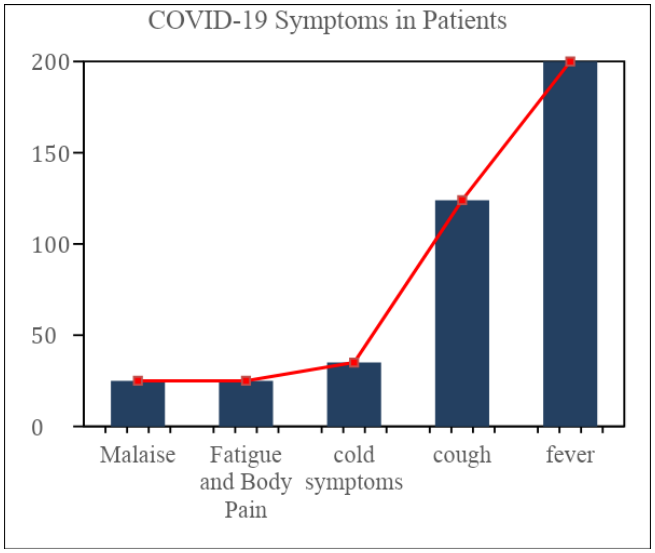

Figure 1 Symptoms in patients.

3.1. Dataset

The dataset employed in this research came from Kaggle under the name "Novel Corona Virus 2019 Dataset" (26). It comprises data compiled from multiple reliable sources like World Health Organization and John Hopkins University; to meet our specific research requirements, we pre-processed it further to meet them; please see Table 1 for details about this dataset's features.

3.2. Data Analysis

This research examined a dataset that includes information on different symptoms observed among patients. Fever, cough, cold, fatigue, body pain, and malaise emerged as some of the most frequent indicators, as illustrated by Figure 1.

Table 1: Dataset description

Column	Description	Values (for categorical variables)	Type
Id	Patient id	NA	Numeric
Location	The location where the patient belong 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 data patient started noticing the symptoms	NA	Date
hosp_vis	Data when the patient visited the hospital	NA	Date
vis_wuhan	Whether the patient 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
Symptom 1, Symptom 2, Symptom 3, Symptom 4, Symptom 5, Symptom 6	Symptoms noticed by the patients	Multiple symptoms noticed by the patients	String, Categorical

3.3. Data-Preprocessing

In this study, the dataset employed contained columns with different data types - Date, String and Numeric as well as categorical variables. To implement machine learning models effectively, all input data must be converted to numeric form; we achieved this through label encoding categorical variables within their columns by assigning unique numbers for every distinct categorical value within that column.

Our dataset also contained missing values that caused issues when used as input to models. To address this, we filled these spaces with the label "NA". Additionally, some patient records with missing values for both "death" and "recover" columns were extracted from the main dataset and combined into a test dataset, while all remaining records became a training dataset.

As the dataset includes columns in date format, these were not directly utilized; feature engineering techniques were applied instead to them in order to capture a difference between "hosp_vis" (hospital visit) and "sym_on" (symptom onset) values, representing how many days have elapsed since symptom

onset and patient hospital visit date; providing valuable data for analysis and modeling purposes. This engineered feature provided invaluable information for analysis and modeling purposes.

IV. IMPLEMENTATION

The objective of this study is to formulate precise predictions pertaining to individual patients, considering a multitude of factors including travel history and demographics. The emphasis is placed on achieving the highest level of accuracy in these predictions to ensure reliability and validity in the outcomes.

In order to meticulously evaluate the efficacy of the predictive model developed within this study, a set of evaluation metrics were deployed, namely True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). These metrics serve as pivotal indicators in assessing the model's proficiency in forecasting patient outcomes accurately. True Positives and True Negatives represent the instances where the model's predictions were correct, identifying the presence and absence of a condition accurately, respectively. Conversely, False Positives and False Negatives depict the instances where the model's predictions were incorrect, either identifying a condition where there isn't one or failing to identify a condition that is present. By analyzing these metrics, a comprehensive understanding of the model's predictive capabilities and areas for improvement can be attained, ensuring the refinement and optimization of the model for enhanced reliability in real-world applications.

4.1. Accuracy

The Accuracy serves as an important indicator of classification model performance on any dataset, calculated by dividing the total number of correct predictions (TP + TN) made by the classifier by the total number of data points (TP + TN + FP + FN). Accuracy serves as an excellent measure for gauging effectiveness; see Equation (1) to calculate it.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad 0.0 < Accuracy < 1.0 \qquad (1)$$

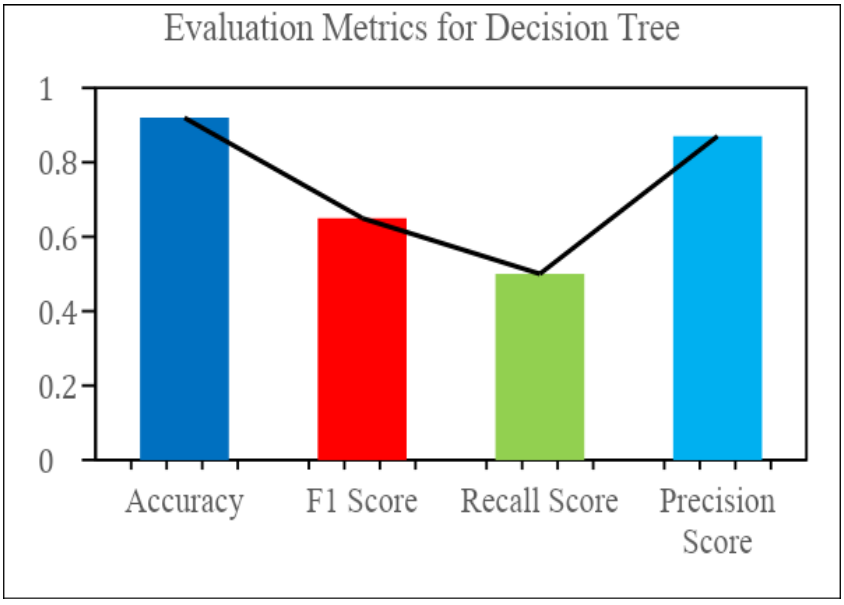

FIGURE 2 Evaluation metrics for decision tree.

4.2. Precision

Precision is an essential measure to assess a classification model's ability to correctly identify positive samples. It is determined by dividing the number of true positive (TP) samples by the sum of true positive and false positive samples (TP+FP)[14], precisely serving as an invaluable gauge for performance evaluation in datasets with imbalanced classes; Equation (2) provides the formula.

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

4.3. Recall

Recall, commonly referred to as sensitivity or true positive rate, is an integral metric used to gauge a model's ability to correctly identify positive samples out of all those which should have been identified as such. A recall is calculated by dividing the total number of true positive (TP) and false negative (FN) samples by 1. Its use becomes increasingly relevant when working with imbalanced class datasets to assess performance in capturing all relevant positive instances; Equation (3) shows how you can calculate it.

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

4.4. F1 score

The F1 score, often referred to as F-score or F-measure, is a measure that encompasses precision and recall to give a comprehensive assessment of a model's performance in classifying COVID-19 patients. It represents the harmonic mean between these values, thus balancing out the tradeoff between these metrics. It has become one of the primary metrics used for measuring model effectiveness; Equation (4) illustrates this calculation which accounts for both precision and recall when making its calculations.

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

V. RESULTS

The dataset, after undergoing meticulous pre-processing, was employed to train a variety of machine learning classification models, encompassing Decision Tree Classifier, Support Vector Classifier, Gaussian Naive Bayes Classifier, and Boosted Random Forest Classifier models. The dataset's imbalanced nature necessitated the use of the F1 score as the principal metric for contrasting the models, and Figures 2-5 illustrate the performance of each classification model in this context. The decision tree, illustrated in Figure 6, is structured to approximate the target variable and is configured with a depth constrained to two levels. Every node within this tree manifests a Gini index below 0.5, underscoring the imbalance inherent in the training dataset. In a similar vein, the decision trees delineated in Figures 7 through 11 also adhere to a two-level depth and consistently exhibit a Gini index below 0.5 at each of their leaf nodes. To augment the performance of the model, the depth of the trees was deliberately restricted to two, and the quantity of decision tree estimators incorporated within the random forest was increased to 100. This approach is instrumental in curbing excessive variance and is conducive to generating predictions with heightened precision. By constraining the depth of the decision trees, the model is prevented from becoming overly complex, reducing the risk of overfitting to the training data and enhancing its generalizability to unseen data. Concurrently, by amplifying the number of decision tree estimators within the random forest, the model gains in robustness and accuracy, as it aggregates the predictions from multiple trees, mitigating the impact of individual tree biases and errors.

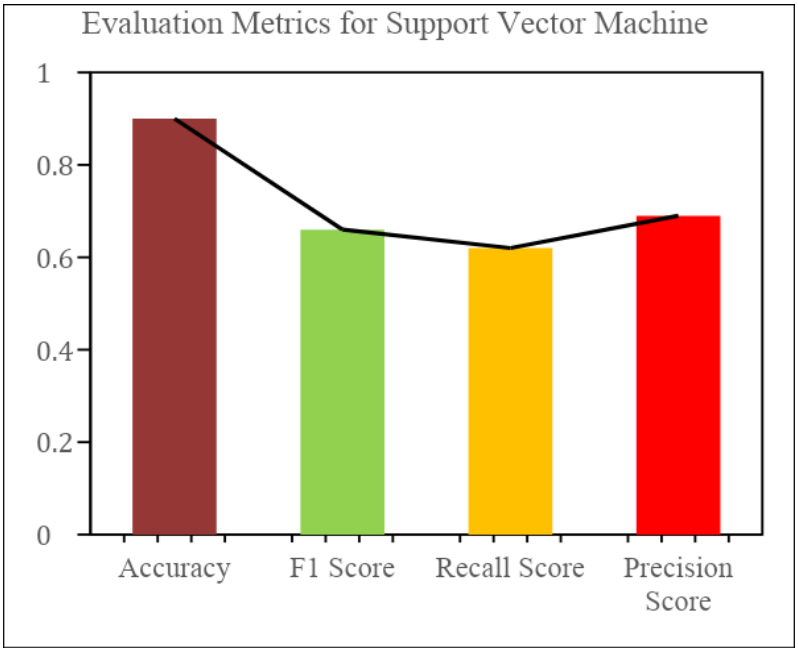

FIGURE 3 Evaluation metrics for Support Vector Machine

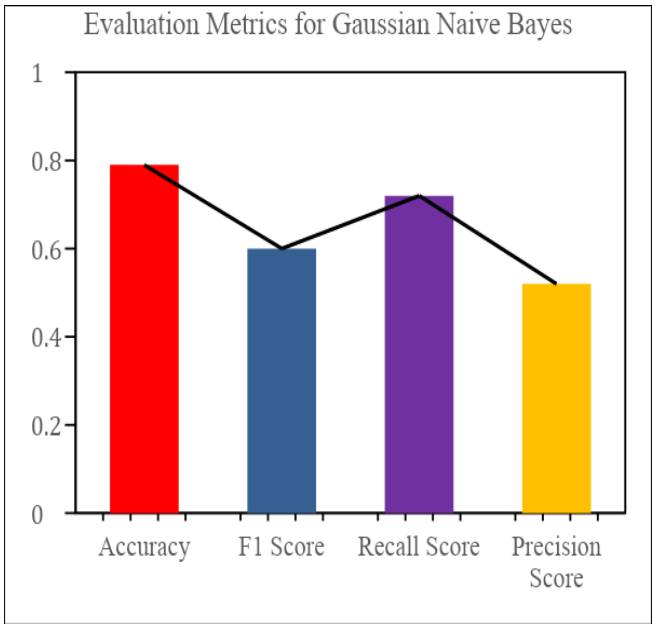

FIGURE 4 Evaluation metrics for Gaussian Naïve Bayes

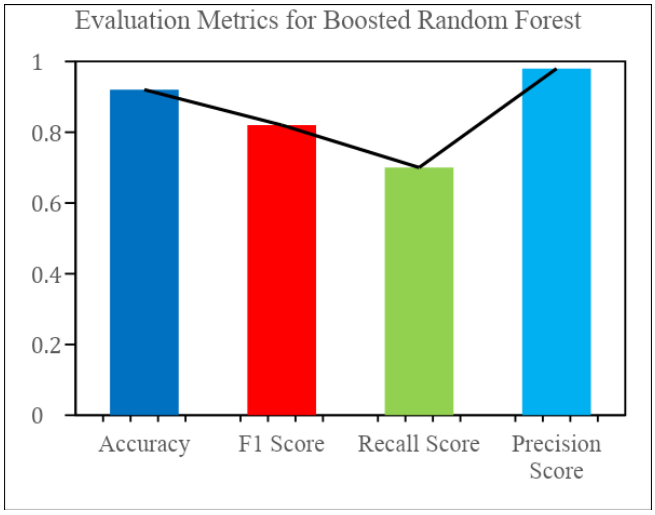

FIGURE 5 Evaluation metrics for Booster Random Forest.

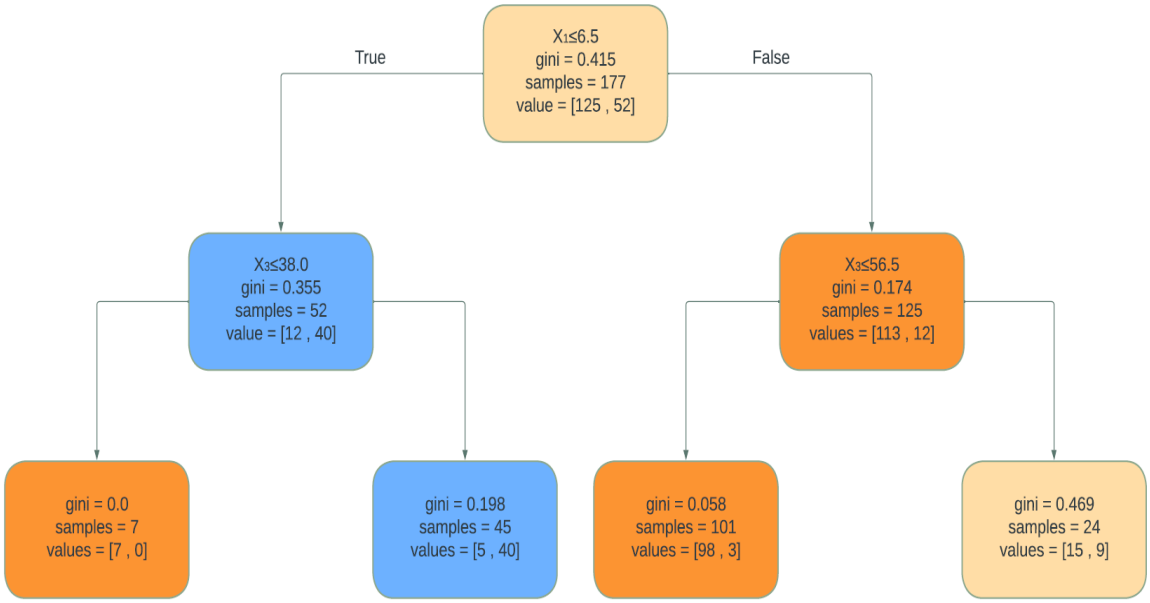

FIGURE 6 Decision Tree

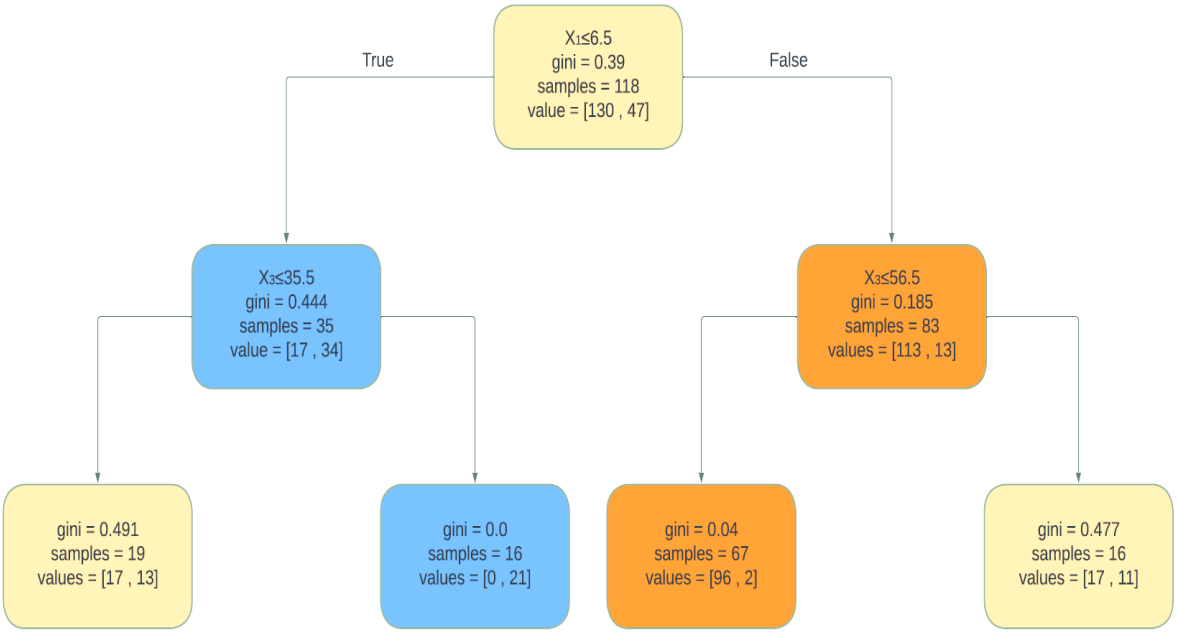

FIGURE 7 Decision Tree 1

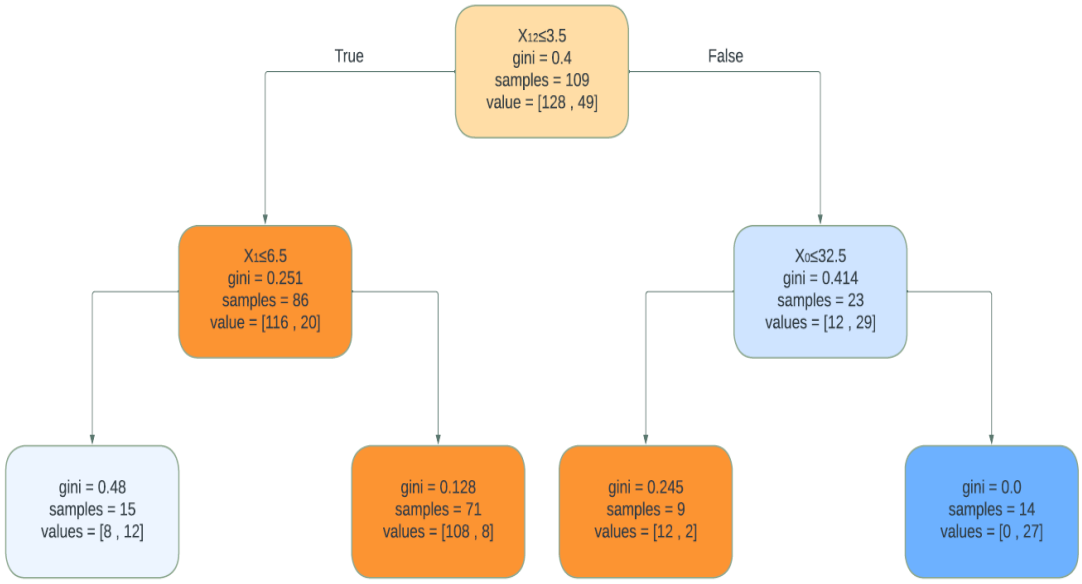

FIGURE 8 Decision Tree 10

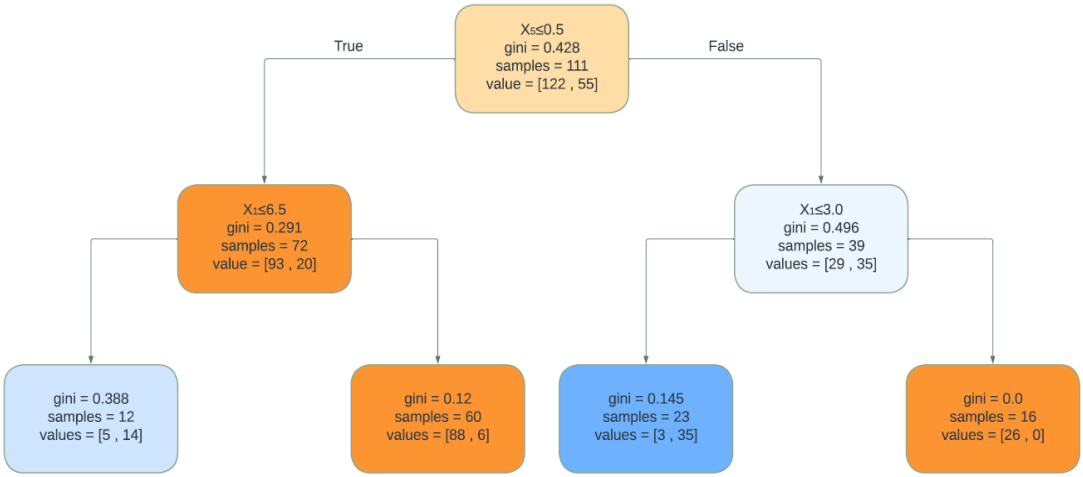

FIGURE 9 Decision Tree 25

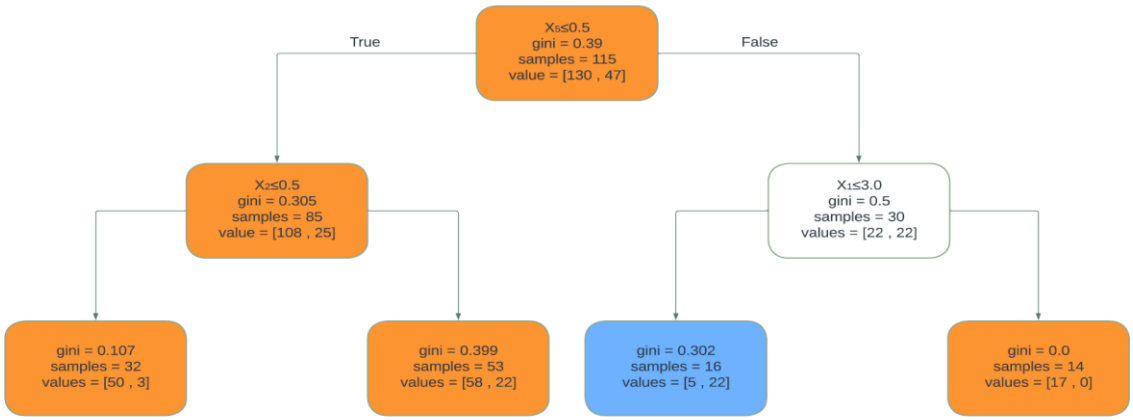

FIGURE 10 Decision Tree 100

Based on these results, the Boosted Random Forest algorithm emerged as the top-performing model and efforts were directed toward fine-tuning this model for optimal performance on this dataset.

5.1. Hyperparameter Optimization

In order to enhance the performance of the Boosted Random Forest Classifier, a grid search was conducted in order to identify optimal parameter combinations. By employing the GridSearchCV() function from the sklearn library, we systematically searched a grid of selected parameters to pinpoint optimal settings; results of the grid search algorithm, including details on selected hyperparameters, are detailed in Table 2.

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	0.86

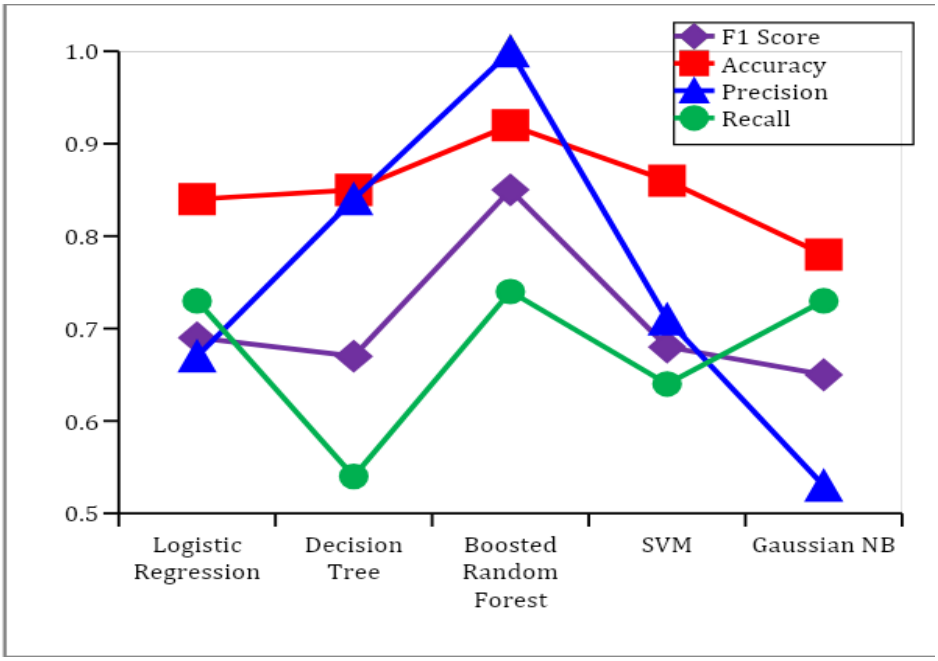

FIGURE 11 Comparison of Model's performance

Table 3 contains evaluation metrics of the Fine Tuned Boosted Random Forest model. These findings show that it exhibited superior performance in predicting COVID-19 patient fatalities, as illustrated by Figure 11. Additionally, all models, including Boosted Random Forest models, are shown here with comparative analyses; see Figure 11 for an illustration of this performance analysis.

VI. FUTURE WORK

In this study, Artificial Intelligence (AI) is pivotal in formulating efficacious treatment methodologies. We introduce an AI model that leverages the Boosted Random Forest algorithm, enhanced with the AdaBoost algorithm, to perform precise predictions on the COVID-19 patient dataset. The model has manifested exemplary robustness, achieving an F1 Score of 0.86, illustrating its adept performance even on datasets with imbalances. A meticulous examination of the dataset revealed a pronounced mortality rate among the natives of Wuhan compared to non-natives and a discernible disparity in mortality between males and females. The majority of the affected patients were observed to be within the age bracket of 20 to 70 years.

The trajectory of forthcoming research is poised to focus on the development of a holistic pipeline. This pipeline will amalgamate computer vision models designed for CXR scanning with models dedicated to processing demographic and healthcare data. The integration of these diverse models aims to facilitate demographic profiling and the incorporation of mobile health applications that leverage these

amalgamated models. One of the overarching objectives is to harness telemedicine to expedite screening and detection processes in regions impacted by COVID-19. This integrated approach is envisioned to fortify the responsiveness and preparedness in addressing future outbreaks, allowing for timely interventions and informed decision-making in managing and mitigating the impacts of such pandemics. The synergy between advanced AI models and diverse healthcare data can potentially unlock new avenues in predictive analytics, enabling more personalized and effective treatment strategies.

REFERENCES

[1]. Mahdy LN, et al. *Automatic x-ray covid-19 lung image classification system based on multi-level thresholding and support vector machine. MedRxiv; 2020. p. 2020.03. 30.20047787.*

[2]. Gizem, Yu H, et al. *Data-driven discovery of a clinical route for severity detection of COVID-19 paediatric cases. IET Cybersyst Robot 2020;2(4):205–6.*

[3]. Too J,Mirjalili S. *A hyper learning binary dragonfly algorithm for feature selection: a COVID-19 case study. Knowl Base Syst 2021;212:106553.*

[4]. Song J, et al. *Maximum likelihood-based extended Kalman filter for COVID-19 prediction. Chaos, Solit Fractals 2021;146:110922.*

[5]. Kumar N, Kumar H. *A novel hybrid fuzzy time series model for prediction of COVID-19 infected cases and deaths in India. ISA (Instrum Soc Am) Trans 2021;In Press, Corrected Proof.*

[6]. Cobre AdF, et al. *Diagnosis and prediction of COVID-19 severity: can biochemical tests and machine learning be used as prognostic indicators? Comput Biol Med 2021;134:104531*.

[7]. Arvind V, et al. *Development of a machine learning algorithm to predict intubation among hospitalized patients with COVID-19. J Crit Care 2021;62: 25–30.*

[8]. Pahar M, et al. *COVID-19 cough classification using machine learning and global smartphone recordings. Comput Biol Med 2021;135:104572.*

[9]. Ebinger J, et al. *A machine learning algorithm predicts duration of hospitalization in COVID-19 patients. Intel Based Med 2021;5:100035*

[10]. Zhang S, et al. *Identification and validation of prognostic factors in patients with COVID-19: a retrospective study based on artificial intelligence algorithms. J Intensive Med 2021;1:103–9.*

[11]. Singh M, et al. *Transfer learning–based ensemble support vector machine model for automated COVID-19 detection using lung computerized tomography scan data.* Med Biol Eng Comput 2021;59(4):825–39.

[12]. Cai H. *Sex difference and smoking predisposition in patients with COVID-19.* Lancet Respir Med. (2020) 8:e20.

AUTHORS

Ravikant Kholwal, a graduate from PDPM IIITDM Jabalpur, has significantly advanced intrusion detection systems, increasing accuracy by 25% through a novel algorithm. With publications in IJEAT and expertise in image classification, he has contributed to diverse projects using Django, ReactJs, and OpenCV. As a Software Engineer and CTO, he is proficient in JavaScript, React.js, AngularJS, Firebase, and is an AWS Certified Cloud Practitioner. Ravikant's blend of research and application skills distinguishes him in computing.