A New Prediction Data Model of High-Risk COVID-19 Patient with Smart Notification (HRCP-SN) Using Machine Learning Algorithm

Ganesh Yenurkar (ganeshyenurkar@gmail.com)
Yeshwantrao Chavan College of Engineering Department of Computer Technology
https://orcid.org/0000-0002-6270-4236

Sandip Mal
VIT Bhopal University

Research Article

Keywords: High-risk patient, Supervised Machine Learning, COVID-19, Risk Prediction, Healthcare, Optimized treatment

Posted Date: December 27th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2082268/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

A web application designed to predict high-risk patients affected by COVID-19 runs a machine learning (ML) model at the backend to generate results. The random forest classification technique (a type of supervised ML approach) is used to predict the high-risk status of patients who are COVID-19 positive and are at the initial stage of infection. To predict high-risk patients, the model uses the patients’ current underlying health conditions, such as age, sex, diabetes, asthma, hypertension, smoking, and other factors. After data preprocessing and training, the model could predict the severity of the patient with an accuracy of approximately in 65–70%. According to some studies, random forest ML models outperform other ML models for solving the challenge of predicting unusual events such as pneumonia, hypertension, diabetes, obesity, and chronic renal disease were the most contributory variables for model implementation. This project will help patients and hospital staff make necessary decisions and actions in advance. This will help healthcare workers arrange resources and hospital areas for high-risk COVID-19 patients. Thus, this study provides an effective and optimized treatment. Using this application and suitable patient data, hospitals can predict whether a patient will require urgent care.

1. Introduction

After emerging of the COVID-19 pandemic outbreak during December 2019 in China, the World Health Organization (WHO) reported it regarding (nCoV breakout) global widespread in March 2020. The novel coronavirus showed itself in a variety of ways, from symptomless infection to mild symptoms to severe disease, including respiratory and multiple organ dysfunction and death. Ever afterward the beginning of the COVID-19 pandemic, health-care assistants got more support, solidarity, and gratitude than they have ever gained [1].

The COVID-19 epidemic is putting great stress on medical resources globally throughout the world, prompting healthcare specialists to wonder who urgently needs treatment. Hospitalization is essential because of the severity of a few COVID-19 instances, and some individuals may require admission to the Intensive Care Unit. This places a significant load on healthcare professionals, who are already overworked in hospitals and need adequate hardware equipment. It can ease this pressure by gaining a better understanding of who is at risk of death because of COVID-19. In this way, it would be possible to identify patients who have a legitimate need for treatment and tailor treatment to these high-risk instances. Patients at elevated risk of severe results, based on their initial health condition of them when assessed positive for COVID-19, would receive priority treatment. The confined resources in hospitals may be better allocated and the burden on hospital staff would decrease.

In these unprecedented times, protection of lives and health of health care personnel is vital during the ongoing COVID-19 epidemic, allowing them for a stronger worldwide response. ML techniques allows to do the evaluation of input-output relationships in complicated systems. Because of its capacity to uncover nonlinear relationships, greater multi-factorial algorithm optimization, and inbuilt model validation approaches, ML models outperform traditional prediction models in risk classification.
Applications of the distinct ML paradigms for factors election and the classification of large datasets may provide great inference tools in the medical field and can be recommended for risk trend analysis/prediction of COVID-19 positive patients [1].

Our major objectives are:

To construct an ML model to predict if a COVID-19 patient will have critical COVID-19 soon, based on their initial health conditions at their admission as age, diabetes, sex, tobacco, and other health factors. To serve this tool as a Python web application to run the ML model on the user input file.

We arranged the proposed research work into different sections. Section 2 describes a literature survey of COVID-19 detection models performed by various researchers in learning methodologies. Section 3 proposes a method in classifying COVID-19 disease labels and the working of machine learning algorithm used. The results of simulations implemented to analyse the proposed method’s performance outcome are discussed in Section 4. It showed Limitations that are occurred of the proposed system in Section 5. The discussion section of the proposed model and its conclusion mentioned in Section 6. And at last, we provided the future scope for the research work under Section 7.

2. Related Work

In a research, different ML algorithms (shown in Table 1) were assessed and prediction of patient’s threat for severe COVID-19 based on their admission condition. The patients with severe COVID-19 at the time of admission were eliminated because of poor saturation of oxygen and partial arterial oxygen pressure (PaO2). The key result was the likelihood of developing a critical illness, which would be defined as mechanical ventilation multiple organ dysfunction syndrome, Intensive Care Unit (ICU) hospitalization, and/or expiration. Three distinct ML techniques were used to expect inpatient’ severeness, and they contrasted their results with the presently recommended predictors and the APACHE II score of risk prediction. There were 162 patients bedded with non-severe COVID-19 among the 6995 patients assessed, and 25 (15.4 percent) of them worsened to severe COVID-19. ML models beat all other metrics in prognosticate the threat of COVID-19 together with APACHE II score (ROC AUC of 0.92 vs 0.72, subsequently), with 88% of sensitivity, 92.7% of specificity, and 92.0% of accuracy. APACHE II score, WBC add up, schedule from inception of sign to entrance, oxygen saturation level along with WBC add up were the most key factors in the ML models. When compared to the most effective techniques available, they show an excellent accuracy of ML models in predicting severe COVID-19 [1].

In a study, illustrates the patient’s main well-being issues like age-groups, diabetes, gender, and other characteristics, it used the mortality estimation through a deep learning model to assess positive for COVID-19 illness. The accuracy, specificity, and sensitivity measures were used to assess and develop the models. The algorithm can estimate if a COVID-19 verified patient will be dead or not based on their information after pre-processing and training. We have compared the metrics of various models. The results have proven that the deep learning model is superior to the other techniques [2].
Besides the general trend analysis, a study examines the five most affected states in India, namely Maharashtra, Tamil Nadu, Karnataka, Andhra Pradesh, and Uttar Pradesh. In India and the adopted states, ARIMA (Autoregressive Integrated Moving Average) and time series, they employed future prediction models to create three categories of predictions: confirmed cases, fatalities, and recovered cases. The findings imply the models used are useful tools for predicting COVID-19 changes. The research also reveals that the ARIMA model outperforms the Prophet Model to forecast outbreaks. The projections may be valuable in strengthening government authorities, health institutions, and hospitals’ preparation to tackle the virus’ widespread spread [3].

To assure the quality of online teaching during the critical era of new coronavirus pandemic prevention and control, the educational administration department must pick the courses and improve oversight and assessment. The online teaching tool platform should receive technical help from the teaching development centre. Teachers and students must shift their perspectives on teaching and learning, and teachers must develop an online teaching assessment system. Relevant departments must study and provide solutions to problems that arise during implementing online instruction in a timely way. They should urge instructors to always keep a reserve of high-quality video resources on hand and to prepare for crises. With its spectacular tagline end software, Salesforce.com, the US online software services pioneer, pioneered the way in PaaS and SaaS platforms. The research on how to use cloud computing services to achieve tremendous success draws on and is inspired by the online teaching activity of colleges and institutions amid an epidemic [4].

Healthcare institutions, particularly in underdeveloped nations like Pakistan, are in danger of going beyond their capacity and limit because of a lack of vaccination and rapid viral transference from individual to individual. It is analytical to control resources appropriately oversee and, in these nations, to limit high-rise dying rate and the destruction it might bring. In this work, they considered a case study of a tiny Pakistani city whose healthcare resources do not oversee with such an epidemic. Because of a lack of resources, a major portion of COVID-19 inpatients had to be referred to metropolitan. To deal with the scarcity of resources, the data is used from COVID-19 inpatients in this tiny city and constructed and implemented an ML classification model to forecast the severity of the illness. I picked SVM to predict patient severity out of the seven algorithms that were considered and assessed. The model has a 60% accuracy rate and divides the severity of the inpatient into moderate and severe levels [5].

A method was proposed to generate an artificial neural network capable of predicting the survival of a patient, which includes archiving in an electronic database of patient health, patient health data comprises multiple data sets, each with at least one of the first parameters related to the heart rate variability data and the second related to the heart rate variability data, heart rate variability. Vital signs data, each with a third additional parameter related to patient survival, provide a network of interconnected nodes from an artificial neural network, nodes comprising many artificial neurons, each of which has at least one associated weighted input; and train the artificial neural network using the patient's health data such that the relevant weight of at least one input of each of the many artificial
neurons is adjusted to the respective parameters first, second, and third of different datasets of patient health data, so that the artificial neural network is trained to make predictions about patient survival [6].

A rationale technology system of the current invention employs both current knowledge and implicit information which may be statistically extracted from the training data to provide a method and equipment for detecting illness and curing a patient. A system for collecting patients’ data from a different place, analysing it in on a trained neural network, creating the diagnostic value, and communicating the diagnostic value to a different location is also included in this technology [7].

There is one invention in which it is more specifically a generic data-mining method for predicting illness progression and detecting high-risk inpatients. A technique for estimating the expansion of the infection and detecting high-risk inpatients includes the steps of providing clinical data and molecular genetic data, pre-processing the data, selecting a predetermined number of variables from the allotted data based on their combined/mutual information content, and automatically generating prediction data using ML [8].

A disease prognosis prediction model is the subject of this invention. When a computer is used to construct a prototype/model which will forecast the prognosis of a predefined disease from its clinical chemistry test value, several real clinical chemistry test values and the disease's actual prognostic value are entered into the computer and employed. The data collection program analyses it, identifies multiple clinical chemistry inspection items that affect the prognosis of the diseases mentioned above, determines the priority of the multiple items on the prognosis of the diseases mentioned above, and establishes and stipulates multiple clinical chemistry inspection items based on the priority. They used this program as the above model to determine the association between the range of clinical chemistry examination values and the prognostic value [9].
<table>
<thead>
<tr>
<th>Ref. No.</th>
<th>Objective</th>
<th>Dataset Used</th>
<th>Methodology Used</th>
<th>Accuracy (%)</th>
<th>Impact</th>
<th>Major Improvement Needed (Yes/ No)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[5]</td>
<td>Machine Learning Classification Algorithms was used for COVID-19 Severity Prediction: Pakistan case study with Limited Health Services</td>
<td>Kaggle dataset.</td>
<td>LDA, SVM, LR</td>
<td>60</td>
<td>Slight</td>
<td>Yes</td>
</tr>
<tr>
<td>[11]</td>
<td>AI is used to Predict COVID-19 Severity, it depends on the Correlation of Exaggerated Monocyte Activation, Excessive Organ Damage, and Hyper inflammatory Syndrome</td>
<td>--</td>
<td>Support Vector Machine (SVM), LR, Random Forest.</td>
<td>&gt; 80</td>
<td>Slight</td>
<td>No</td>
</tr>
<tr>
<td>[12]</td>
<td>To predict COVID-19 instances using a mixed machine learning and beetle antennae search technique</td>
<td>WHO dataset.</td>
<td>CESBAS-ANFIS, FPASSA-ANFIS</td>
<td>&gt; 90</td>
<td>Strong</td>
<td>No</td>
</tr>
<tr>
<td>[13]</td>
<td>Searching of library books using RFID transaction</td>
<td>--</td>
<td>RFID</td>
<td>92</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>Ref. No.</td>
<td>Objective</td>
<td>Dataset Used</td>
<td>Methodology Used</td>
<td>Accuracy (%)</td>
<td>Impact</td>
<td>Major Improvement Needed (Yes/ No)</td>
</tr>
<tr>
<td>---------</td>
<td>---------------------------------------------------------------------------</td>
<td>-----------------------------------</td>
<td>-------------------------------------------</td>
<td>--------------</td>
<td>--------</td>
<td>-----------------------------------</td>
</tr>
<tr>
<td>[14]</td>
<td>Hybrid method for the Prediction of sales Model in Tourism Industry and Hotel Recommendation.</td>
<td>Twitter and Facebook Dataset</td>
<td>social matrix factorization and base matrix</td>
<td>88</td>
<td>Slight</td>
<td>Yes</td>
</tr>
<tr>
<td>[15]</td>
<td>Topic Detection in Twitter Dataset Using Python</td>
<td>Twitter Dataset</td>
<td>Random forest</td>
<td>95</td>
<td>Strong</td>
<td>Yes</td>
</tr>
</tbody>
</table>

3. Methodology

Following figure-1 depicts the system architecture.

- This app is deployed on Heroku, which is a Salesforce platform.
- Python powers this web application. I used the flask as the web framework.
- We use Unicorn as the application server. We use the flask mail module for email sending purposes.
- We build Web pages using HTML, CSS, JavaScript. We use bootstrap as CSS and JavaScript framework.
- Random Forest ML algorithm runs in the backend to generate results

We showed the flowchart in Fig. 2. The flowchart shows the sequence of how the hospital admin can use the web application. This flowchart is sequential and is an easy-to-understand format.

3.1 Supervised Machine Learning Algorithm: Random Forest Algorithm (RFA)

We employ this algorithm in this project to classify the patients randomly. Random Forest works best if we need to classify some data based on some factors with no prior information. Thus, the random forest classifier worked best for this project.

For classification and regression analysis, a random forest supervised ML algorithm is used. It is an utmost ductile along with accessible algorithm. A random forest is encompassing of decision trees. It means that if there is a more tree, having the extra powerful forest.

Random forest generates decision trees randomly on choosing our data samples, called bootstrapping, and gets forecast from each tree to select the leading solution using polling. Various applications of random forest include suggestion engines, image classification, and attribute options. Other applications like analyse loyal loan applicants, identify fraudulent activity, and predict diseases, etc.
Random Forest Algorithm (RFA) can estimate and classify the data. It is also the most flexible and user-friendly algorithm. A decision tree is a type of tree that is found in a forest. I thought a forest to be powerful if there were extra trees. Random forest algorithm (RFA) uses bootstrapping randomly for picked data samples to generate decision trees, then get will get predictions from each tree and the best answer voted. feature importance, it may serve as a powerful indicator. For a range of tasks, we may use random forests via identify fraud transactions, predict the possibility of having disease, etc.

3.2 Binary Classification

A set of elements divided into two categories using a classification rule. This experiment employed binary classification, which is used to divide patients into two groups: those who are at elevated risk and those who are not.

3.3 Dataset

For this project, the raw dataset [10] was got from the Kaggle website. We processed the data using Python. Python modules used are Pandas (for handling CSV files), Matplotlib and Seaborn (to plot graphs), Sci-kit learn (for random forest classifier, metrics calculation, feature selection), Imblearn (for over-sampling), and NumPy. Jupiter’s notebook was used as the platform to run python scripts. The raw dataset [10] got obtained from the Kaggle website was then pre-processed to remove any null values, inconsistencies, outliers, unnecessary columns, converted the data into integer form, we rearranged columns to get desired format, etc. After all this, the refined and pre-processed dataset [10] was ready for the training.

3.4 Machine Learning

I do ML using Python. The refined dataset [10] is divided into x and y i.e., dependent variables (factors) and independent variables (target output). Then, x is split into a training metadata and test metadata. Then, the random forest framework is fitted with the dependent variable and the independent variables. This trains our model. Then the prediction is made on the input file (patient’s data). We evaluate the model using ROC AUC values and accuracy values. A graph is plotted between n_estimators (number of decision trees to be generated) and ROC AUC value, as shown in Fig. 4.

By this plot, the correct value of n_estimators can be estimated to be used in the model. For this project, the value of n_estimators are kept default, which is 100. Unique features used to train the model and we depict the output feature in Fig. 5.

3.5 System’s working

- From Fig. 5 it is shows that the training dataset with input-output features initially. The training dataset [10] contains columns: id, sex, in tubed, age, pregnancy, diabetes, pneumonia, hypertension, COPD, asthma, renal chronic, obesity, cardiovascular, tobacco.
- Then, using the patient’s data got from the hospital, a prediction is made. The 16 output/prediction is a binary classification in terms of 1 & 0. This 1 & 0 is converted into YES and NO, respectively.
Then, the result is generated. For this, we merge the prediction with the patient’s id, email id, and name. Hence, the result file will contain four columns: id, email, name, high-risk.

- After it generates the result file, an email can be sent to the patients. The email contains a message and the result file as an attachment.

### 3.6 Python web application

- The web application is based on python.
- We built our web application using the Python Flask application framework. Flask is a Python-based web development framework. Armin Reacher developed Flask, who is also the founder of the Pocco, a Python community. We focus the Flask template engine on the Jinja2 template engine and the Werkzeug WSGI toolkit. Both initiatives are the brainchild of Pocco.
- For our web application, we employed Gunicorn as the application server. Green Unicorn is a Python WSGI HTTP Server for UNIX developed by Gunicorn. For Python web applications, WSGI stands for Web Server Gateway Interface. All prominent web frameworks, including Flask and Django, operate with Gunicorn and other WSGI servers.
- Our web application uses the Bootstrap CSS and JavaScript framework. Bootstrap is a popular choice used for developing the mobile-first and responsive websites. It is completely free to download and use. Mark Otto and Jacob Thornton created bootstrap on Twitter.
- We use a custom email template for the body of the email.

### 3.7 Measurements

#### 3.7.1 Accuracy

Accuracy is a machine learning measurement that helps us to understand how many correct predictions our model can make. We represent it as a percentage (%). The higher the accuracy, the better is our model. The formula for accuracy is:

\[
Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}
\]

#### 3.7.2 AUC

We abbreviate the term area under the curve as AUC. The ROC curve in Fig. 12 is the one being discussed, having specificity on the horizontal-axis and sensitivity on the vertical-axis. The model’s interaction is best illustrated by an ROC curve because there is always a trade-off between these two metrics. A model with a high value for both will have a curve that approaches 90 degrees. For the desired curve like this, the maximum AUC value is 1.

#### 3.7.3 Precision
A classifier’s ability is referred for precision to prevent misclassifying a defeatist event as a good one. For each class, we would observe that it’s a proportion of true positives to the sum of true positives and false positives. On the other hands we make the optimistic predictions with the precision.

$$Precision = \frac{TP}{TP + FP}$$

### 3.7.4 Recall

Recall is a term that describes the capacity of an analysis to find every true class. We intend it using the proportion of true positives to the addition of true positives and false negatives for each classification. Recall is the percentage of correctly detected positives.

$$Recall = \frac{TP}{TP + Fn}$$

### 3.7.5 F1 Score

The weighted harmonic mean of accuracy and recall is an F1 score. F1 score of 1.0 is being considered as the inflated and 0.0 is considered as the lowest. F1 ratings are lower than accuracy assessments because of consideration of precision and recall. Of thumb, use the weighted mean of F1 instead of using overall accuracy to compare classifier models.

$$F1Score = \frac{2 * (Recall + Precision)}{(Recall + Precision)}$$

### 3.7.6 Support

Support is a count of the current phenomenon of the class in the specified dataset [10]. Fundamental concerns with the classifier’s reported scores are exposed by using imbalanced support in the training dataset, which will require classified sampling or rebalancing.

### 3.7.7 Macro Average and Weighted Average

The macro average gives each prediction the same weight when calculating loss, but if the data is uneven and we want to give specific predictions more weight (based on their proportion), the ‘weighted’ average is employed instead. First, Exploratory Data Analysis (EDA) is performed on the pre-processed data. Figure 6 shows the distribution of patients having various symptoms. We see that most patients consume tobacco (19%), have a renal chronic disease (16%), and have obesity (12%). These three types of people are almost 50% of the total patients. Patients having cardiovascular disease, hypertension, asthma, cold, diabetes, pneumonia cumulatively comprise almost 18% of the total patients. Also, almost 34% of patients are having some other diseases.
Most of the population lies in the age group of 40–80 years of age. So, we can conclude that a patient having age between 40–60 years has a high chance of having severe COVID-19 in near future. This correlates with the original dataset [10], in that the people having age greater than 30 or 40 years are more prone to risk having COVID-19 infection. These people are also a major portion of the working people of a nation, as shown in Fig. 7.

We have conduct ourselves with integrity, fidelity, and honesty. We have openly taken responsibility for our actions, and only make agreements, which we intend to keep. We will not intentionally engage in or participate in any form of malicious harm to another person or animal.

4. Results

Total 104606 patients' data was used to train the random forest ML model. We have presented a technique that can forecast the severity of the COVID-19 disease in inpatients. A severity prediction ML model is a useful tool for healthcare personnel in the medical field for managing resources of hospitals. The allocation of resources to a critical patient can have influence between their life and death [2]. The results got from this project are described below.

From Fig. 8, we can see that there are a greater number of males than females in the dataset [10]. Although as we can see in Fig. 10 and Fig. 11, gender does not have significant feature importance in determining the high-risk status of any patient. Before data pre-processing, there was the same number of males and females in the dataset [10]. After data pre-processing (specifically by removing null values), a lot of entries were removed from the original dataset [10] which resulted in the rise in the difference between the number of males and females. The pre-processed dataset [10] contains approx. 42000 females and approx. 63000 males.

From Fig. 9, twenty-nine% of patients are high-risk patients and 71% of patients are non-high-risk patients. Since there are significantly a smaller number of patients who have high-risk compared to the non-high-risk patients, this leads to the imbalanced classification problem. So, to train the model, the SMOTE (Synthetic Minority Over-sampling Technique) technique was used.

SMOTE method is used to tackle imbalanced classification. We can define SMOTE as an over-sampling technique in which minority class artificial samples are created by creating minority class copies through the original dataset. Here, high-risk patients are the minority class in which we generate synthetic samples. Synthetic samples simply mean the copies of original minority class samples. The problem of over-fitting is avoided using this method rather than just doing random over-sampling. Figure 10 represents the training dataset [10], which contains dependent variables (factors). Here, in the sex column, the Females are represented by one, and Males are represented by 2. In the remaining columns except for age group, Yes is represented by one, and No is represented by 0. Patients are divided into four age groups: 1–0 to 12 age (Child), 2–13 to 18 age (Adolescence), 3–19 to 59 age (Adult), 4–60 age and above (Senior Adult). Classifier algorithms are more biased towards the feature having more unique values (high cardinality features). They give more importance to these high cardinality features, which are
not good. This is an inherent behaviour of all the classification algorithms. For that purpose, we convert the age feature into the age group feature. I keep all the data in integer form for simplicity.

Figure 11 represents patients’ current health data. We train the ML model using this as an input file to predict high-risk patients.

Figure 12 shows the feature importance bar graph based on the mean decrease in impurity. For feature selection, the threshold value of 0.04 is used to get the top features. Using the threshold value as 0.04, we got the top six features, namely age group, covid_res, intubated, pneumonia, hypertension, diabetes. We retrain the model using the selected notable features.

Figure 14 represents the ROC curve got from the retrained model. The more the graphs tend towards the top-left corner, the better the classifier is. The AUC score got is 0.67 approx. which is slightly better than the previous AUC score. The feature importance as a pie chart as shown in the Fig. 13 below.

Figure 15 represents the confusion matrix of the model. Count values/percentages are used to summaries the number of accurate and inaccurate predictions.

We show the output of the model in Figure 16. For simplicity, 1 is replaced by YES, and NO replaced 0. Then this output is merged with the patient’s id, email id, and name. We represent the classification report of the model in Figure 18. I export the final output as a CSV file shown in Fig. 17. The output dataset [10] contains an independent variable (target output) with column name as high-risk. High-risk column contains 1 & 0. If high-risk is 1, the patient will have an elevated risk of COVID-19 soon. Else, the patient will not have an elevated risk.

Now, how hospital admin can use the web application is given below. The web application enables hospital admin to quickly upload patients’ data. After we upload the file, the random forest model works at the backend and generates the result file as result.csv filename. We can download the result by clicking the download button. By clicking on the email button, an email will be sent to the patients. We used a custom HTML email template for this purpose. The email contains a warning message, a link to show a list of nearby hospitals in Google Maps, and the result file and health report as an attachment. The patient’s health report is shown in Fig. 19.

5. Limitations

- Sometimes, the input data may not be correct and genuine.
- After it shows the results to patients, it will depend solely on the patient to act based on the report given by the hospital.
- We limited the model to the quantifiable factors related to a particular patient. It does not take all the affecting factors.
- This model may not give 100% correct predictions based on current COVID-19 data.
6. Discussion And Conclusion

Innumerable studies are now being conducted that employ ML algorithms for the prediction of high-risk patients who are committed to COVID-19 disease. Maximum number of research had an approach to the patients’ data from the hospital medical records. [10]. Our findings differ from those of the previous study [1], [2], [5] in terms of technique. We currently used only one ML algorithm, i.e., Random Forest Classifier. The earlier studies used various algorithms to assess which algorithm is performing best for their dataset [10] like Random Forest Classifier (RFC), SVM, Auto-encoder, ANN, LR, Regression Tree, etc. In our study, pneumonia and hypertension showed a greater contribution towards predicting high-risk patients, as shown in Fig. 11. As we can see from Table 1, five models are referenced. The models which have accuracy greater than or equal to 90% are considered having a strong impact and other models are considered having a slight impact. So, the models having a slight impact would need improvement to achieve a better score. Our ML model performed better than the reference model, having an accuracy of 60% [5]. Our model showed accuracy between 60%-70%.

We develop an ML model to predict a person’s severity after they have assessed positive for COVID-19. Our ML model showed an accuracy of approximately in between 65–70%. The model only predicted severity because of a COVID-19 diagnosis, but it can help doctors decide whether to admit someone positive for COVID-19. The coronavirus has the potential to have a noteworthy influence on affected people's quality of life. In the future, we can develop a model which can forecast not only the severity but also the severity of the growth of the disease. This step will encourage people to get medical help as soon as possible, preventing the sickness from weakening future dispositions in the infected person. This will prevent many people from being admitted to the Intensive Care Unit (ICU) if they can seek care elsewhere.

This model currently only predicts the result based on the patients’ symptoms, but in the future, we can also include various other factors and parameters in the system like wellbeing habit (physical extract), occupation, symptoms, and continual disease. When this information will be combined with the current availability of the medical assets in a certain range, we may create a correct arrangement for the total amount of restorative assets necessary.

- In the future, we can integrate this project with other medical-related apps.
- COVID-19 wave prediction can be added to this project in the future.
- Different ML algorithms can be used and assessed to predict the high-risk patients.

Declarations

Acknowledgement

This research would not be possible without the exceptional support of my supervisor. We are thanking to our colleagues from yeshwantrao Chavan College of engineering who provided insight and expertise that
assisted the research, although they may not agree with all the interpretations/conclusions of this paper.

We should not submit this research paper to more over one journal for simultaneous consideration. Also, I will commit myself with integrity, fidelity, and honesty. I will not intentionally engage in or take part in any form of malicious harm to another person or animal.

Funding

As there is no external funding received for this research. Also, we have no conflicts of interest to disclose. We discovered the raw dataset from the Kaggle website (https://www.kaggle.com/tanmoyx/covid19-patient-precondition dataset select=covid.csv).

References


Figures
Figure 1

Shows Proposed System Architecture
Figure 2

Shows System Flowchart
Figure 3

Shows High Level System Overview
Figure 4

Shows AUC score vs estimator's plot
Figure 5

Shows Input and Output features of the random forest ML model
Figure 6

Shows Percentage of patients
Figure 7

Shows Age distribution
Figure 8

Shows Gender distribution
**Figure 9**

Shows High-risk vs non-high-risk patients
Figure 10

Shows Training Dataset

<table>
<thead>
<tr>
<th>id</th>
<th>name</th>
<th>email</th>
<th>sex</th>
<th>intubed</th>
<th>pneumonia</th>
<th>age_group</th>
<th>diabetes</th>
<th>copd</th>
<th>asthma</th>
<th>hypertension</th>
<th>other_disease</th>
<th>cardiovascular</th>
<th>obesity</th>
<th>renal_chronic</th>
<th>tobacco</th>
<th>covid_res</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>167380</td>
<td>Chaimers Valer</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>085948</td>
<td>Lucine Gaynithote</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0001603</td>
<td>Rajiv Malik</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>186608</td>
<td>Shreyas Rajkumar</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1.23E+36</td>
<td>Juli Koobagade</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>000691</td>
<td>YCCE Student</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>060734</td>
<td>Zew Mokken</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>12695c</td>
<td>Allie Zarlet</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>15486</td>
<td>Myllam Rupay</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>16614e</td>
<td>Alexa Jenkyn</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 11

Shows Patient’s Health Data (Input File)

![RF Feature Importance](image)

**RF Feature Importance**

- age_group
- covid_res
- intubed
- pneumonia
- hypertension
- diabetes
- sex
- obesity
- other_disease
- renal_chronic
- cardiovascular
- tobacco
- copd
- asthma

Figure 12

Shows Feature importance- Bar graph
Figure 13

Shows Feature importance- Pie chart
Figure 14

Shows ROC Curve
Figure 15

Shows Model's confusion matrix
<table>
<thead>
<tr>
<th></th>
<th>high_risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>NO</td>
</tr>
<tr>
<td>1</td>
<td>NO</td>
</tr>
<tr>
<td>2</td>
<td>YES</td>
</tr>
<tr>
<td>3</td>
<td>YES</td>
</tr>
<tr>
<td>4</td>
<td>YES</td>
</tr>
<tr>
<td>5</td>
<td>YES</td>
</tr>
<tr>
<td>6</td>
<td>NO</td>
</tr>
<tr>
<td>7</td>
<td>NO</td>
</tr>
<tr>
<td>8</td>
<td>NO</td>
</tr>
<tr>
<td>9</td>
<td>NO</td>
</tr>
</tbody>
</table>

**Figure 16**

Shows The output of the ML model
<table>
<thead>
<tr>
<th>id</th>
<th>name</th>
<th>email</th>
<th>high_risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Chalmers Vasler</td>
<td><a href="mailto:cvasler0@unblog.fr">cvasler0@unblog.fr</a></td>
<td>NO</td>
</tr>
<tr>
<td>1</td>
<td>Lucine Gaythwaite</td>
<td><a href="mailto:lgaythwaite1@live.com">lgaythwaite1@live.com</a></td>
<td>NO</td>
</tr>
<tr>
<td>2</td>
<td>Prajwal Hatwar</td>
<td><a href="mailto:hatwarprajwal@gmail.com">hatwarprajwal@gmail.com</a></td>
<td>YES</td>
</tr>
<tr>
<td>3</td>
<td>Shreyas Rajurkar</td>
<td><a href="mailto:shreyasrajurkar13@gmail.com">shreyasrajurkar13@gmail.com</a></td>
<td>YES</td>
</tr>
<tr>
<td>4</td>
<td>Juli Khobragade</td>
<td><a href="mailto:julikhobragade923@gmail.com">julikhobragade923@gmail.com</a></td>
<td>YES</td>
</tr>
<tr>
<td>5</td>
<td>YCCE Student</td>
<td><a href="mailto:18010032@ycce.in">18010032@ycce.in</a></td>
<td>YES</td>
</tr>
<tr>
<td>6</td>
<td>Zed Muckeen</td>
<td><a href="mailto:zmuckeen6@washingtonpost.com">zmuckeen6@washingtonpost.com</a></td>
<td>NO</td>
</tr>
<tr>
<td>7</td>
<td>Allie Zarfat</td>
<td><a href="mailto:azarfat7@blogspot.com">azarfat7@blogspot.com</a></td>
<td>NO</td>
</tr>
<tr>
<td>8</td>
<td>Myriam Rapsey</td>
<td><a href="mailto:mrapsey8@businessinsider.com">mrapsey8@businessinsider.com</a></td>
<td>NO</td>
</tr>
<tr>
<td>9</td>
<td>Alexa Jenkyn</td>
<td><a href="mailto:ajenkyn9@ezinearticles.com">ajenkyn9@ezinearticles.com</a></td>
<td>NO</td>
</tr>
</tbody>
</table>

**Figure 17**

Shows Result of the patients (output file)
Figure 18

Shows Classification report of the ML model.

<table>
<thead>
<tr>
<th>class</th>
<th>precision</th>
<th>recall</th>
<th>f1 score</th>
<th>support</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.83</td>
<td>0.69</td>
<td>0.75</td>
<td>14802</td>
</tr>
<tr>
<td>1</td>
<td>0.46</td>
<td>0.66</td>
<td>0.54</td>
<td>6120</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>metrics</th>
<th>accuracy</th>
<th>support</th>
</tr>
</thead>
<tbody>
<tr>
<td>accuracy</td>
<td>0.68</td>
<td>20922</td>
</tr>
<tr>
<td>macro avg</td>
<td>0.65</td>
<td>20922</td>
</tr>
<tr>
<td>weighted avg</td>
<td>0.72</td>
<td>20922</td>
</tr>
</tbody>
</table>

Model's ROC AUC value: 0.670087739865572
Figure 19

Shows Health report

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- CovidDataset.csv