Evolutionary Intelligence
https://doi.org/10.1007/s12065-022-00704-3

SPECIAL ISSUE

Vaccine hesitancy in the post-vaccination COVID-19 era: a machine learning and statistical analysis driven study

Himanshu Gupta[1] · **Om Prakash Verma**[1]

Received: 6 July 2021 / Revised: 30 November 2021 / Accepted: 4 February 2022

Abstract

Background The COVID-19 pandemic has badly affected people of all ages globally. Therefore, its vaccine has been developed and made available for public use in unprecedented times. However, because of various levels of hesitancy, it did not have general acceptance. The main objective of this work is to identify the risk associated with the COVID-19 vaccines by developing a prognosis tool that will help in enhancing its acceptability and therefore, reducing the lethality of SARS-CoV-2. **Methods:** The obtained raw VAERS dataset has three files indicating medical history, vaccination status, and post vaccination symptoms respectively with more than 354 thousand samples. After pre-processing, this raw dataset has been merged into one with 85 different attributes however, the whole analysis has been subdivided into three scenarios ((i) medical history (ii) reaction of vaccination (iii) combination of both). Further, Machine Learning (ML) models which includes Linear Regression (LR), Random Forest (RF), Naive Bayes (NB), Light Gradient Boosting Algorithm (LGBM), and Multilayer feed-forward perceptron (MLP) have been employed to predict the most probable outcome and their performance has been evaluated based on various performance parameters. Also, the chi-square (statistical), LR, RF, and LGBM have been utilized to estimate the most probable attribute in the dataset that resulted in death, hospitalization, and COVID-19. **Results:** For the above mentioned scenarios, all the models estimates different attributes (such as cardiac arrest, Cancer, Hyperlipidemia, Kidney Disease, Diabetes, Atrial Fibrillation, Dementia, Thyroid, etc.) for death, hospitalization, and COVID-19 even after vaccination. Further, for prediction, LGBM outperforms all the other developed models in most of the scenarios whereas, LR, RF, NB, and MLP perform satisfactorily in patches. **Conclusion:** The male population in the age group of 50–70 has been found most susceptible to this virus. Also, people with existing serious illnesses have been found most vulnerable. Therefore, they must be vaccinated in close observations. Generally, no serious adverse effect of the vaccine has been observed therefore, people must vaccinate themselves without any hesitation at the earliest. Also, the model developed using LGBM establishes its supremacy over all the other prediction models. Therefore, it can be very helpful for the policymakers in administrating and prioritizing the population for the different vaccination programs.

Keywords COVID-19 · Machine Learning · Predictive Analysis · SARS-CoV-2 · Statistical Analysis · VAERS

Abbreviation

CDC	The United States Centers for Disease Control and Prevention
LGBM	Light Grading Boosting Machine
LR	Logistic Regression
MERS	Middle East Respiratory Syndrome
MLP	Multilayer feed-forward Perceptron
NB	Naive Bayes
RF	Random Forest
SARS	Severe Acute Respiratory Syndrome
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus-2
VOC	Variant of Concerns
VOI	Variants of Interests
WHO	World Health Organization

✉ Om Prakash Verma
vermaop@nitj.ac.in

Himanshu Gupta
guptah.nitj@gmail.com

[1] Department of Instrumentation and Control Engineering, Dr. B R Ambedkar National Institute of Technology, Jalandhar, India

Published online: 09 March 2022

1 Introduction

The Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) also known as COVID-19, has created an unprecedented worldwide health emergency. The first case of COVID-19 was reported in late December 2019, Wuhan, China and, exploded into every corner of the world in a flash. Because of this, World Health Organization (WHO) declared this a global pandemic on 11th March 2020 [1]. Though its origin and source are still unknown, there has been a considerable discussion on its origin and scientists believe that bat could be the most likely primary reservoir [2]. It shares approximately 79.5% and 50.0% genomic homology with Severe Acute Respiratory Syndrome (SARS) and the Middle East Respiratory Syndrome (MERS), the other two members of the coronavirus family, respectively [3]. Both SARS and MERS raised international concerns as they have been associated with a high mortality rate of 9.6% and 36.0%, amongst the diagnosed people, respectively [4]. Therefore, at the onset of COVID-19 all the governments have imposed unparalleled mitigation steps to control its spread however, it continues to ravage the world with more than 181 million cases and 3.9 million deaths as of June 2021 [5]. Further, to minimize the impact of this unprecedented invisible force and help the policymakers, many literatures have been found which either uses predictive modelling to forecast the peak [6] or technology (IoT) to develop a smart and contactless industry [7].

Like any other virus, SARS-CoV-2 also has the mutation capability because of which WHO has identified 11 variants as Variants of Interests (VOI) out of which 4 have been assessed as both VOI and Variant of Concerns (VOC) [8]. The VOIs have been identified as responsible for community transmission and VOCs for the drastic change in COVID-19 epidemiology or decrease in the effectiveness of available diagnostics, vaccines, therapeutics, or other mitigation strategies. The Epsilon (B.1.427/B.1.429), Zeta (P.2), Eta (B.1.525), Theta (P.3), Iota (B.1.526), Kappa (B.1.617.1), and Lambda (C.37) have been assigned as VOI whereas, Alpha (B.1.1.7), Beta (B.1.351), Gamma (P.1), and Delta (B.1.617.2) as both VOI and VOC.

Because of this, the research fraternity burned the midnight oil for the unprecedented development of the COVID-19 vaccine for public use. Consequently, 288 vaccine candidates have been developed by 25th June 2021 [9]. Out of these, 184 are in pre-clinical trials and, 36 are in Phase-I trials, 28 in the combined Phase I/II trials, 10 in Phase II trials, 7 in the combined Phase II/III trials, 18 in Phase III, and 5 in Phase IV of development. Amongst them, 17 (16.35%) are RNA vaccines, 10 (9.62%) are DNA vaccines, 17 (16.35%) are non-replicating vector vaccines, 4 (3.85%) are replicating vector vaccines, 16 (15.38%) employs inactivated virus, 2 (1.92%) are live-attenuated virus vaccine, 33 (31.73%) are protein subunit vaccine, and 5 (4.81%) uses virus-like particles. Further, most of the vaccines that are in Phase III and IV have shown more than 90% efficacy in preventing the deadly killing spree of COVID-19 [10]. Therefore, 18 vaccines have been approved by at least one national regulatory authority for public use. These include two RNA vaccines, eight conventional inactivated vaccines, five viral vector vaccines, and three protein subunit vaccines. The approved RNA vaccines have been developed by Pfizer–BioNTech (BNT162) and Moderna (mRNA-1273) whereas, Sinopharm (BBIBP-CorV and WIBP-CorV), Sinovac (CoronaVac), Bharat Biotech (Covaxin), The Chumakov Centre (CoviVac), Shifa Pharmed (COVIran Barakat), Minhai-Kangtai (KCONVAC), and Research Institute for Biological Safety Problems (QazCoVac) developed vaccines using an inactivated virus. Further, Gamaleya Research Institute of Epidemiology and Microbiology developed Sputnik V and Sputnik Light whereas, Oxford–AstraZeneca, CanSino Biologics, and Johnson & Johnson developed ChAdOx1-s, Ad5-nCoV, and Ad26.COV2.S as viral vector vaccines respectively. Also, protein subunit vaccines (EpiVacCorona, CIGB-66, and ZF2001) have been developed by Vector Institute, Center for Genetic Engineering and Biotechnology, and Anhui Zhifei Longcom Biopharmaceutical Co. Ltd. respectively [11].

Currently, almost all countries have started their vaccination program. However, due to the shortage of vaccines, it began with the prioritizing to those who have more susceptible to the adverse effect of COVID-19 such as elderly people, individuals with specific chronic diseases, and front-line medical persons [12]. In this line, India has started the world's biggest free vaccination program on 16th January 2021 with the target of 30 million front-line health care workers [13]. At present, over 41 million doses of vaccine have been administered daily and as of 27th June 2021, more than 2.8 billion have already been administered [14]. Therefore, approximately 22.6% of the global population have already received at least one dose of the COVID-19 vaccine. However, the majority of this comes from high-income countries, China, and India whereas, only 0.9% population in low-income countries have been vaccinated at least once. This represents that even in the current global pandemic, every country wants to secure its citizens first. More specifically, there is a large gap between the number of doses administered per 100 people worldwide. The countries like United Arab Emirates have 153 doses for every 100 people whereas, Chad have only 0.1. Therefore, there is an urgent requirement of population based vaccine distribution so that the vaccination for all becomes possible in stipulated time.

Further, besides the development and distribution of vaccines, the willingness of people to vaccinate themselves plays a crucial role to eradicate such a global and devastating virus [15]. However, it has been found that people have various levels of hesitancy because no medicine is free from side effects [16]. Based upon the previous experience (H1N1 eruption in 2009), the major path block in general acceptance has been the concern regarding safety and trust [17]. Similarly, the lack of trust over health authorities has been witnessed in the vaccine trials of HPV and HIV in Europe and the United States [18]. Nevertheless, the vaccines authorized for public use have been evaluated through exhaustive clinical and public trials but, those developed in outbreaks do not always have sufficient public trials. Therefore, it becomes very difficult to anticipate the adverse reactions of rapidly deployed vaccines, such as the vaccines developed in the ongoing pandemic. It has been generally found that the COVID-19 vaccines have some adverse allergic reactions and side effects however, these reactions have been mostly evidenced in individuals with pre-existing chronic comorbidities such as diabetes mellitus, cardiovascular disease, and allergic to a specific compound. Sometimes, some rare risk factors might have also been witnessed but considering the limitation of time along with the size and nature of trials, they cannot be perceived in clinical trials. Furthermore, the side effects of a vaccine are separate issues and do not indicate its effectiveness. Also, the clinical and demographic information of individuals has a direct impact on the effectiveness of any vaccine.

Although, there have been very few reported adverse reactions for COVID-19 vaccines, however, in rare cases anaphylaxis (a life-threatening allergic reaction), being developed within minutes to hours after vaccination, like reactions have been observed [19]. Also, some fatalities have been reported after the COVID-19 vaccination however, it has been evidenced that the rate of these breakouts is very low and the role of COVID-19 vaccines in these fatalities is still under investigation [20]. Therefore, the United States Centers for Disease Control and Prevention (CDC) assessed the symptoms after vaccination for close surveillance of any direct or indirect effects of vaccinations. Though, it has been found that the ratio of adverse effects to the number of vaccinations is very low, they cannot be overlooked. They not only provide useful information to anticipate the unwanted outcomes but also may help in achieving the general acceptability of the vaccine.

The Machine Learning (ML) models have shown their expertise in characterizing the hidden patterns of data and therefore, have been employed in various complex classification tasks [21]. Therefore, in the present investigation, ML methodology has been developed to identify the individuals with serious complications of vaccination. This will not only assist the authorities in vaccinating them with a safe medical environment to avoid any breakout but also help in developing greater trust for vaccination programs. Therefore, it will make COVID-19 vaccination much safer so that even the last man gets involved in these vaccination drives with enthusiasm. In summary, the main contribution of this work has been summarized as:

1. To the best knowledge and belief of authors, this has been the very first study that analyses the impact of COVID-19 vaccines employing more than 354 thousand samples.
2. To investigate the requirement of early hospitalization in SARS-CoV-2 patients which may help in reducing the lethality of the disease.
3. To identify and analyze the most probable causes in the individuals' medical history that resulted in adverse reactions to vaccination.
4. To explore the prominent symptoms that may result in the need for close observation after vaccination.
5. To analyze the most significant factors that resulted in breakouts even after the vaccination.
6. To develop the ML models for the prediction and classification of individuals most susceptible to the adverse effects of vaccination and therefore, may require high medical attention.

The rest of this paper has been organized as follows: The materials and methods being employed for the present investigation have been presented in Sect. 2. This section also discusses about the dataset used for the analysis (2.1), proposed framework (2.2), and simulation setup and metrics (2.3). Then, the detailed analysis being done in three parts along with the obtained investigational results has been presented in Sect. 3. Finally, the concluding remarks of this investigation work have been summarized in Sect. 4.

2 Materials and methods

This section focuses on the dataset and methodology being employed for the present investigation. In this regard, Sects. 2.1, 2.2, and 2.3 explain the dataset utilized, proposed framework and, simulation setup and metrics used to assess the performance of the developed models respectively.

2.1 Data collection

The raw dataset of individuals who have been vaccinated between 1st January 2021 to 11th June 2021 and also reported adverse reactions has been acquired from the Vaccine Adverse Event Reporting System (VAERS) website [22]. The VAERS has been established in the 1990s with the aim to detect possible safety problems in the

USA and, is co-managed by CDC. The collected data has three files in csv file format describing the general data, vaccination status, and symptoms. This acquired dataset contains individuals who have been vaccinated for various diseases such as COVID-19, Flu, Influenza, etc. However, for the present investigation, individuals vaccinated for only SARS-CoV-2 have been considered and the rest has been omitted. Therefore, the dataset being utilized consists of more than 354 thousand unique individuals. This acquired and cleaned dataset have various attributes of individual's information such as age, sex, current illness, medical history, allergic history, date and type of vaccine, onset and recovery of illness, number of hospitalization days, life-threatening illness, disability status, symptoms after vaccination, laboratory diagnostics after the onset of disease, etc. It has been found that some of these attributes are in text format (such as medical history, laboratory diagnostics, etc.) whereas, others are in numerical values (such as age, number of hospitalization days, etc.). Therefore, all these attributes have been converted into numerical values to have a better understanding of the features. The description of various features in the VAERS dataset has been illustrated in Table 1. Further, to have a better visualization of the dataset, the density distribution of features for the probable outcome (death or alive) has been presented in Fig. 1.

Further, a correlation plot has been obtained and illustrated in Fig. 2 to have a better understanding of available attributes on the outcome of the pandemic. It has been observed that attributes like A, S, H, HD, C, E, and L have a greater influence on the lethality of the ongoing pandemic whereas, others do not.

2.2 Proposed framework

The proposed methodology for efficient and accurate estimation of various complexities has been demonstrated in Fig. 3. This includes feature extraction from raw data by employing string matching, preprocessing and cleaning of the dataset, statistical analysis, sampling and feature estimation, classification, and performance evaluation. The overall framework has been subdivided into five compartments: Feature extraction, Preprocessing and Exploratory Data Analysis (EDA), Statistical test, ML models, and Performance parameters.

2.2.1 Feature extraction

The acquired raw dataset contains most of the important features in textual format however, for any analysis they must be converted into separate entities. Therefore, all the text data has been converted into attributes by employing the string matching technique. Further, it has been analyzed that the initial correlation plot (Fig. 2) does not indicate about the significant relationship of various attributes (especially, M and Al) and the current outbreak. However, the previously reported studies revealed that outbreak has a direct association with patient's medical and allergic history. Therefore, all the unique entries of disease in the medical history of patients have been counted. Though, in healthcare even the most scarce entity is of utmost importance and cannot be neglected as it may have an indispensable influence on a particular individual's life. However, because of the very large size of data and the required computational burden to process each and every disease, diseases with greater than 500 counts in medical history have been considered as attributes

Table 1 Description of cleaned VAERS dataset

SN	Features	Description	Range	Mean	Standard Deviation
1.	Age (A)	Age in Years	0–119	49.66	18.31
2.	Sex (S)	Gender information (0: Females, 1: Male and 2: Unknown)	0–2	0.31	0.51
3.	L_Threat (L)	Life-threatening illness	0–1	0.02	0.13
4.	Hospital (H)	Hospitalized	0–1	0.06	0.24
5.	Hospital Days (HD)	Number of days Hospitalized	0–120	0.19	1.51
6.	Disable (Di)	Disability status	0–1	0.01	0.12
7.	Num_Days (ND)	Number of days after vaccination	0–456	4.46	14.15
8.	Other Medicine (OM)	Currently using any other medicine	0–1	0.68	0.46
9.	Current Illness (C)	Illnesses at time of vaccination	0–1	0.11	0.32
10.	Prior vaccination (P)	Any prior vaccination	0–1	0.06	0.23
11.	Allergic History (Al)	Any known allergic history	0–1	0.35	0.48
12.	Medical History (M)	Number of diseases	0–16	0.51	0.99
13.	Emergency (E)	Visited emergency or not	0–1	0.14	0.34
14.	Birth Defect (BD)	Any known birth defect	0–1	0.00	0.02
15.	Died (D)	Died	0–1	0.01	0.12

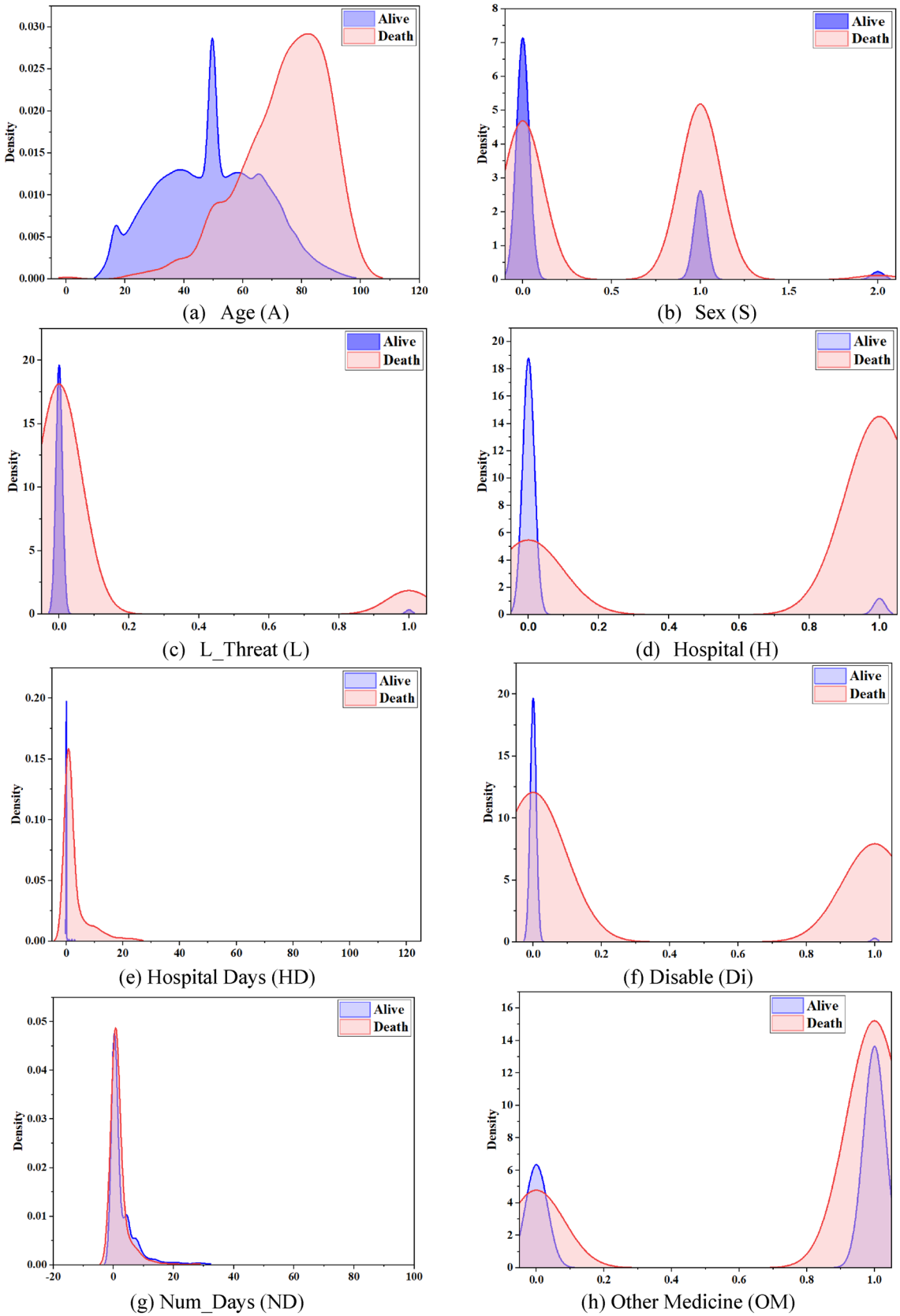


Fig. 1 Impact of attributes on the COVID-19 outbreak

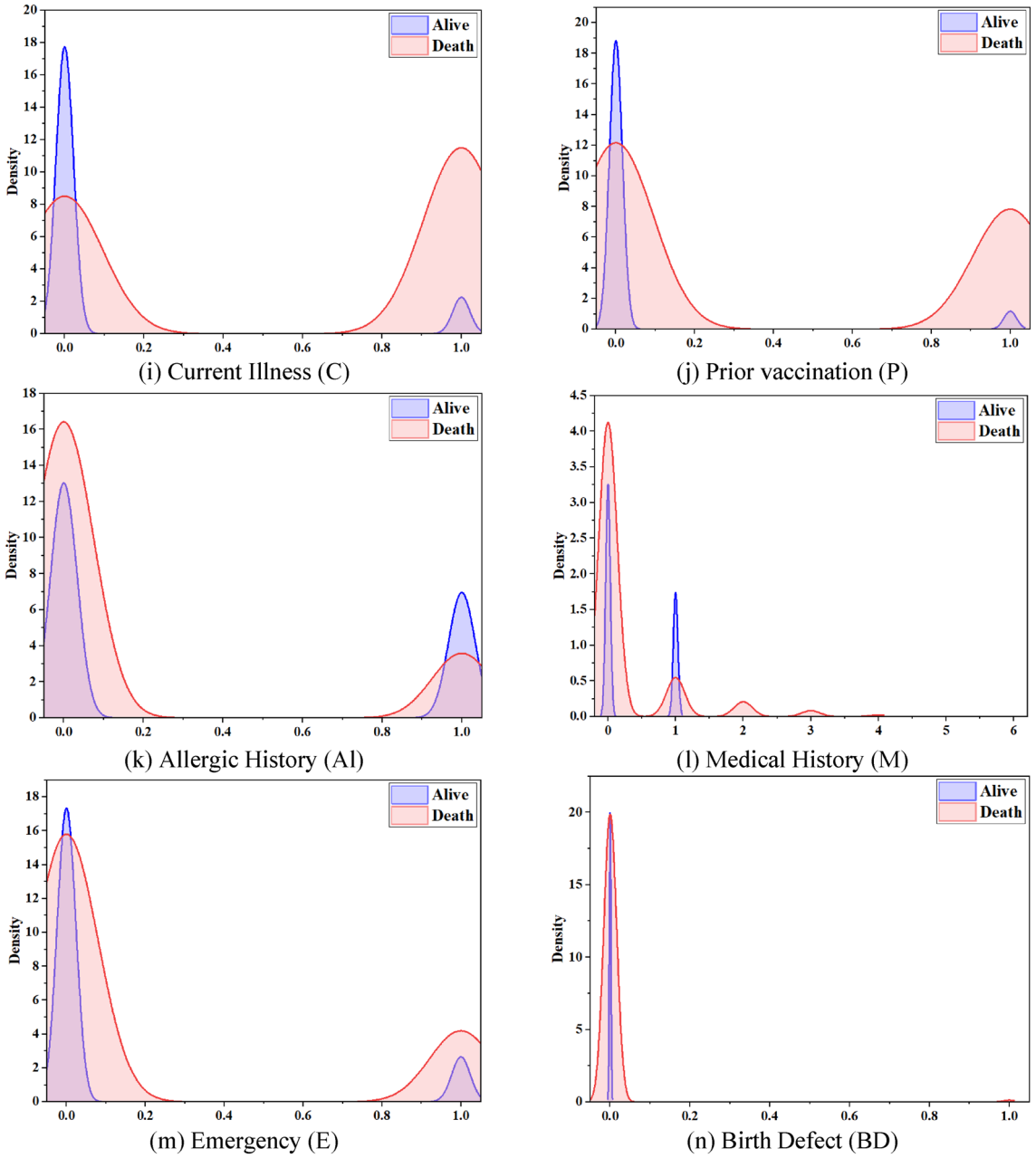


Fig. 1 (continued)

and the rest neglected. Therefore for the present investigation, 23 diseases (Diabetic Mellitus, Thyroid, Pain, Obesity, Migraine, Kidney Disease, Hypertension, Hyperlipidemia, High cholesterol, Heart Disease, GERD, Depression, Dementia, Covid-19, COPD, CANCER, Atrial, Fibrillation, Asthma, Arthritis, Anxiety, Anemia, Other Abnormalities, and Gout) from the medical history of patients have been also considered as attributes. In the raw dataset, besides the patient's historical data file, two other files have been found named as VAERSVAX and VAERSSYMPTOMS. The VAERSVAX contains information regarding vaccination such as type of vaccine, manufacturer, number of doses, vaccination site whereas, VAERSSYMPTOMPS contains facts and figures of adverse reactions of vaccines as reported by individuals. Based upon the frequency of symptoms, a total of 49 symptoms have been employed to develop the dataset used here for the analysis. Further, it has been found that all these three files are connected by a unique VAERS id. Therefore, after feature identification and extraction these files have been merged into a single file. In summary, the dataset being utilized in this analysis, have 85 different features for over 354 thousand samples.

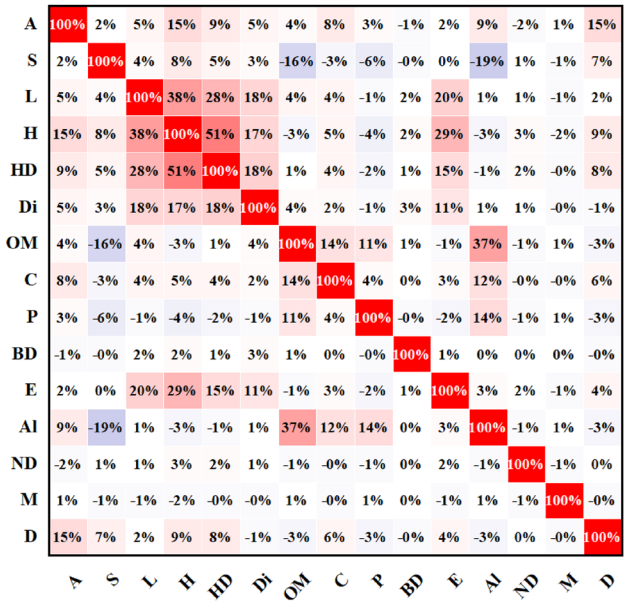


Fig. 2 Correlation plot between different attributes of VAERS dataset

2.2.2 Preprocessing and exploratory data analysis (EDA)

In this data-driven world, the outcome of any analysis vastly depends upon the quality of data being utilized. Therefore, preprocessing and EDA becomes the primary task for any data-driven investigation. In preprocessing, the various aspects (such as outliers, missing values, irrelevant values, replica, etc.) of the dataset have been examined whereas, EDA helps to understand the data by visualizing it. It has been observed that the data has many irrelevant and missing values of the attributes. Therefore, in this work, outlier rejection (*OR*) along with filling missing values (*MV*) has been employed to clean the dataset.

The values that are extremely deviated from other observations of any attribute have been referred to as outliers. They must be omitted from the dataset because in the data-driven analysis algorithms become very peculiar to the range and distribution of the attributes. In the present work, to detect and omit the outliers, quartiles have been employed and it has been mathematically presented as in Eq. (1) [23].

$$OR(k) = \begin{cases} k, & if Q_1 - 1.5 \times IQR \leq k \leq Q_3 + 1.5 \times IQR \\ reject, & otherwise \end{cases} \tag{1}$$

where, k symbolizes the presence of the feature vector in m-dimensional feature space ($k \in \mathbb{R}^m$). Q_1,Q_3, and IQR signifies the first, third, and interquartile range of the features such that $Q_1, Q_3, IQR \in \mathbb{R}^m$ respectively.

Further, it has been observed that many attributes have missing and unknown values however, neither these data

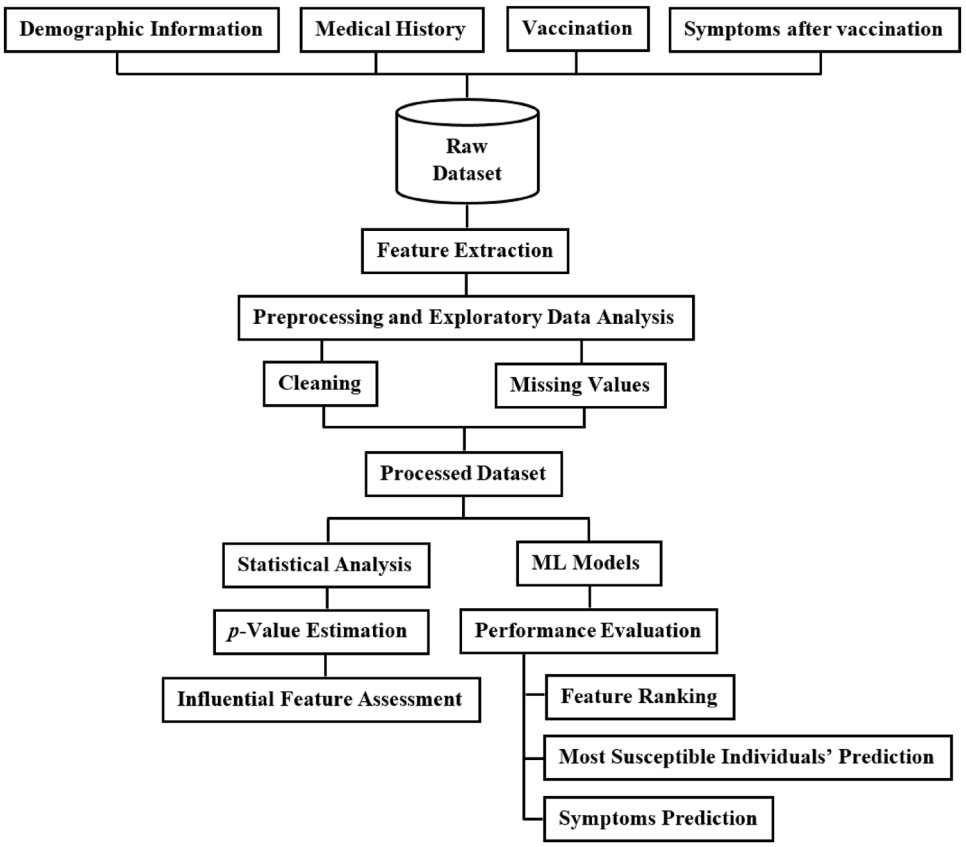


Fig. 3 Layout of the proposed methodology for estimating serious complications of the SARS-CoV-2 vaccine

points can be ignored because it will drastically reduce the size of the dataset nor can be filled by random and arbitrary values as it will affect the outcome. Therefore, to handle this issue median by target (death) methodology has been employed and formulated as in Eq. (2).

$$MV(k) = \begin{cases} median(k), & if k = missed\ or\ null\ or\ not\ avaliable \\ k, & otherwise \end{cases} \tag{2}$$

2.2.3 Statistical analysis

Generally, statistical methods have been employed to test the hypothesis in a dataset. Out of the many available statistical methods, this work employs a very popular chi-square (χ^2) test to find the association of extracted attributes in the breakout, even after vaccination with the confidence level of 95%. This test has been most commonly used to evaluate the test of independence. The test of independence analyses the association between various attributes of the dataset and the outcome (target). Therefore, it may help in the identification of the most crucial factors because of which the current pandemic becomes so deadly. Mathematically, the relationship between contributing factors in the outcome has been calculated by Eq. 3.

$$\chi^2 = \sum \frac{(O-E)^2}{E} \tag{3}$$

where, O and E refer to the original and expected outcomes. The attributes with higher values of χ^2 have been considered as independent whereas, smaller values of χ^2 represents the higher association. Therefore, attributes with p-value < 0.05 have been deliberated as crucial attributes of COVID-19.

2.2.4 ML models

The SARS-Cov-2 virus continuously changes its characteristics because of which it has many deadly mutations. Therefore, even after one year from the onset of the pandemic, researchers have tried to exactly identify the factors contributing to its lethality. In this work, most popular ML models such as Logistic Regression (LR), Random Forest (RF), Naive Bayes (NB), Light Grading Boosting Machine (LGBM), and Multilayer feed-forward Perceptron (MLP) have been employed for this purpose. These models have been selected because of their tremendous performance in various tasks [23, 24].

i. *Logistic Regression*

LR has been considered as one of the most simple yet, effective ML models. It has been employed to determine the relationship between the required dependent parameter and the other independent parameters in the dataset. Generally, it employs logistic function to estimate the probabilities of outcome and then, based upon the maximum likelihood estimation it categorizes the outcome. Mathematically, it has been expressed by Eq. 4 [24].

$$P(y) = \frac{1}{1+e^{-x}} \tag{4}$$

ii. *Random Forest*

RF ensembles a large number of individual and uncorrelated decision trees. Therefore despite a single prediction, the model predicts the output based upon the most likely prediction of several trees. It employs bootstrapping which has been done by random resampling of the training dataset [25]. This approach works extremely well for the higher dimensional dataset with acceptable accuracy. If the model has been represented by F_{RF}, the parameters of the training dataset by (x,y), pseudo-residuals by r, regularization by γ, and differentiable loss function by L then the procedure of RF can be divided into two parts: model initialization and computation of residuals, which has been mathematically represented by Eq. 5 and Eq. 6 [4].

$$F_0 = \underset{\gamma}{argmin} \sum L(y,\gamma) \tag{5}$$

$$r = \left[\frac{\delta L(y, F_{RF})}{\delta F_{RF}}\right] \tag{6}$$

iii. *Naive Bayes*

NB has been considered as a supervised classification algorithm that provides class-specific conditional probabilities based on Bayes rules. In NB, all the attributes have been assumed as independent of each other therefore, it is computationally inexpensive and easy to implement, yet powerful classification algorithm. Although the assumption of independent features seems impractical, it still produces results with fair accuracy. It first induces a distribution based upon Eq. 7 and then, the unknown instance has been classified by determining the maximum probability as expressed by Eq. 8 [26].

$$Pr\left(c, x_1, ..., x_n\right) = Pr(c) \prod_{i=1}^{n} Pr(x_i|c) \tag{7}$$

$$c^* = \underset{c}{argmax}\, Pr(c|x_1, ..., x_n) \tag{8}$$

where, c, m, and $Pr(c)$ represent class, the value of an attribute, and class prior probability. Both, $Pr(c|x_1,\ldots, x_n)$ and $Pr(x_i|c)$ signifies the conditional probabilities.

iv. *Light Grading Boosting Machine*

LGBM is a gradient boosting algorithm that employs a tree-based learning framework. As compared to other tree-based frameworks, it grew trees vertically (leaf-wise) whereas, others horizontally (level-wise). Therefore, it can reduce the losses more efficiently, and because of the lighter version, it can handle very large datasets with less computational complexities.

v. *Multilayer feed-forward Perceptron*

The MLP is a type of neural network which may have one input layer, multiple hidden layers, and one output layer. It is a mathematical model which aims to mimic the functioning of human brains. All these layers contain several artificial neurons that have been connected with each other in a unidirectional manner by mesh arrangements [27]. It employs activation functions through which the information from input to output has been processed. This work utilizes Keras' sequential library of TensorFlow for the development of the MLP model which has one input layer, four hidden layers, and one output layer as illustrated in Table 2 and Fig. 4.

The MLP produces an M-dimensional output vector for P-dimensional input vector subjected to $f(k) : \mathbb{R}^P \rightarrow \mathbb{R}^M$. Further, the output of each processing unit for n neurons can be mathematically represented by Eq. (9).

$$f(k) = \varnothing\left(\sum_n w_n k_n + b\right) \tag{9}$$

where, w_n, k_n, b, and $\varnothing$ represents the weights, input, bias, and the activation function respectively. This work employs L2 regularization to avoid overfitting and during the entire training the cost function as formulated by Eq. (10), has been minimized using Adam optimizer.

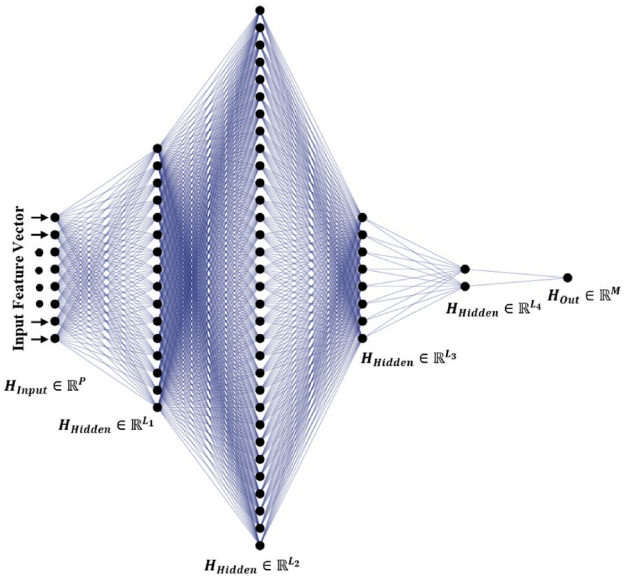


Fig. 4 Developed MLP architecture

$$L = \sum_{i=0}^{P} (y_i - \sum_{j=0}^{M} k_{ij} w_j)^2 + \lambda \sum_{j=0}^{M} {w_j}^2 \tag{10}$$

where, L and λ signify the cost function and regularization parameter respectively. Also, the hyperparameters for the developed model have been empirically (hit and trial) chosen and depicted in Table 3.

2.3 Simulation setup and evaluation metrics

The present work employs various APIs of Python and Keras for the programming and implementation of the developed models on Pycharm using Python 3.8 programming environment. The simulation has been done on Ubuntu 20.04 OS, Intel(R) Core (TM) i7-9750H CPU @ 2.60 GHz processor with 8 GB RAM and 4 GB NVIDIA GeForce GTX 1650 graphics card.

The present work employs Precision, Accuracy, Recall, and F_1 score to critically analyze the performance of the

Table 2 Developed MLP model summary

Layer (type)	Output Shape	Number of Parameters	Activation Function
Dense	(None, 8)	696	relu
Dense	(None, 16)	144	selu
Dense	(None, 32)	544	selu
Dense	(None, 8)	264	relu
Dense	(None, 2)	18	relu
Dense	(None, 1)	3	sigmoid

Table 3 Hyperparameters of the developed MLP model

SN	Hyperparameter	Description
	Hidden Layers	4
	Hidden Layer Neurons	16, 32, 8, 2
	Learning rate	0.01
	Epochs	1000
	Activation Function	Selu, Relu, and Sigmoid

developed models [28]. The Precision represents the accurateness of the predictions whereas, Recall has been used to represent the number of true positives that have been correctly identified. Accuracy depicts the percentage of true predictions and F_1 score indicates the balance between precision and recall. Mathematically, these performance evaluation metrics can be computed using Eqs. (11)-(14).

$$Precision = \frac{N_{TP}}{N_{TP} + N_{FP}} \tag{11}$$

$$Accuracy = \frac{N_{TP} + N_{TN}}{N_{TP} + N_{TN} + N_{FP} + N_{FN}} \times 100 \tag{12}$$

$$Recall = \frac{N_{TP}}{N_{TP} + N_{FN}} \tag{13}$$

$$F_1 score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{14}$$

where, N_{TP}, N_{FP}, N_{TN}, and N_{FN} signifies the number of true positives, false positives, true negatives, and false negatives respectively.

3 Result and discussion

The present work investigates the significant contributing factors because of which COVID-19 becomes so lethal. Therefore, the investigation has been carried out in three scenarios: (i) Based upon medical history only (ii) Based upon the reaction of vaccination only (iii) Based upon both medical history and adverse reaction. Further, the noteworthy contributing features have been identified by both statistical analysis and developed ML models (LR, RF, and LGBM). Then, all the developed ML models have been used to predict the important key outcomes of interest (Death, Hospitalization, and COVID-19 positive).

3.1 Scenario 1: based upon medical history only

This analysis has been done by utilizing the important features found in the medical history of the individual patients. The dataset has 38 attributes and by analyzing this, it has been revealed that out of the total 354,451 entries 5,062 people died between 1st January 2021 to 11th June 2021. Therefore, the mortality rate in that duration has been computed as 1.43%. Among them, only 1,274 (25.17%) and 60 (1.18%) have been identified as hospitalized and COVID-19 positive respectively. Also, during this period the total number of patients hospitalized and tested COVID-19 positive has been estimated as 21,926 (6.19%) and 3,486 (0.98%) of the total samples respectively. This analysis has been illustrated in Table 4 for better understanding. This result clearly indicates that the USA has passed its peak and in addition to that their large-scale vaccination program helped them in reducing the lethality of this outbreak.

It has been also computed that in this span only 0.19% (10) individuals died out of the total reported deaths who have been hospitalized and tested COVID-19 positive. This reveals that most of the SARS-CoV-2 positive people recovered themselves but some required hospitalization therefore, early hospitalization would be the key to reducing the lethality. However, admission of all patients in hospitals would increase the burden of already overcrowded hospitals. Therefore, the top 10 most significant contributing factors have been estimated by employing the chi-square statistical method and developed ML models which is as represented in Table 5. For Death and Hospitalized outcome, only LR identifies pre-existing diseases amongst top-10 contributing factors whereas, others identify medical history. However, for COVID-19, almost all the models estimate that pre-existing diseases play a crucial role. This clearly indicates that COVID-19 targets the immune system of humans and those with pre-existing diseases have lesser immunity. Therefore, the population with an earlier history of serious diseases has become the soft target of this deadly virus.

Further, the ML models have been employed to predict the possible outcome of interest. The performance parameters for all the models on the test dataset have been presented in Table 6. The best and worst value of precision have been achieved by LGBM (0.51) and MLP (0.02) for death prediction. Similar trends have been witnessed for hospitalization outcome. However, for predicting COVID-19, all the developed models faces difficulties in achieving acceptable precision. This may be because of the very small ratio (0.98%) of COVID-19 patients in the available dataset. Further, the best values of recall for death, hospitalization, and COVID-19 have been estimated by MLP as 0.93, 0.85, and 0.92 whereas, the worst by LR (0.01), NB (0.30), and LR (0.00) respectively. According to F_1 score, RF, LGBM, and NB dominate all the other models for death, hospitalization and COVID-19 by a minimum margin of 22.22%, 1.22% and 33.33% respectively. Also, except MLP, all the developed models have been found as sufficiently accurate.

These models have been also compared on the basis of region of convergence (ROC) curve and Precision-Recall

Table 4 Analysis of scenario 1

Total = 354,451	Died	Hospitalized	Covid positive
Died	5,062	1,274	60
Hospitalized	1,274	21,926	209
Covid positive	60	209	3,486

Table 5 Estimation of most significant attributes in scenario 1

Outcome	Rank	Models			
		Chi-square	LR	RF	LGBM
Death	(1)	A	A	A	ND
	(2)	S (M)	C	H	A
	(3)	HD	S (M)	S	HD
	(4)	C	COVID-19	C	S (M)
	(5)	E	H	P	OM
	(6)	OM	E	HD	C
	(7)	Al	High Cholesterol	E	Al
	(8)	P	Anxiety	Al	E
	(9)	L	Diabetes	OM	Al
	(10)	Di	HD	ND	L
Hospitalized	(1)	A	HD	HD	ND
	(2)	S (M)	L	E	A
	(3)	HD	E	A	E
	(4)	E	A	Di	S (M)
	(5)	L	S (M)	L	L
	(6)	Di	Di	S (M)	C
	(7)	C	C	P	Al
	(8)	P	Kidney Disease	OM	Di
	(9)	ND	Heart Disease	ND	HD
	(10)	Al	Birth Defect	C	P
COVID-19 positive	(1)	ND	Pain	ND	ND
	(2)	Other Abnormalities	Heart Disease	Pain	A
	(3)	Pain	Other Abnormalities	Hyperlipidemia	Heart Disease
	(4)	Heart Disease	Cancer	Kidney Disease	Hyperlipidemia
	(5)	Dementia	Hyperlipidemia	Heart Disease	Kidney Disease
	(6)	Hyperlipidemia	Kidney Disease	Other Abnormalities	Cancer
	(7)	Kidney Disease	Diabetes	Diabetes	Pain
	(8)	Cancer	Atrial Fibrillation	Cancer	Diabetes
	(9)	Atrial Fibrillation	Dementia	Dementia	OM
	(10)	Diabetes	Thyroid	Anemia	Other Abnormalities

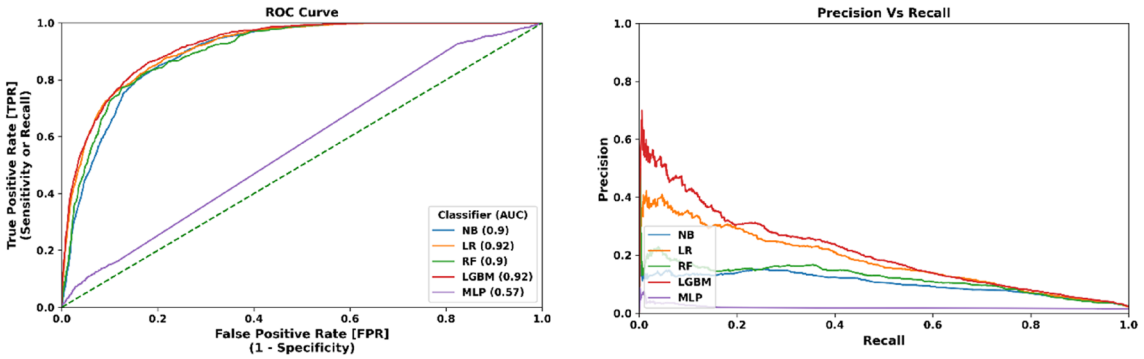
Table 6 Outcome prediction in scenario 1

SN	Parameter	Outcome	ML models				
			LR	RF	NB	LGBM	MLP
1	Precision	Death	0.41	0.06	0.15	0.51	0.02
		Hospitalized	0.97	0.60	0.65	0.97	0.50
		COVID-19	0	0.01	0.13	0.08	0.01
2	Recall	Death	0.01	0.84	0.06	0.04	0.93
		Hospitalized	0.71	0.81	0.30	0.72	0.85
		COVID-19	0	0.52	0.03	0.01	0.92
3	Accuracy	Death	0.98	0.80	0.98	0.99	0.18
		Hospitalized	0.98	0.96	0.95	0.98	0.94
		COVID-19	0.99	0.71	0.99	0.99	0.09
4	F_1 score	Death	0.03	0.11	0.09	0.08	0.03
		Hospitalized	0.82	0.69	0.41	0.83	0.63
		COVID-19	0	0.03	0.04	0.01	0.02

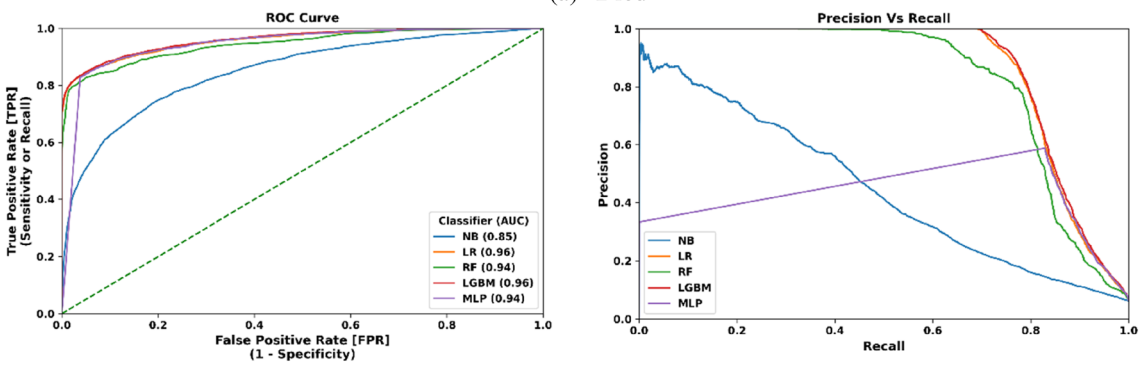
(PR) curve for all the mentioned outcomes. These curves have been illustrated in Fig. 5 for more clarity which revealed the effectiveness of LGBM for most of the outcomes. The average area under the curve (AUC) obtained by LGBM, considering all the three outcomes, has been computed as 0.84 whereas, 0.83, 0.83, 0.80, and 0.67 for LR, RF, NB, and MLP respectively. Therefore, LGBM outperforms LR, RF, NB, and MLP by a margin of 1.20%, 1.20%, 5.00%, and 25.37% respectively.

3.2 Scenario 2: based upon the reaction of vaccination only

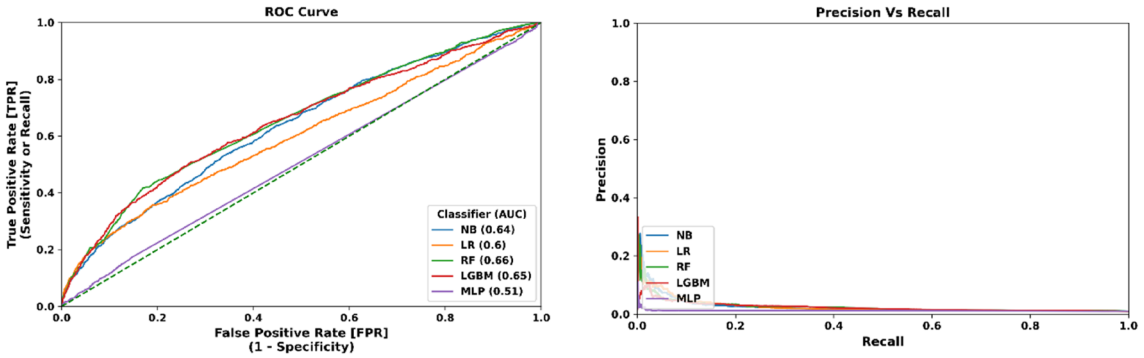
To analyze the adverse reaction of the vaccine, the data available in VAXSYMPTOMS has been employed. Out of many available symptoms, most repeatedly occurred symptoms have been identified and then, based upon this, the dataset consisting of 64 attributes has been utilized. On examining the reaction profile of the patient' it has been found that compared to the total number of deaths reported in the medical history, less people have died who have been vaccinated at least once. This indicates the effectiveness of the vaccines being used against SARS-CoV-2. Further, for quantitative analysis, the obtained counts of the interesting



(a) Died



(b) Hospitalized



(c) COVID-19 positive

Fig. 5 Scenario 1: ROC and PR curve for **a** Death **b** Hospitalization **c** COVID-19 positive

features have been presented in Table 7. It has been analyzed that the number of patients who tested positive for COVID-19 has been increased significantly with an increment of 2,467 (70.77%) which may be because the vaccines developed immunities in the body and during that development phase individuals might get the symptoms responsible for the SARS-CoV-2 virus. Further, out of the total deaths in this span, only 1,097 (24.34%) people have been hospitalized. These hospitalized patients constitute about 5.00% of the total people being hospitalized. This again indicates that the fatality rate associated with the SARS-CoV-2 virus can be further reduced provided the right candidate gets hospitalized at the right time. Also, higher deaths among COVID-19 positive patients have been witnessed but this higher rate is because of the adverse reaction of vaccination or due to any other reason (medical history) is still an open question.

Further, the most influential parameters being identified by various models and giving rise to features of interest have been depicted in Table 8. It has been observed that apart from existing medical conditions, the most common adverse reaction because of which people admitted in hospitals is COVID-19, chest pain, thrombosis, and dyspnoea as estimated by chi-square, LR, RF, and LGBM respectively. Further, A, Cough, and ND have been identified as the most probable causes of SARS-CoV-2 infection even after the vaccination. Apart from the medical history, the most

Table 7 Analysis of scenario 2

Total = 354,451	Died	Hospitalized	Covid positive
Died	4,506	1,097	268
Hospitalized	1,097	21,926	1,443
Covid positive	268	1,443	5,953

Table 8 Estimation of most influential attributes in scenario 2

Outcome	Rank	Models			
		Chi-square	LR	RF	LGBM
Death	(1)	A	Birth Defect	A	ND
	(2)	S (M)	Abortion Spontaneous	S (M)	A
	(3)	HD	Intensive Care	Headache	HD
	(4)	Cardiac Arrest	Loss of Consciousness	Pain	S (M)
	(5)	Intensive Care	L	Rash	Pain
	(6)	Pain	Hyperhidrosis	Dizziness	E
	(7)	C	Feeling Abnormal	Cardiac Arrest	OM
	(8)	Headache	Muscular Weakness	Chills	Al
	(9)	Chills	Chest pain	Injection Site Swelling	C
	(10)	E	Injection Site Pain	Injection Site Erythema	L
Hospitalized	(1)	A	HD	E	A
	(2)	COVID-19	E	HD	L
	(3)	Intensive Care	L	L	S (M)
	(4)	Dyspnoea	Chest Pain	A	OM
	(5)	Injection Site Pain	COVID-19	Thrombosis	E
	(6)	S (M)	Thrombosis	Dyspnoea	Dyspnoea
	(7)	Headache	Dyspnoea	Di	COVID-19
	(8)	Chills	Intensive Care	S (M)	Thrombosis
	(9)	D	A	Injection Site Pain	Pain
	(10)	Condition Aggravated	S (M)	Injection Site Erythema	HD
COVID-19 positive	(1)	A	Cough	Cough	ND
	(2)	HD	Intensive Care	Rash	A
	(3)	Cough	Cardiac Arrest	OM	HD
	(4)	OM	Diarrhea	HD	C
	(5)	Pain	Dyspnoea	Dizziness	OM
	(6)	Intensive Care	Malaise	Injection Site Pain	Al
	(7)	Dyspnoea	OM	Injection Site Erythema	Cough
	(8)	Rash	C	Injection Site pruritus	Pain
	(9)	Dizziness	Al	Pain	S (M)
	(10)	Injection Site Pain	Increased Blood Pressure	Arthralgia	Pyrexia

significant contributing symptom for death as acknowledged by chi-square, LR, and RF have been observed as cardiac arrest, birth defect, and headache whereas, LGBM did not consider any of the symptoms as influential for death.

Again, ML models have been developed to predict the outcomes of interest. It has been observed that for predicting death, LR outperforms other models and achieved a precision of 0.84 whereas, both RF and MLP dominate by achieving a recall of 0.95. In terms of accuracy and F_1 score, all the developed ML models except RF, have been able to attain acceptable values and therefore, perform satisfactorily. Similarly, all the developed ML models attain fairly good values of these metrics and therefore, have been considered to predict the individual's required hospital assistance after vaccination. However, for COVID-19 prediction the results of these developed models differ significantly. The LGBM produces the maximum value of F_1 score (0.65) with precision, recall, and accuracy of 0.70, 0.61, and 0.99 respectively whereas, the worst F_1 score (0.30) has been obtained by RF. The obtained results for all the performance parameters have been listed in Table 9.

To further analyze the prediction capability of these developed models, they have been critically examined on the basis of ROC and PR curve as represented in Fig. 6. The average AUC obtained by LR, RF, NB, LGBM, and MLP has been computed as 0.97, 0.96, 0.94, 0.97, and 0.97. Therefore in terms of AUC, all the developed models perform satisfactorily however, based on the PR curve LGBM proves its effectiveness.

3.3 Scenario 3: based upon both medical history and adverse reaction

Generally, it has been considered that the existing medical condition of individuals has a direct impact on any post-vaccination symptoms. Therefore, the post-vaccine symptoms cannot be considered independent of the patient's medical history. Consequently, another dataset has been framed containing all the crucial yet common features available in the raw dataset. As mentioned earlier, this dataset contains a total of 85 attributes with over 354 thousand samples. Although a similar kind of study has been found in the literature but, it employs very few samples [10]. Therefore, as per the best knowledge of the authors, this study has been considered as one of the most rigorous analyses for COVID-19 based upon the medical history of patients and various symptoms generated after vaccination as reported by the vaccinated individuals.

After appending all the three files available in the raw dataset, it has been exposed that the dead status of some individuals has been only mentioned in one file. Therefore by carefully including missing data of individuals, a total of 5,327 (1.50%) have been found died out of which 1,363 (25.59%) and 340 (6.38%) have hospitalized and tested SARS-CoV-2 positive respectively. This again reveals that a large number of the population did not get the basic treatment and if they have been admitted, the outcome would be different. Also, only 1,443 (6.58%) SARS-CoV-2 positive patients have been admitted to the hospitals as compared to 21,926 total hospitalizations. This reflects that a large portion of people has been admitted because of several other reasons and not because of COVID-19 and these results have been illustrated by Table 10.

Further, the most dominant feature has been again investigated and illustrated in Table 11. It has been found that LR estimated cardiac arrest as the dominant feature whereas, others identified A because of which people died in the duration of this study. Also, most of the models have considered HD as a primary reason because of which people are admitted to the hospitals. Similarly, the most common reaction of vaccination has been identified as a cough that majorly contributes to COVID-19.

Table 9 Performance analysis of developed ML models in scenario 2

SN	Parameter	Outcome	ML models				
			LR	RF	NB	LGBM	MLP
1	Precision	Death	0.84	0.25	0.73	0.81	0.61
		Hospitalized	0.96	0.78	0.53	0.96	0.52
		COVID-19	0.66	0.18	0.50	0.70	0.19
2	Recall	Death	0.93	0.95	0.92	0.90	0.95
		Hospitalized	0.73	0.77	0.48	0.75	0.88
		COVID-19	0.48	0.81	0.66	0.61	0.86
3	Accuracy	Death	0.99	0.96	0.99	0.99	0.99
		Hospitalized	0.98	0.97	0.94	0.98	0.94
		COVID-19	0.99	0.94	0.98	0.99	0.94
4	F_1 score	Death	0.93	0.39	0.81	0.85	0.74
		Hospitalized	0.83	0.78	0.50	0.84	0.65
		COVID-19	0.55	0.30	0.57	0.65	0.31

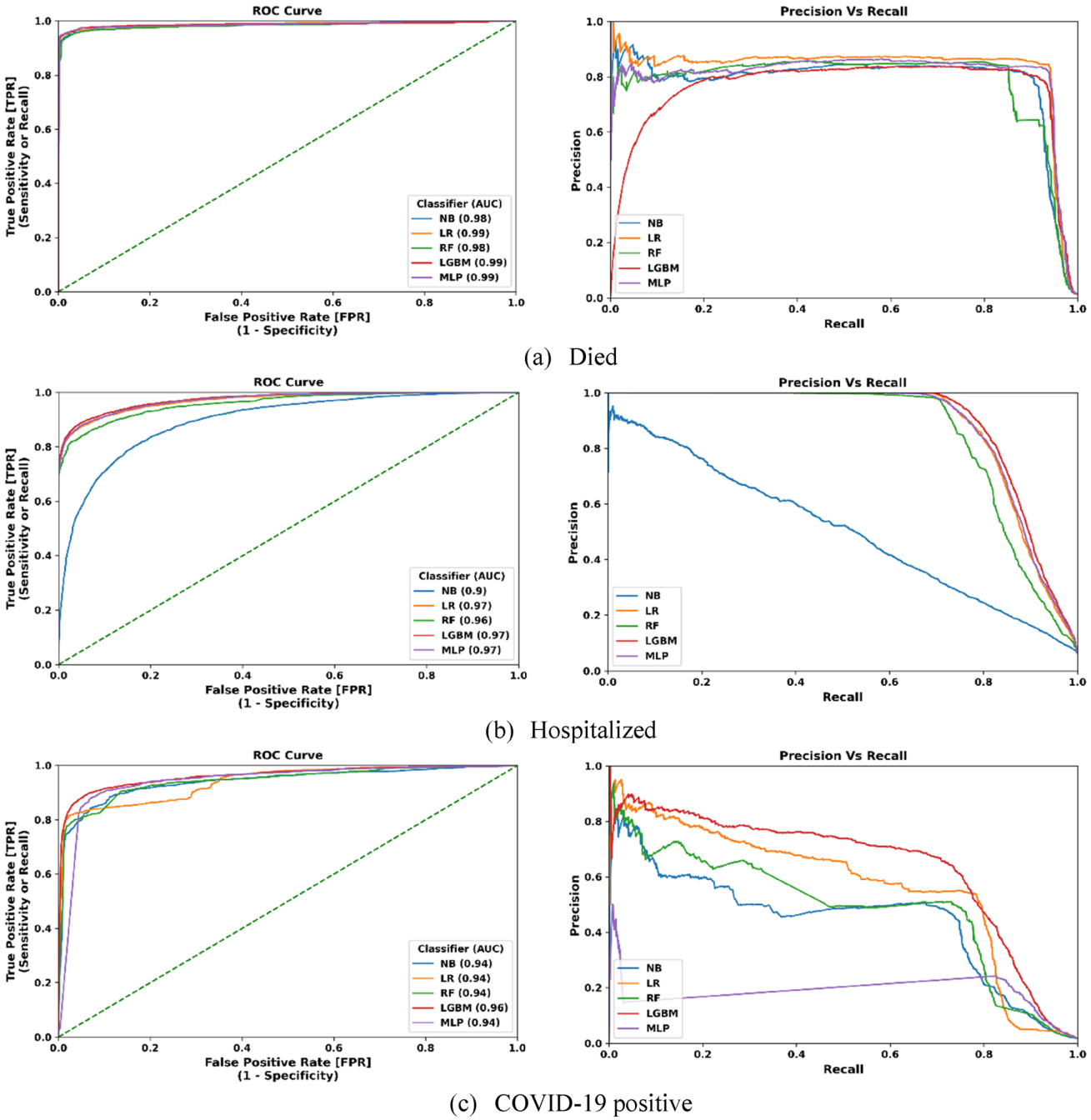


Fig. 6 Scenario 2: ROC and PR curve for **a** Death **b** Hospitalization **c** COVID-19 positive

Table 10 Analysis of scenario 3

Total = 354,451	Died	Hospitalized	Covid positive
Died	5.327	1,363	340
Hospitalized	1,363	21,926	1,443
Covid positive	340	1,443	5,953

Finally, the performance of all the developed ML models has been exhaustively analyzed and tabulated in Table 12. It has been observed that LGBM dominates all the other developed models and achieved state-of-the-art performance for most of the outcomes. It provides extraordinary results for the prediction of death over LR, RF, NB, and MLP by a margin of 36.84%, 116.67%, 13.04%, and 136.36% in terms of F_1 score respectively. However, slightly lower value of precision as compared to LR has been also computed.

Table 11 Analysis of most dominant feature in scenario 3

Outcome	Rank	Models			
		Chi-square	LR	RF	LGBM
Death	(1)	A	Cardiac Arrest	A	A
	(2)	S (M)	A	H	S (M)
	(3)	HD	C	Pain	C
	(4)	Cardiac Arrest	S (M)	Headache	Pain
	(5)	Intensive Care	Abortion Spontaneous	S (M)	H
	(6)	Pain	Dyspnoea	Chills	ND
	(7)	C	Malaise	Rash	HD
	(8)	Headache	Intensive Care	HD	OM
	(9)	E	Chest Pain	Dizziness	E
	(10)	Chills	Feeling Abnormal	Injection Site Pain	Cardiac Arrest
Hospitalized	(1)	A	HD	HD	HD
	(2)	S (M)	E	E	L
	(3)	HD	L	L	S (M)
	(4)	Cardiac Arrest	Chest Pain	A	OM
	(5)	Intensive Care	Thrombosis	Thrombosis	Dyspnoea
	(6)	Headache	COVID-19	Dyspnoea	E
	(7)	E	Dyspnoea	Injection Site pruritus	Al
	(8)	Chills	Intensive Care	Injection Site Pain	A
	(9)	COVID-19	A	Di	Pain
	(10)	Dyspnoea	S (M)	S (M)	ND
COVID-19 positive	(1)	A	Cough	Rash	Cough
	(2)	HD	C	Cough	ND
	(3)	Cough	S (M)	OM	Pain
	(4)	OM	Dyspnoea	Dizziness	HD
	(5)	Pain	Malaise	Injection Site Pain	A
	(6)	Intensive Care	Diarrhea	Pain	OM
	(7)	Dyspnoea	Pyrexia	A	C
	(8)	Rash	A	Injection Site Erythema	S (M)
	(9)	Dizziness	Intensive Care	Injection Site Swelling	E
	(10)	Injection Site Pain	High Cholesterol	HD	Dizziness

Table 12 Performance evaluation of developed ML models in scenario 3

SN	Parameter	Outcome	ML models				
			LR	RF	NB	LGBM	MLP
1	Precision	Death	0.68	0.07	0.18	0.65	0.06
		Hospitalized	0.96	0.76	0.54	0.96	0.49
		COVID-19	0.69	0.16	0.45	0.64	0.05
2	Recall	Death	0.11	0.86	0.32	0.17	0.82
		Hospitalized	0.72	0.79	0.49	0.75	0.88
		COVID-19	0.25	0.88	0.59	0.49	0.74
3	Accuracy	Death	0.98	0.81	0.97	0.98	0.80
		Hospitalized	0.98	0.97	0.94	0.98	0.94
		COVID-19	0.98	0.92	0.94	0.96	0.74
4	F_1 score	Death	0.19	0.12	0.23	0.26	0.11
		Hospitalized	0.83	0.78	0.51	0.84	0.63
		COVID-19	0.36	0.27	0.51	0.55	0.09

Further, these models, except RF and MLP, struggles to achieve promising values of recall however, acceptable accuracy has been acquired by all the developed models. For the prediction of the need for hospitalization, LGBM outpaces other models by a minimum margin of 1.20% in terms of F_1 score. Further, LR shares the similar values of precision and accuracy with LGBM whereas, MLP outclasses LGBM in terms of recall by a margin of 11.39%. Similarly, the LGBM successfully predicted most of the samples for COVID-19 outcome in the test dataset with descent F_1 score and promising accuracy.

Based upon the ROC and PR curve (Fig. 7) also, it has been determined that LGBM estimates the required outcome with significant values of precision, recall, accuracy, and F_1 score whereas, RF and MLP struggle the most.

Therefore, on the basis of the above analysis it has been revealed that in general, the ML model developed by employing LGBM provides significant prediction in all the cases as compared to all the other developed models. However, as the data has been found as highly biased towards negative class (only 1.50% of positive samples for any outcome and scenario) therefore, in many cases, the models become overfit and unable to produce satisfactory predictions. It has been also discovered that the number of deaths does not largely associate with COVID-19 during the studied interval, at least in the USA. Therefore, with the very high vaccination rate, the USA has passed its peak and recovering at a decent pace. Also, during the present investigation, no serious adverse reaction of the vaccine has been found therefore, currently the existing medical conditions of individuals resulted in the random breakout, not the vaccines. Surprisingly, no model considers diabetes, hypertension like diseases as influential factors in any of the above-mentioned scenarios. This clearly breaks the myth that these people are most susceptible to the adverse reaction of the vaccine and encourages them to come forward for vaccination programs.

4 Conclusion

In the present work, a rigorous analysis of the ongoing pandemic has been accomplished by employing both statistical analysis and ML frameworks in three parts: only historical data, only post-vaccination symptoms, and incorporating both historical and post-vaccination symptoms with more than 354 thousand samples. The major findings of this work have been summarized as:

(i) The people in the age group of 50–70 have been found as most susceptible to the SARS-CoV-2.
(ii) The male population has been identified as more vulnerable than to female population.
(iii) The population with a history of life-threatening diseases such as cardiac diseases, allergies, dyspnoea, etc. should be vaccinated in close observation.
(iv) The existing medical history worked as a catalyst because of which SARS-CoV-2 exploded in every corner of the globe. However, in rare cases, extreme adverse reactions of the vaccines have been also noticed which cannot be ignored. Therefore, it requires further future efforts.
(v) The most common post-vaccination symptoms have been identified as large hospital stays, rash, injection site discomfort, dizziness, dyspnoea, chills, headache, etc. Most of these major symptoms have been found normal and do not indicate towards any sign of serious and immediate concern.
(vi) The developed ML models perform brilliantly especially, when the medical history along with symptoms have been used for prediction. This may help the policymakers in identifying the most vulnerable population and therefore, priority-based administration of the vaccine.

Though, the present work enlightens various aspects of COVID-19 yet, influenced by the USA population. Therefore, before generalizing, more focused studies are required which will be done in the future on the availability of the required dataset. Further, deep learning models such as recurrent neural networks may also be employed to extract more hidden patterns in order to better understand and manage the COVID-19 dynamics and therefore, enhance the general acceptability of its vaccines.

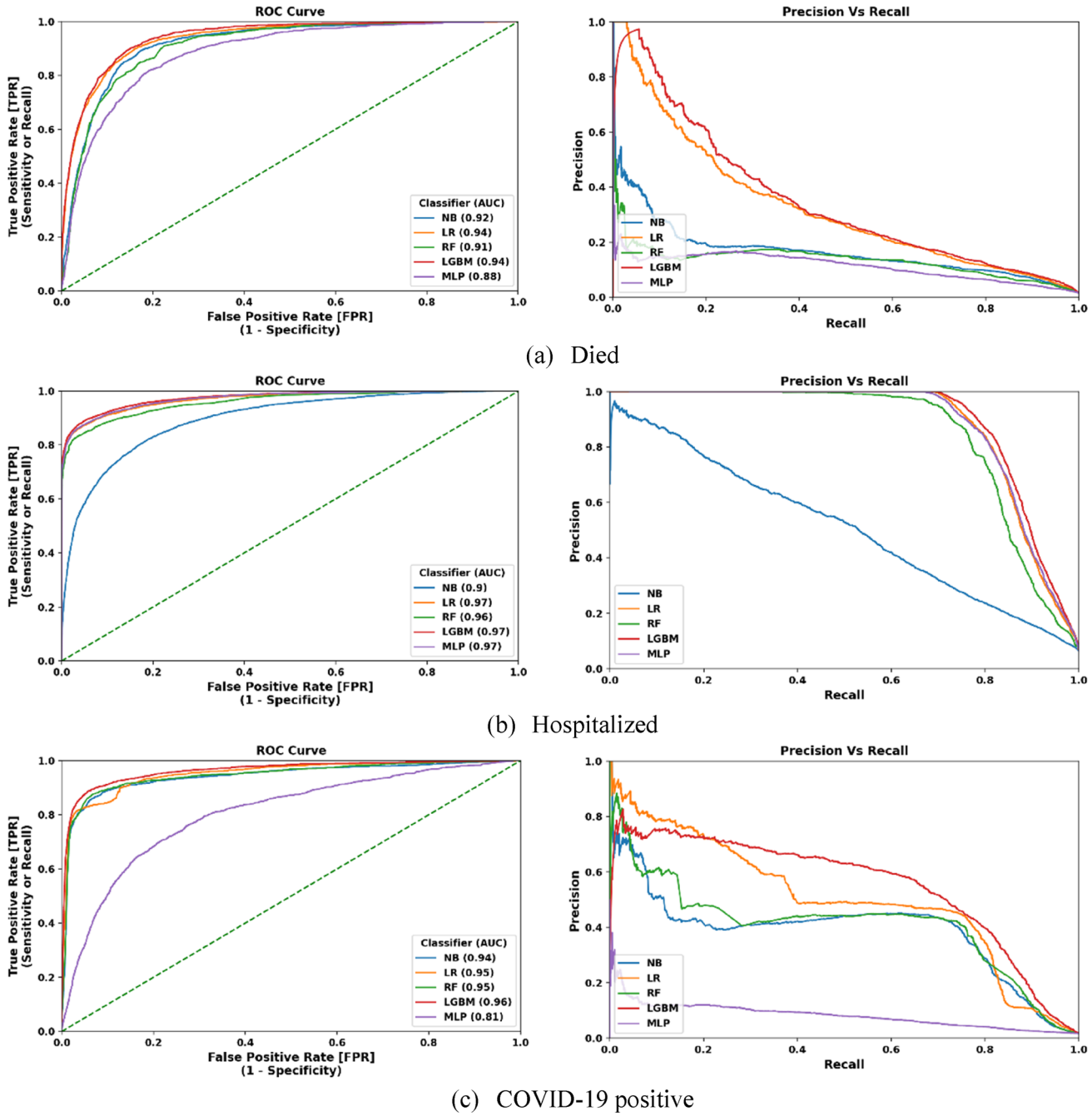


Fig. 7 Scenario 3: ROC and PR curve for **a** Death **b** Hospitalization **c** COVID-19 positive

Funding None.

Conflicts of interest The authors declare no conflict of interest.

References

1. Gupta H, Kumar S, Yadav D et al (2021) Data analytics and mathematical modeling for simulating the dynamics of COVID-19 epidemic—a case study of India. Electronics 10:127. https://doi.org/10.3390/electronics10020127
2. Shereen MA, Khan S, Kazmi A et al (2020) COVID-19 infection: origin, transmission, and characteristics of human coronaviruses. J Adv Res 24:91–98

3. Zhu Z, Lian X, Su X et al (2020) From SARS and MERS to COVID-19: a brief summary and comparison of severe acute respiratory infections caused by three highly pathogenic human coronaviruses. Respir Res 21:1–14
4. Ahamad MM, Aktar S, Rashed-Al-Mahfuz M et al (2020) A machine learning model to identify early stage symptoms of SARS-Cov-2 infected patients. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2020.113661
5. COVID Live Update: 181,190,692 Cases and 3,925,285 Deaths from the Coronavirus - Worldometer. https://www.worldometers.info/coronavirus/. Accessed 26 Jun 2021
6. Dash S, Chakraborty C, Giri SK, Pani SK (2021) Intelligent computing on time-series data analysis and prediction of COVID-19 pandemics. Pattern Recognit Lett 151:69–75. https://doi.org/10.1016/J.PATREC.2021.07.027
7. Rahman A, Chakraborty C, Anwar A et al (2021) SDN–IoT empowered intelligent framework for industry 4.0 applications during COVID-19 pandemic. Cluster Comput 3:1–18. https://doi.org/10.1007/S10586-021-03367-4/TABLES/5
8. Tracking SARS-CoV-2 variants. https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/. Accessed 26 Jun 2021
9. COVID-19 vaccine tracker and landscape. https://www.who.int/publications/m/item/draft-landscape-of-covid-19-candidate-vaccines. Accessed 27 Jun 2021
10. Ahamad MM, Aktar S, Uddin MJ, et al Adverse effects of COVID-19 vaccination: machine learning and statistical approach to identify and classify incidences of morbidity and post-vaccination reactogenicity. https://doi.org/10.1101/2021.04.16.21255618
11. COVID-19 vaccine tracker. https://vac-lshtm.shinyapps.io/ncov_vaccine_landscape/. Accessed 26 Jun 2021
12. Russo AG, Decarli A, Valsecchi MG (2021) Strategy to Identify priority groups for COVID-19 vaccination: a population based cohort study. Vaccine 39:2517–2525. https://doi.org/10.1016/j.vaccine.2021.03.076
13. Kumar VM, Pandi-Perumal SR, Trakht I, Thyagarajan SP (2021) Strategy for COVID-19 vaccination in India: the country with the second highest population and number of cases. npj Vaccines 6:1–7
14. Coronavirus (COVID-19) Vaccinations - Statistics and Research - Our World in Data. https://ourworldindata.org/covid-vaccinations. Accessed 27 Jun 2021
15. Hafizh M, Badri Y, Mahmud S, et al (2021) COVID-19 Vaccine willingness and hesitancy among residents an qatar: a quantitative analysis based an machine learning. https://doi.org/10.1080/10911359202119736 42.
16. Solís Arce JS, Warren SS, Meriggi NF et al (2021) (2021) COVID-19 vaccine acceptance and hesitancy in low- and middle-income countries. Nat Med 278(27):1385–1394. https://doi.org/10.1038/s41591-021-01454-y
17. Eastwood K, Durrheim DN, Jones A, Butler M (2010) Acceptance of pandemic (H1N1) 2009 influenza vaccination by the australian public. Med J Aust 192:33–36. https://doi.org/10.5694/j.1326-5377.2010.tb03399.x
18. Troiano G, Nardi A (2021) Vaccine hesitancy in the era of COVID-19. Public Health 194:245–251
19. Shimabukuro TT, Cole M, Su JR (2021) Reports of anaphylaxis after receipt of mRNA COVID-19 vaccines in the US-December 14, 2020-January 18, 2021. JAMA - J Am Med Assoc 325:1101–1102
20. Haas EJ, Angulo FJ, McLaughlin JM et al (2021) Impact and effectiveness of mRNA BNT162b2 vaccine against SARS-CoV-2 Infections and COVID-19 cases, hospitalisations, and deaths following a nationwide vaccination campaign in Israel: an observational study using national surveillance data. Lancet 397:1819–1829. https://doi.org/10.1016/S0140-6736(21)00947-8
21. Kumar S, Yadav D, Gupta H et al (2021) A Novel Yolov3 algorithm-based deep learning approach for waste segregation: towards smart waste management. Electron 10:1–20. https://doi.org/10.3390/electronics10010014
22. VAERS - Data. https://vaers.hhs.gov/data.html. Accessed 28 Jun 2021
23. Gupta H, Varshney H, Sharma TK et al (2021) Comparative performance analysis of quantum machine learning with deep learning for diabetes prediction. Complex Intell Syst 1:3. https://doi.org/10.1007/s40747-021-00398-7
24. Khanday AMUD, Rabani ST, Khan QR et al (2020) Machine learning based approaches for detecting COVID-19 using clinical text data. Int J Inf Technol 12:731–739. https://doi.org/10.1007/s41870-020-00495-9
25. Sarica A, Cerasa A, Quattrone A (2017) Random Forest Algorithm for The Classification of Neuroimaging Data in Alzheimer's Disea: A Systematic Review. Front. Aging Neurosci. 9
26. Salmi N, Rustam Z Naïve Bayes Classifier Models for Predicting the Colon Cancer. https://doi.org/10.1088/1757-899X/546/5/052068
27. Miller AS, Blott BH, Hames TK (1992) Review of Neural Network Applications in Medical Imaging and Signal Processing. Med Biol Eng Comput 30:449–464
28. Gupta H, Verma OP (2021) Monitoring and surveillance of urban road traffic using low altitude drone images: a deep learning approach. Multimed Tools Appl. https://doi.org/10.1007/s11042-021-11146-x