CovidXAI: Explainable AI Assisted Web Application for COVID-19 Vaccine Prioritisation

Deepraj Chowdhury*[1] | Saranda Poddar[2] | Soham Banarjee[1] | Riya Pal[3] | Abrar Gani[4] | Caroline Ellis[5] | Rajesh Chand Arya[6] | Sukhpal Singh Gill[7] | Steve Uhlig[7]

[1]Department of Electronics and Communications Engineering, International Institute of Information Technology, Naya Raipur, Chhattisgarh, India
[2]Department of Computer Science Engineering, RCC Institute of Information Technology, Kolkata, India
[3]Department of Environmental Science, University of Kalyani, Kalyani, India
[4]Department of Trauma and Orthopaedics, St George's University Hospitals NHS Foundation Hospital, London, UK
[5]Department of Anaesthesia, Ashford and St Peter's Hospitals NHS Foundation Trust, London, UK
[6]Department of Cardiac Anaesthesia, Hero DMC Heart Institute, unit Dayanand Medical College and Hospital, Ludhiana, Punjab, India
[7]School of Electronic Engineering and Computer Science, Queen Mary University of London, London, UK

Correspondence
*Deepraj Chowdhury, Department of Electronics Communication Engineering, International Institute of Information Technology (IIIT), Naya Raipur, India, 493661. Email: deepraj19101@iiitnr.edu.in

Abstract

COVID-19 vaccines have a limited supply, and there is a huge gap between supply and demand, leading to disproportionate administration. One of the main conditions on which balanced and optimal vaccine distribution depends are the health conditions of the vaccine recipients. Vaccine administration of front-line workers, the elderly, and those with diseases should be prioritized. To solve this problem, we proposed a novel architecture called CovidXAI, which is trained with a self-collected dataset with 24 parameters influencing the risk group of the vaccine recipient. Then, Random Forest and XGBoost classifiers have been used to train the model – having training accuracies of 0.85 and 0.87 respectively, to predict the risk factor, classified as low, medium, and high risk. The optimal vaccine distribution can be done using the derived from the predicted risk class. A web application is developed as a user interface, and Explainable AI (XAI) has been used to demonstrate the varying dependence of the various factors used in the dataset, on the output by CovidXAI.

KEYWORDS:
Explainable AI; COVID-19; Vaccine distribution; Web application; Random Forest; XGBoost

1 | INTRODUCTION

In a large country like India, having a high population density, besides social distancing [1], strategic vaccine distribution is needed to achieve herd immunity. The objective of this work is to build a web application that helps to identify the more vulnerable masses in receiving the vaccine so that the spread of Covid can be controlled. Previous research works investigate correlations between either increased age or other health parameters (comorbidities and other health conditions) or healthcare worker status, to the severity of the COVID infection. None of them combines or deals with all these factors at once to predict the risk group of an individual. CovidXAI considers a dataset consisting of 24 factors, including age, health parameters/diseases/disorders, healthcare worker status, etc., and builds a model to predict the risk group of an individual. A web application is developed to

[0]**Abbreviations:** XAI, Explainable Artificial Intelligence; ML, Machine Learning

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/itl2.381

offer an interactive user interface. Machine Learning has been used to predict the risk group of an individual. The working of the Machine Learning (ML) model is further justified and made transparent using Explainable AI (XAI) to show the contribution of each of the parameters to the model prediction results.

Our Contributions: The main contributions of this paper are: 1) A novel architecture called CovidXAI is proposed which uses machine learning model to predict the risk group of the user based on the given input, 2) Based on the risk factor, the user is classified into low, medium, or high risk group, 3) The urgency of vaccination of the user hence can be derived from the predicted risk group, 4) XAI is used to demonstrate the influence of each of the 24 parameters on the output risk factor in order to make the model's workings visible and 5) A web-application is developed to offer an interactive user interface for CovidXAI architecture. The paper is structured as follows: Section 2 consists of related works. Section 3 gives the proposed methodology. Sections 4 and 5 consist of performance evaluation, conclusion, and future work, respectively.

2 | RELATED WORK

Ejaz et al.[2] conducted detailed research on the correlation of different comorbidities (diabetes, obesity and asthma) with higher chances of lung damage and death. Wolff et al.[3] also conducted research regarding the risks and severity of patients infected with COVID-19, and how existing comorbidities lead to increased risks. Nguyen et al.[4] have found that front-line healthcare workers have a threefold increased risk of contracting COVID-19. The authors used a COVID symptom study app for its users to report daily information on symptoms and COVID-19 testing and used the data to find out the comparison results of the general community and healthcare workers. Bubar et al.[5] used an age-stratified susceptible, exposed, infectious, recovered (SEIR) model to conduct research on and evaluate COVID-19 vaccine Prioritisation strategies. The authors found out that the mortality rate was mostly minimized when the vaccine was prioritized for adults more than 60 years old. Sujatha et al.[6] compared results of different Machine Learning (ML) classifiers to predict whether a candidate's life is at risk on taking the COVID-19 vaccine, and considers whether taking the vaccine might actually be helpful to candidates with certain demographics. To date, no studies have examined the relationship with Covid-19 infection severity and other risk variables including pre-existing comorbidities or other illnesses, which can increase mortality and lung injury. Health parameters (comorbidities and health conditions) and health-care worker status are taken into consideration in existing works. A person's risk group cannot be accurately predicted by any previous work that does not include all of these variables at the same time. Prioritization of vaccinations using XAI and ML has not been attempted before. The proposed solution (CovidXAI) considers all 24 factors, including age, health parameters, healthcare worker status, etc., and builds a model to predict the risk group of an individual. The working of the model is further justified and made transparent using Explainable AI (XAI) to show the contribution of each of the parameters to the model prediction results. Comparing CovidXAI to other works is shown in Table 1.

TABLE 1 Comparison of CovidXAI with Related Works.

Author	Focus of Study	Main Findings	ML	XAI
Ejaz et al.[2]	COVID-19 and comorbidities	Increased mortality rate caused by comorbidities on COVID-19 infected patients	×	×
Wolff et al.[3]	Study of risk factors effecting severe COVID-19 infection	Effect of pre-existing comorbidities and chances of developing new ones	×	×
Nguyen et al.[4]	Increased risk of health-care workers contracting COVID-19	Comparison of COVID-19 infection rates of front-line health-care workers and the general community	×	×
Bubar et al.[5]	Prioritisation of COVID-19 vaccines	Prioritisation of higher age groups to reduce mortality rate using SEIR model based on age stratification	×	×
Sujatha et al.[6]	Suitable candidate prediction for COVID-19 vaccination	Prediction of danger to candidates on taking COVID-19 vaccine	×	×
CovidXAI (this work)	Risk factor on the basis of 24 factor including health condition, antibody status and job status	Prioritisation of vaccine on the basis of 24 different factors	✓	✓

(Abbreviations - ×:= method does not support the property, and ✓:= method supports the property)

3 | PROPOSED METHODOLOGY

The proposed solution to the problem is a prototype consisting of a machine learning(ML) model, which runs at the back-end of a web application, which ensures a smooth user experience. Figure 1 shows the concept of the prototype. The user is asked to enter a series of data such as age, various health conditions, gender, frontline worker status, which are the factors used to train the model. This data is collected at the front-end of the application, and fed to the trained ML model in the back-end, which is hosted in the cloud. The trained model processes the data and helps in predicting the respective risk group of the user. Then the predicted risk group, which is classified as 'low risk', 'medium risk', and 'high risk' is sent back to the front-end of the application, which is displayed to the user. To further justify how the model is working, and to make the working of the model transparent, the usage of XAI has been demonstrated. XAI clearly shows the varying impact of each of the attributes used to train the model, on the output.

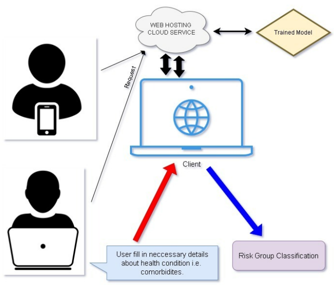

FIGURE 1 Conceptual diagram of CovidXAI

Dataset: The dataset, which has been referred to in this paper[7], is used to train the model. It has been made from data collected from 2512 individuals. This training data has 25 attributes upon which it is based. They are - gender, pregnancy, age, cancer, down syndrome, COPD, heart condition, sickle cell, BMI group, antibody test, immunocompromised state, smoking, type 2 diabetes, asthma, type 1 diabetes, hypertension, liver disease, frontline worker status, chronic kidney disease, cerebrovascular disease, neurological disorder, pulmonary fibrosis, thalassemia, and decompensated cirrhosis. All the attributes are entered as boolean (1 and 0, corresponding to yes and no, respectively), except BMI group and age. BMI groups have been classified into three – overweight, normal, and underweight. The target variable – the risk group, is based upon the values of these 24 variables. The resultant risk factor is a decimal between 0 and 1, which is classified into 3 categories – low risk(**0.0 to 0.4**), medium risk(**0.41 to 0.7**), and high risk(**0.71 to 0.99**). The more the risk factor, the higher the mortality rate in case of COVID infection.

Machine Learning Models: For predicting the risk factor of any individual, a data-driven approach has been considered in this prototype. There is no mathematical approach for predicting the risk group of a person. To collect data, help was taken from medical professionals, and they have been confirmed by the same based on clinical experience and published literature on COVID-19. The long term results of COVID-19 infection are not yet published as the disease itself is 2 years old. To predict the risk factor, machine learning (ML) has been used. Among various ML algorithms, Random Forest and XGBoost classifiers have been implemented[7]. 1) **Random Forest** is a supervised learning algorithm, which makes use of an ensemble of decision trees, which leads to more accurate results.This dataset also has a high number of parameters, so the risk of overfitting is more. 2) **XGBoost** is is a faster ensemble classifier that uses the gradient boosting framework and it uses CART (Classification and Regression Trees) - where each node contains a score instead of a single decision.

Explainable AI: In healthcare applications, the use of AI is not trustworthy enough for healthcare professionals, as the machine learning models are essentially a black box to medical experts[8] due to the poor explainability of the working of these models. XAI helps to explain the working of machine learning models, and shows the impact of every feature or attribute used to train the model. Simply put, it shows how much the model predictions change if we change the value of a single feature. The use of Shapley values has been made to implement the usage of XAI.

- **SHAP:** SHapley Additive exPlanations (SHAP)[9] values are a measure of the feature importances and is a method to explain individual predictions. It is based on the game theoretically optimal Shapley values. Its goal is to explain the prediction of an instance by computing the contribution of each feature to the prediction. In Figure 6 , the SHAP values of

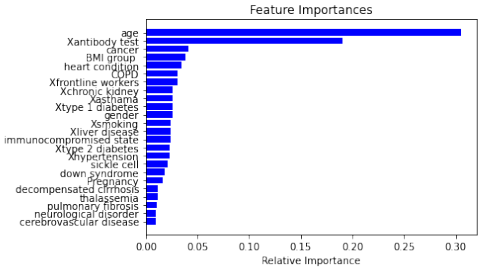

FIGURE 2 Random Forest Feature Importance

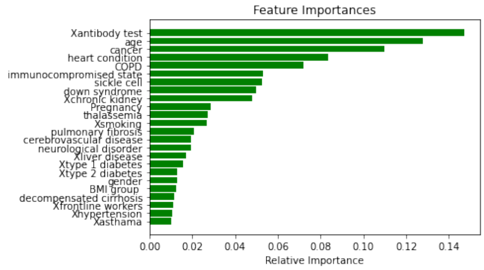

FIGURE 3 XGBoost Feature Importance

each feature shows clearly the contribution of each feature to the predictions made by the Random Forest model. Figure 8 shows the average impact of different features on the model output. Figures 7 and 9 show the same for XGBoost model.

Web Application: The web application is made using streamlit, which acts as a user interface. The user gives the input of 24 parameters and gets the result. The model itself is coded in a python notebook environment – and is contained separately as a python notebook(.ipynb). The model is then dumped in a pickle (.pkl) file. Then the inputs entered by the user are entered into the model which is in the form of a pickle file and has been imported in the .py streamlit file. After the data is fed to the model, the risk group is automatically classified as low, medium, or high risk. The high-risk group indicates that the user needs urgent vaccination and vice versa. Then this whole application can be deployed with the help of a cloud service, after which anyone with the link to the application can use it.

4 | PERFORMANCE EVALUATION

CovidXAI is a data-driven approach which has been adopted due to the lack of a mathematical procedure in predicting the Prioritisation group of an individual. Random Forest and XGBoost classifiers are used in the model for the prediction of the user's risk group. The risk factor denotes the user's probability of mortality on contracting COVID-19. Person with higher risk needs to be vaccinated before than the person with lesser risk. The following are the performance metrics of the classifiers used to classify the risk groups from the various parameters: Table 2 shows the different performance metrics of the Random Forest(RF) and XGBoost(XGB) classifiers, used in the proposed prototype - indicating how well the models perform while making predictions on the test data.

TABLE 2 Class-wise Performance metrics of both classifiers

Metric	Accuracy		Precision		Recall		f1-score	
Class	**RF**	**XGB**	**RF**	**XGB**	**RF**	**XGB**	**RF**	**XGB**
0.0(Low)	0.9	0.93	0.92	0.92	0.90	0.93	0.91	0.93
1.0(Medium)	0.71	0.68	0.69	0.80	0.71	0.68	0.70	0.74
2.0(High)	0.86	0.98	0.83	0.81	0.86	0.91	0.85	0.86
Overall	0.85	0.87	0.82	0.84	0.82	0.84	0.82	0.84

Based on the predictions of the built model, Figures 2 and 3 are plots showing the impact of an attribute in a prediction by the model. Figure 4 and 5 shows the confusion matrix of the random forest classifier and XGBoost classifier, which conveys the information of true prediction and false prediction by the model. Figures 6 and 7 demonstrate the usage of XAI, for RF and XGB, respectively, with the help of SHAP values to clearly explain the impact of the different features on the risk group classifications - the horizontal location shows whether the effect of that value is associated with a higher or lower risk prediction.

Figures 8 and 9, for RF and XGB, respectively, show the impact of each factor in the prediction of classes 0, 1, 2, that is low, medium, and high risk, respectively. In Figures 10 and 11, in the case of Random Forest and XGBoost respectively, the SHAP values show how much each factor contributed to the model's prediction when compared to the mean prediction.

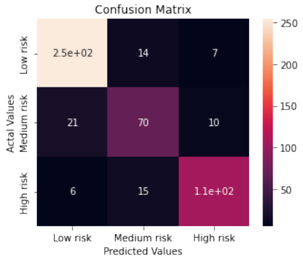

FIGURE 4 Confusion Matrix of Random Forest Classifier

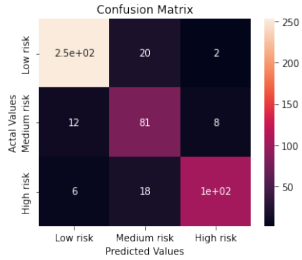

FIGURE 5 Confusion Matrix of XGBoost Classifier

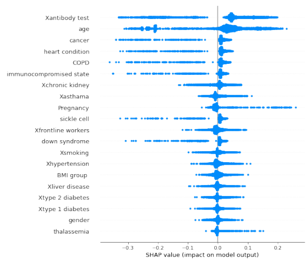

FIGURE 6 Plot showing SHAP values of Random Forest Classifier

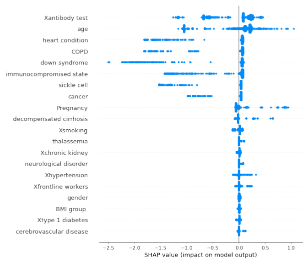

FIGURE 7 Plot showing SHAP values of XGBoost Classifier

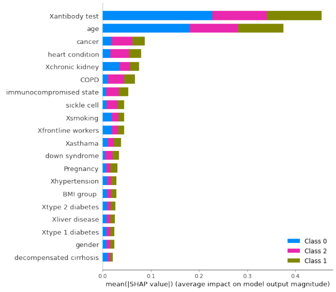

FIGURE 8 Plot showing average impact of different features on the Random Forest model output

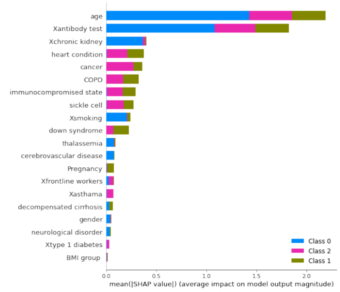

FIGURE 9 Plot showing average impact of different features on the XGBoost model output

Large positive or negative SHAP values indicate that the feature had a significant impact on the model's prediction. Compared to Covacdiser[7], CovidXAI performs better with a maximum precision of 84% compared to 82% of Covacdiser.

5 | CONCLUSIONS AND FUTURE WORK

Mass vaccination has been found and proven to be the most effective method to controlling this pandemic. Highly populated and developing/undeveloped countries are still facing a significant shortage of vaccine doses. Imposing lockdowns harms the economy, and cannot be seen as a viable alternative to mass vaccination. CovidXAI serves as an effective solution to this problem by identifying the risk grouping of population. It takes into consideration 24 factors to classify the risk group of an individual, which can be used to classify his/her urgency in receiving the vaccine. The long-term analysis of these risk factors is not available as COVID-19 disease is just 2 years old. CovidXAI helps to allocate more vaccines for the more vulnerable population - reducing mortality, helping in achieving herd immunity faster, and preventing the virus from spreading faster. It has been aimed to add more features to make better predictions and increase the viability of the model. To increase the accuracy

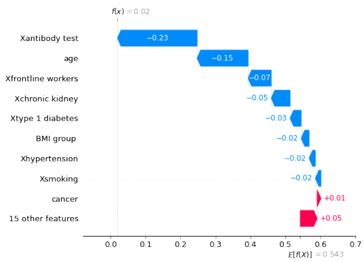

FIGURE 10 Waterfall Plot showing SHAP values of each feature in Random Forest Classifier

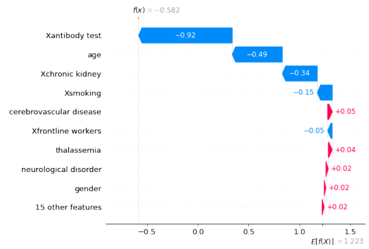

FIGURE 11 Waterfall Plot showing SHAP values of each feature in XGBoost Classifier

of the model and to increase the efficiency of the prototype, deep learning algorithms can be used. User accessibility can be improved by making a mobile application for android and iOS devices by utilizing the concept of Internet of Things (IoT)[10].

DATA AVAILABILITY STATEMENT

The dataset used for the experimentation can be accessed from https://github.com/deepraj2001/COVID-VACCINE/.

References

1. Yadav S, Gulia P, Gill NS, Chatterjee JM. A Real-Time Crowd Monitoring and Management System for Social Distance Classification and Healthcare Using Deep Learning. *Journal of Healthcare Engineering* 2022; 2022.
2. Ejaz H, Alsrhani A, Zafar A, et al. COVID-19 and comorbidities: Deleterious impact on infected patients. *Journal of infection and public health* 2020; 13(12): 1833–1839.
3. Wolff D, Nee S, Hickey NS, Marschollek M. Risk factors for Covid-19 severity and fatality: a structured literature review. *Infection* 2021; 49(1): 15–28.
4. Nguyen LH, Drew DA, Graham MS, et al. Risk of COVID-19 among front-line health-care workers and the general community: a prospective cohort study. *The Lancet Public Health* 2020; 5(9): e475–e483.
5. Bubar KM, Reinholt K, Kissler SM, et al. Model-informed COVID-19 vaccine prioritization strategies by age and serostatus. *Science* 2021; 371(6532): 916–921.
6. Sujatha R, Venkata Siva Krishna B, Chatterjee JM, et al. Prediction of suitable candidates for covid-19 vaccination. *Intelligent Automation and Soft Computing* 2022: 525–541.
7. Chowdhury D, Banerjee S, Dey A, Biswas D, Bhattacharyya S. Covacdiser: A Machine Learning-Based Web Application to Recommend the Prioritization of COVID-19 Vaccination. In: Springer. 2022 (pp. 105–117).
8. Gill S. AI for next generation computing: Emerging trends and future directions. *Internet of Things* 2022; 19: 1-34.
9. Parsa AB, Movahedi A, Taghipour H, Derrible S, Mohammadian AK. Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. *Accident Analysis & Prevention* 2020; 136: 105405.
10. Shao T, Chowdhury D, Gill SS, Buyya R. IoT-Pi: A machine learning-based lightweight framework for cost-effective distributed computing using IoT. *Internet Technology Letters* 2022: e355.

How to cite this article: Deepraj C., Saranda P., Soham B., Riya P., Gani A., Ellis C., Arya R. C, Gill S.S, and Uhlig S. (2022), CovidXAI: Explainable AI Assisted Web Application for COVID-19 Vaccine Prioritisation, *Internet Technology Letters, 2012;00:1–6.*