Predicting COVID-19 post-vaccination mortality in persons with cardiovascular disease risk factors using explainable AI*

Taiwo Kolajo[1,2,†], Olawande Daramola[1,*,†]

[1]University of Pretoria, Hatfield Campus, Pretoria, South Africa
[2]Federal University Lokoja, PMB 1154, Lokoja, Nigeria

Abstract

Coronavirus disease (COVID-19) vaccination was adopted worldwide due to the advent of the COVID-19 pandemic in 2019. However, many post-vaccination adverse events, such as death and severe illness, were reported. So far, the specific case of post-vaccination adverse events pertaining to persons with cardiovascular risk factors and comorbidities has not been explored empirically, which limits the understanding of the underlying causes of adverse reactions to vaccination by this category of persons. This paper explored Explainable AI (XAI) to identify the critical determinants of post-vaccination mortality in persons with cardiovascular risk factors. To do this, we extracted 16657 records of persons with cardiovascular risk factors from the Vaccine Adverse Event Reporting System (VAERS) open dataset (from 2020 to May 2024). We then employed predictive modelling using a process that involved four stages. The first stage involved extracting relevant data from VAERS, data preprocessing, and handling class imbalance. In the second stage, we conducted a comparative performance evaluation of seven machine learning (ML) algorithms (Logistic Regression — LR, K-Nearest Neighbour — KNN, Deep Multilayer Perceptron — Deep MLP, Support Vector Machines — SVM, Random Forest — RF, Extreme Gradient Boosting — XGBoost, and Categorical Boost — CatBoost). In the third stage, we compared the performance of two stacked ensemble models composed of six base models, using Catboost and XGBoost as the meta-learners in each case. The fourth stage involved using SHAPley Additive Explanations (SHAP) to interpret the predictions of the best-performing model. The result showed that CatBoost has the best performance among the base ML models (Acc = 0.96, F1=0.96, AUC = 0.96), while Stacked ensemble - XGBoost had the best overall performance (Acc = 0.96, F1=0.96, AUC = 0.99). Also, we found the important predictors of post-vaccination mortality in persons with cardiovascular comorbidity. Generally, older age, a higher number of days spent in the hospital increases the risk of mortality, while the absence of current illness, life-threatening condition, hospitalization, prolonged hospitalization, disability, birth defect, doctor visit, and emergency care; and vaccination dose completion will enhance the probability of survival. However, the presence of diabetes, high cholesterol, high blood pressure, and other illnesses increases the risk of mortality. This study's findings contribute to a better understanding of critical factors that could enable better handling of adverse events related to post-vaccination in persons with cardiovascular disease comorbidity.

Keywords

post-vaccination mortality, COVID-19, cardiovascular disease risk factors, explainable AI

1. Introduction

Cardiovascular disease poses a serious risk to the quality of life of vulnerable populations, particularly middle-aged and older individuals. There is a long history of significant morbidity, death, and financial losses associated with cardiovascular illnesses and their consequences [1][2]. Precision diagnosis is difficult to achieve and places a significant strain on the healthcare system and the economy due to the prevalence of various chronic diseases.

In December 2019, China saw the start of the global pandemic known as Coronavirus disease (COVID-19), which was brought on by SARS-CoV-2. Over 776 million confirmed cases and 7 million deaths were reported worldwide by September 2024 [3]. Global health and the economy were significantly impacted

HC@AIxIA 2024: 3rd AIxIA Workshop on Artificial Intelligence For Healthcare
*Corresponding author.
†These authors contributed equally.
taiwo.kolajo@up.ac.za;taiwo.kolajo@fulokoja.edu.ng (T. Kolajo); wande.daramola@up.ac.za (O. Daramola)
0000-0001-6780-2495 (T. Kolajo); 0000-0001-6430-078X (O. Daramola)

by the COVID-19 pandemic [4][5][6]. Younger people without significant underlying medical conditions may also experience fatal complications. However, elderly patients with underlying chronic disorders like cardiovascular disease are thought to be at greater risk for death due to immunocompromised conditions [7][8].

Among the many instruments at our disposal for lowering morbidity and death, vaccinations are by far the most successful [9]. The reluctance of people to get vaccinated frequently results from worries about possible adverse consequences, which differ depending on the individual [10][11]. There are two main types of COVID-19 vaccinations, which are nucleic acid vaccine and viral vector vaccine [12]. Viral vaccines based on nucleic acids can be either deoxyribonucleic acid (DNA) or ribonucleic acid (RNA) vaccines, which work by using the host's cellular transcriptional and translational machinery to create viral proteins that the immune system may then detect [13]. Because they are relatively easy to produce, nucleic acid vaccines are better than other kinds of vaccines [14]. COVID-19 vaccines belong to RNA subcategory (however, a variant which is of DNA is being developed). Examples are Pfizer-BioNTech and Moderna. The viral vectors serve as a vehicle for delivering the target immunogen in viral vectored vaccinations [15]. The vector allows the body to build an immune response by delivering viral genes that produce antigens against an infectious pathogen. Examples include AstraZeneca, Janssen, and CanSino. The most frequent cardiac incident following COVID-19 vaccination was found to be myocarditis, with an overall prevalence of about 1.62 percent. mRNA vaccines caused myocarditis in almost 90% of post-COVID19 vaccination cases; in contrast, vector-based and/or inactivated vaccines caused fewer cases of this condition [16][17].

Because Artificial Intelligence (AI) is being utilized to help healthcare professionals forecast patient outcomes, it is continuously transforming biomedical research and healthcare management [18]. It is difficult for the general public and medical professionals to understand the outcomes of AI systems, which makes their acceptability problematic. Explainable AI (XAI) is the product of research attempts to make the decisions or outcomes of AI systems clearly comprehensible. XAI aids in comprehending the reasoning behind the results of machine learning algorithms [19][20]. Finding possible weaknesses and locating the sources of errors more quickly can be facilitated by explaining how AI systems operate internally [21][22].

On the other hand, ethics and transparency are necessary for people to have faith in medical devices that use AI [23]. Furthermore, by highlighting the combination of rigorous validation of AI decision-support systems in real-life clinical settings and valuable explanations, explainability promotes ethically sound medical decision-making [24]. It follows that the healthcare sector is one of the sectors where user acceptability of AI algorithms depends on both explainability and accuracy [25]. Explainability can also be used to evaluate whether the system makes fair decisions and allows users to confirm that it does not rely on noise or artefacts in the training data, especially in situations where the training data may present a partial or biased image of the population [26]. Moreover, explanations can help us better grasp what the algorithm is optimized for and the associated trade-offs, as well as provide fresh perspectives on what the AI system has learnt from the data [27]. This paper uses explainable AI to understand the post-vaccination mortality in persons with cardiovascular risk factors.

The rest of the paper is structured as follows: Section 2 discusses the related work. Section 3 presents the methodology. Section 4 presents the results of the research. The discussion was presented in Section 5. Section 6 discusses the limitations of the study while Section 7 presents the conclusion and further work.

2. Related work

Authors in [28] examined vaccination adverse events utilizing the Vaccine Adverse Event Reporting System (VAERS) database through the application of ontology and machine learning. In order to provide easy access to side effect information for patients, healthcare practitioners, and officials, a relational/graph database was established. Machine learning algorithms also predict important symptoms that lead to hospitalization and treatment. Using the VAERS dataset, [29] developed a

prognostic tool to determine the risks connected to COVID-19 vaccinations. Hospitalization, mortality, and COVID-19 outcomes were predicted using machine learning models such as Multilayer Feed-forward Perceptron, Random Forest, Naive Bayes, Light Gradient Boosting Algorithm, and Linear Regression. Males between the ages of 50 and 70 and those with significant diseases were found to be the most vulnerable. Authors in [30] used patient data to identify common factors in adverse reactions and develop strategies to reduce their incidence. They found that factors such as prior illnesses, hospital admission, and SARS-CoV-2 reinfection were significantly associated with poor patient reactions. Preexisting conditions like age, gender, and medication use were also significant predictors. Machine learning classifiers trained with medical history were successful in predicting complication-free vaccinations with an accuracy score above 90%.

Authors in [31] employed machine learning to classify COVID-19 vaccination adverse effects in people with sensitivities to foods, animals, and weather. Machine learning classifiers have been utilized to predict bad responses to COVID-19 vaccinations in patients with allergies. This has been helpful in identifying those who are more likely to have side effects. Authors in [32] studied COVID-19 vaccine adverse events by gender, age, manufacturer, and dose using the Vaccine Adverse Event Reporting System datasets. They found higher frequency in women, but different characteristics of vaccine adverse events, including gender, manufacturer, age, and underlying diseases, were associated with fatal cases.

Authors in [33] propose a multi-label classification method for Vaccine Adverse Event (VAE) detection, utilizing term- and topic-based label selection strategies. They use one-vs-rest, problem transformation, algorithm adaptation, and deep learning methods. Experimental results show that topic-based PT methods improve accuracy by up to 33.69%, OvsR methods achieve optimal accuracy of 98.88%, and AA methods increase accuracy by 87.36%. The method enhances model accuracy and VAE interpretability. Authors in [9] employed machine learning algorithms to predict and determine the severity of adverse responses to COVID-19 and influenza vaccinations. 2111 participants used wearables like the Garmin Vivosmart 4 and smartphone applications to participate in the study. The XGBoost model exhibited ROC values of 0.69 and 0.74 for detecting and predicting mild to severe side effects, respectively.

In order to identify ongoing research efforts using machine learning techniques to comprehend comorbidity processes and make clinical predictions taking these intricate patterns into account, [18] conducted a review. Four predictive analytics tasks were identified: risk prediction, network analysis, clustering, and illness comorbidity data extraction. The results show that certain machine learning-driven applications interpret the model and identify important risk factors for the development of comorbidity while also addressing intrinsic data inadequacies in healthcare datasets.

Authors in [34] investigated the impact of COVID-19 immunization in intubated patients with acute respiratory distress syndrome associated with COVID-19 in Taiwan. Data on patients who were intubated between May 1, 2022, and October 31, 2022, owing to COVID-19 pneumonia were retrospectively evaluated by the authors. A person was deemed completely immunized if they received two or more doses of the vaccine. There were 84 patients in all (40 completely immunized and 44 controls). The two groups had comparable baseline characteristics on the day of intubation, such as age, comorbidities, and Sequential Organ Failure Assessment (SOFA) score. There was no statistically significant difference in the Intensive Care Unit (ICU) death rate between the completely vaccinated and control groups. Body mass index (BMI) and SOFA score had a strong correlation with ICU mortality.

Numerous adverse occurrences following vaccination, including fatalities and serious illnesses, have been documented in the literature [35][36][37]. This paper examines a particular instance that has not been experimentally studied: post-vaccination adverse events involving individuals who had comorbidities and cardiovascular risk factors. The findings of this paper will foster an understanding of the underlying factors that could lead to mortality when adverse events due to post-vaccination occur for this category of patients.

3. Methodology

This section presents the details of methods/techniques used in predicting post-vaccination mortality in persons with cardiovascular risk factors using XAI. This was done in the following stages. Stage 1 contains data collection, preprocessing, and feature engineering. In stage 2, we have models' selection, hyperparameter tuning, model training and model evaluation. Architecture design, hyperparameter tuning, training, and evaluation of the Stacked ensemble were done in stage 3. In stage 4, we provided model interpretation and explainability. The process workflow is presented in Figure 1.

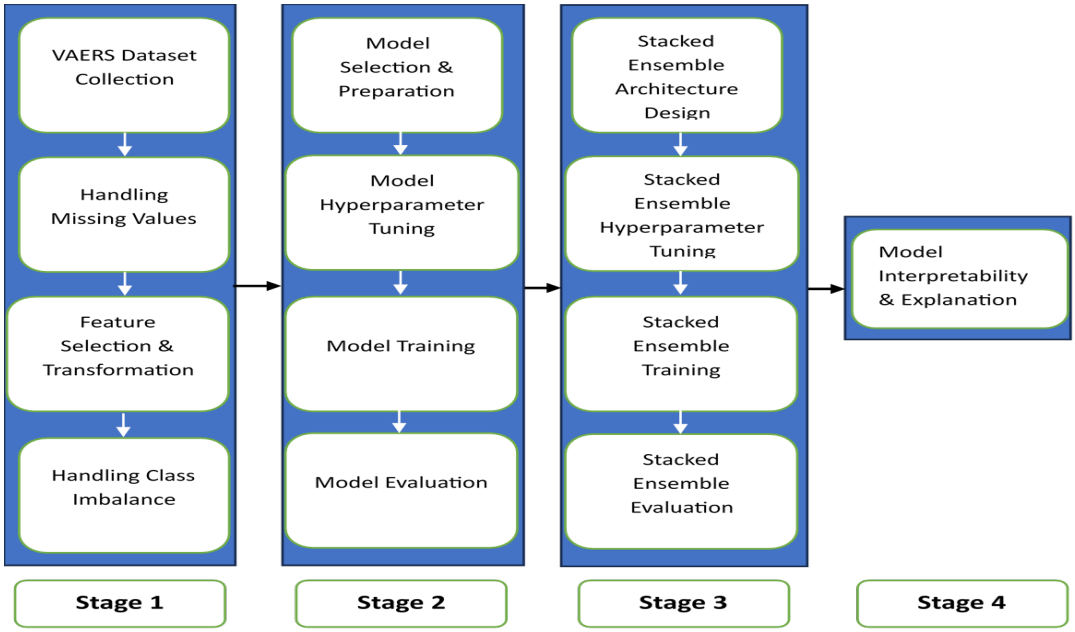

Figure 1: The post-vaccination mortality in persons with cardiovascular risk factors process workflow

3.1. Data collection

The reported incidences of COVID-19 post-vaccination adverse events were extracted from the Vaccine Adverse Event Reporting System (VAERS), which can be accessed through the following link: https://vaers.hhs.gov/data/datasets.html was used in this study. The dataset over five (5) years from 2020 to May 2024 was downloaded in CSV format. Each year dataset has three (3) CSV files tagged VAERSDATA (history/profile), VAERSYMPTOMS (symptoms), and VAERSVAX (vaccine). All the three datasets for each year were merged and named as 2020COVIDCOMBINED, 2021COVIDCOMBINED, 2022COVIDCOMBINED, 2023COVIDCOMBINED, and 2024COVIDCOMBINED.

Based on our interest in examining patients with cardiovascular disease risk factors, we extracted data on instances where any two of high blood pressure, diabetes, and high cholesterol had been acknowledged as a preexisting condition by a patient who reported post-vaccination adverse events. We then selected the records from each year and then merged them for all the five (5) years under consideration. To get relevant patient records, each of the symptoms was expanded in terms of their medical terminologies as follows: High blood pressure/Hypertension; Diabetes/Type 1 Diabetes/Type 2 (includes instances where similar/closely related words like diabetic, diabetes mellitus, diabetes/Type 1 diabetic/Type 2 were used); High cholesterol/Hyperlipidemia/Hypercholesterolemia. We had a total of 16657 records with 52 features. Out of the 52 features, there were 34 categorical features and 18 numerical features.

3.2. Data preprocessing

The records of patients less than 18 years of age (11 records) were removed because the focus was on adult patients. Patient records with unknown vaccine (50 records) and patients with unknown vaccine dosage (1552 records) were also removed. Redundant features irrelevant to our analysis or empty fields were also dropped. We introduced additional fields, as shown in Table 1.

Table 1
Additional variables inferred from the dataset

Additional Field	Value	Justification
DOSE_COMPLETE	Yes/No	For DOSE_COMPLETE, we input Y for a complete dose and N for an incomplete dose. For Moderna, Pfizer-Biontech, and Novavax vaccines, the patient must take 2 primary doses for a complete dosage. For Janssen, Pfizer-Biontech Bivalent, and Moderna Bivalent, only 1 primary dose is required. For patients who took more than the required dose, we removed their records (3,843 records) because they did not follow the prescription.
POST_VAX_SYMPTOM	Yes/ No	We input Y for at least one symptom shown by the patient. Otherwise, N was inputted. Our focus was to look at the three cardiovascular disease risk factors (high blood pressure, diabetes, and high cholesterol). We created four additional fields for high blood pressure (HBP), diabetes (DIABETES), high cholesterol (H_CHOL), and other illnesses (OTHER_ILL)
HBP	Yes/No	Since our focus was to look at the three cardiovascular disease risk factors (high blood pressure, diabetes, and high cholesterol, we extracted the value automatically from the patient's HISTORY field of the dataset
DIABETES	Yes/ No	
H_CHOL	Yes/ No	
OTHER_ILL	Yes/No	Due to the nature of the dataset, which is user-generated, there was no uniformity in the reporting by the users. Extracting whether a patient has other illnesses automatically becomes difficult since there are many other illnesses reported. We manually checked the entries one after the other to determine whether a patient had other illnesses apart from at least two combinations of the three diseases of interest (high blood pressure, diabetes, and high cholesterol.
VAX_NAME	A/B/C/ D/E/ F	We represented each type of vaccine with A, B, C, D, E, and F for easy representation and analysis. A: JANSSEN; B: MODERNA; C: MODERNA BIVALENT; D: NOVAVAX; E: PFIZER-BIONTECH; F: PFIZER-BIONTECH BIVALENT

3.2.1. Handling missing data

To handle the missing values in the dataset, we used univariate imputation or data removal techniques as follows:

1. AGE_YRS: There are 173 cases of missing age values. We calculated the mean (67.24) and median (68) of the patients' age from the existing observations. Since the mean and median are close, we picked the median value and inserted it in place of the missing observations.

2. SEX: We removed 33 records in which the sex was undefined (U) from the dataset.
3. SYMPTOM_TEXT: In the observed data, only five (5) missing values occurred in this field. We decided to remove the five patients' records.
4. L_THREAT, ER_VISIT, HOSPITAL, X_STAY, DISABLE, BIRTH_DEFECT, OFC_VISIT, ER_ED_VISIT: The response to these fields is either Y (Yes) or N (No). In the dataset, case Y was filled but case N was left blank. For our analysis, we auto-filled N (No) for all the empty cells in all the fields listed here.
5. HOSPDAYS: This field indicates the number of days a vaccine recipient was hospitalized; if no day was spent, it is left empty. We inputted 0 in the empty spaces.
6. OTHER_MEDS: This text field provides a narrative of any prescription or non-prescription drugs taken by the vaccine recipient at the time of vaccination, as reported in the form's stated field. The empty fields (1899) were filled with "None".
7. PRIOR_VAX: This field contains details about previous vaccine events registered in the form's listed field. The empty cells were replaced with "None."
8. CUR_ILL: Illnesses at time of vaccination—This text field includes a narrative of any diseases present at the time of vaccination, as stated in the form's designated field. For our analysis, we used 'Y' to represent the presence of disease and 'N' for none.
9. PRIOR_VAX: This field contains details about previous vaccine events that were registered in the form's listed field. We used "Y" for previous vaccine events and "N" for none.

3.3. Feature engineering

There are 22 attributes consisting of 19 categorical features and 3 numerical features (See Table 2).

Table 2
Features of the Dataset

#	Column	Data type
1	AGE_YRS	Numerical
2	SEX	Categorical
3	L_THREAT	Categorical
4	HOSPITAL	Categorical
5	HOSPDAYS	Numerical
6	X_STAY	Categorical
7	DISABLE	Categorical
8	RECOVD	Categorical
9	CUR_ILL	Categorical
10	HBP	Categorical
11	DIABETES	Categorical
12	H_CHOL	Categorical
13	OTHER_ILLN	Categorical
14	PRIOR_VAX	Categorical
15	BIRTH_DEFECT	Categorical
16	OFC_VISIT	Categorical
17	ER_ED_VISIT	Categorical
18	POST_VAX_SYMPTOM	Categorical
19	VAX_DOSE_SERIES	Numerical
20	DOSE_COMPLETE	Categorical
21	VAX_NAME	Categorical
22	DIED	Categorical

The categorical features were transformed using one-hot encoding, and Min-Max scaling was applied to the numerical features. LASSO Regression and Random Forests algorithms were used for feature selection. We split the data into train and test data in the ratio 70:30, respectively. The GridSearchCV

was used to get the best hyperparameter for alpha needed for LASSO Regression. The result shows 1e-05 for alpha as the best parameter for the GridSearchCV. The two feature selection algorithms used were compared using log-likelihood and AUC-ROC scores. Lasso Log-Likelihood (DIED_N): -0.17485207557765708, (DIED_Y): -0.1748520755782875; Random Forest Log-Likelihood (DIED_1): -0.0006035625338038024 (DIED_2): -0.0006035625338038024 as shown in Figure 2. The AUC-ROC scores for LASSO Regression and Random Forests are presented in Figure 3.

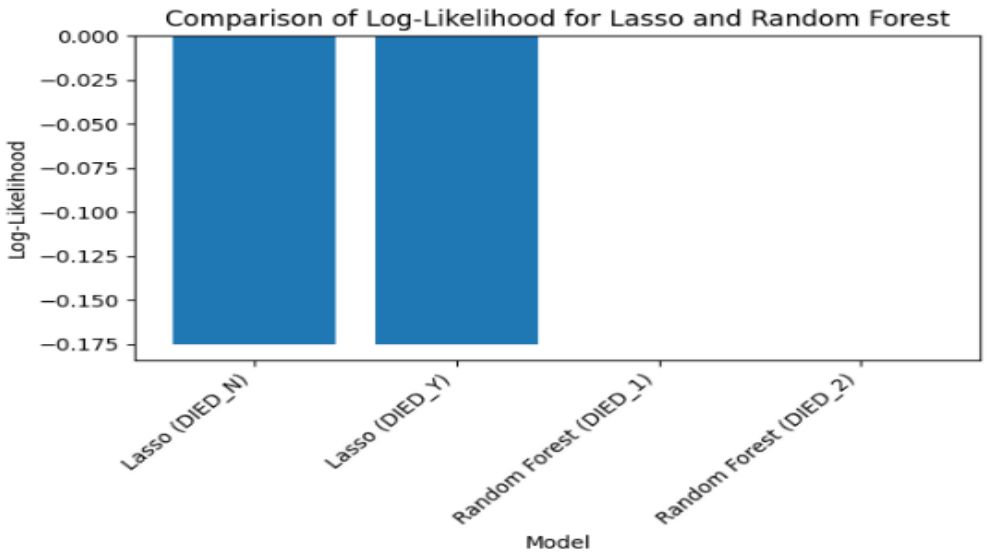

Figure 2: Comparison of log-likelihood for Lasso and random forests

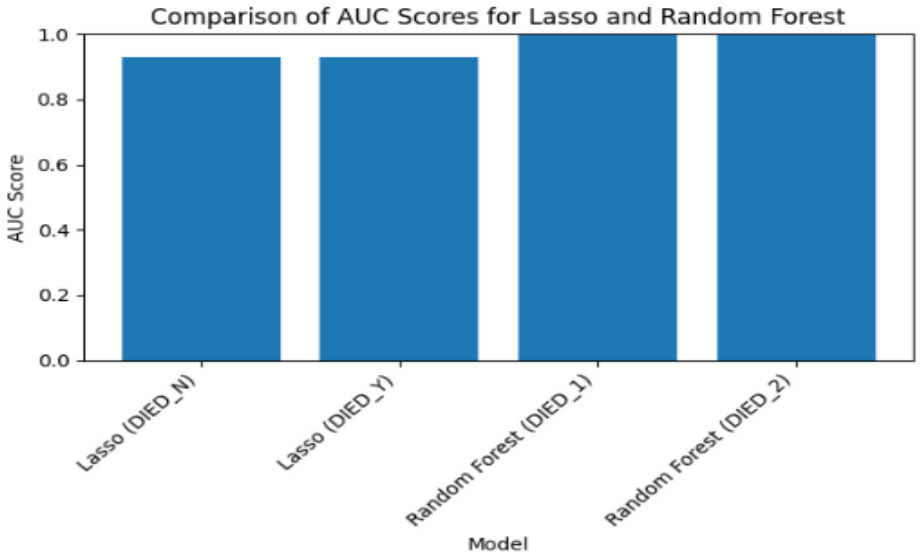

Figure 3: Comparison of AUC-ROC score for Lasso and random forests

The RF selected the top 19 features and LASSO also has 19 distinct selected features. The results of Log-likelihood (which shows the measure of goodness of fit), and AUC-ROC Scores showed that RF performed better than Lasso. Therefore, we chose features selected by RF for further analysis.

Based on the features selected, the original and distinct features from Excel in CSV format were then loaded. At the end of the features selection, we had a total of 20 distinct features: 18 categorical variables (including the target variable) and 2 numerical variables, as shown in Table 3. Table 3 provides the selected features and their descriptions.

Table 3
Selected Features

Categorical variable	Feature name/description	Data value	Frequencies
SEX	Gender type	Male/Female	4835, 6242
L_THREAT	Life-threatening illness experience	Yes/No	326, 10751
HOSPITAL	Hospitalized or not	Yes/No	4160, 6917
X_STAY	Prolongation of existing hospitalization or not	Yes/No	11, 11066
DISABLE	Disability or not	Yes/No	313, 10764
RECOVD	Was the vaccine recipient healed from the adverse incident or not	Yes/No/Unknown	3913, 4493, 2671
CUR_ILL	Illness at the time of vaccination or not	Yes/No	1752, 9325
HBP	Has high blood pressure or not	Yes/No	11040, 37
DIABETES	Has diabetes or not	Yes/No	6958, 4119
H_CHOL	Has high cholesterol or not	Yes/No	6947, 4130
OTHER_ILLN	Has other illness or not	Yes/No	9142, 1935
PRIOR_VAX	Had prior vaccination or not	Yes/No	516, 10561
BIRTH_DEFECT	Has congenital anomaly, birth defect or congenital disability or not	Yes/No	6, 11071
OFC_VISIT	Visited Doctor or other healthcare provider office or not	Yes/No	2870, 8207
ER_ED_VISIT	Visited emergency care or not	Yes/No	2492, 8585
DOSE_COMPLETE	Vaccine recipient completed the dose or not	Yes/No	5317, 5760
VAX_NAME	Name of vaccine received	A/B/C/D/E/F	768, 5089, 35, 2, 5141, 42
DIED	Vaccine recipient died or not	Yes/No	1035, 10042

Numerical variable	Feature name/description	Data range	Mean	Standard deviation
AGE_YRS	The age of the vaccine recipient in years	18-103	66.62	13.29
HOSPDAYS	Number of days Hospitalised	0-91	2.00	5.02

3.3.1. Handling class imbalance

Since our data is largely skewed based on our target variable, we have a class imbalance as presented in Figure 4 with Alive (N) = 10042 and Died (Y) = 1035 records.

We used Random Oversampling and SMOTE techniques to balance the datasets. We compared the performance of Random Oversampling with SMOTE using AUC_ROC and Average Precision metrics. SMOTE performed better in terms of AUC-ROC and Average Precision. Hence, SMOTE was selected as the resampling technique used to balance the dataset (see Table 4).

Table 4
Comparison between Random Over Sampling and SMOTE

Technique	Original dataset	Resampled dataset	AUC-ROC	Average precision
Random Over Sampling	N=10042, Y=1035	0: 10042, 1: 10042	0.93	0.90
SMOTE	N=10042, Y=1035	0: 10042, 1: 10042	0.94	0.91

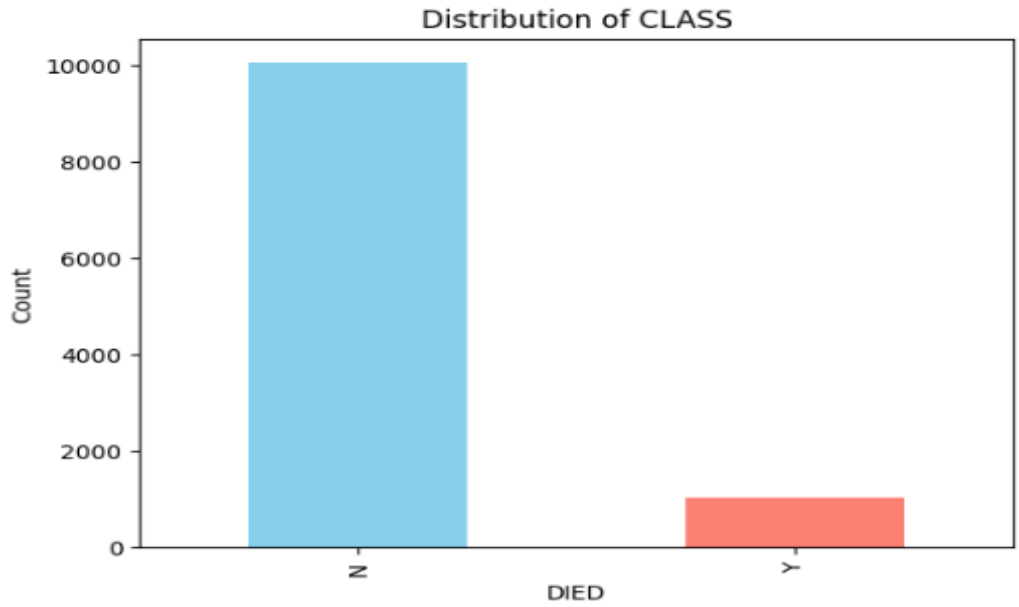

Figure 4: Class distribution of the dataset

3.4. Model training

We trained seven (7) ML algorithms on the dataset. The algorithms are Logistic Regression (LR), K-Nearest Neighbour (KNN), CatBoost, XGBoost, Support Vector Machine (SVM), Deep Multilayer Perception (DMLP), and Random Forests. We also trained two stacked ensembles, each of these had 6 base models and a meta-learner. The stacked ensembles are Stack Ensemble-CatBoost (CatBoost as the meta-learner) and Stack Ensemble-XGBoost (XGBoost as the meta-learner). We had a total of 20084 records after resampling with SMOTE. We split the data into training and testing sets in the ratio of 70:30, that is (14058: 6026), respectively.

3.4.1. Hyperparameter tuning

To select the best hyperparameters for the seven models and the stacked ensembles, we performed hyperparameter tuning for the seven models and the two stack ensembles using GridSearchCV and Bayesian Optimization, respectively. For the stack ensembles, the best hyperparameters were selected out of fifty (50) trials. The best hyperparameters for each model are presented in Table 5.

3.4.2. Addressing overfitting

In the hyperparameter optimization, we employed k-fold cross validation, which averages the scores across all iterations to get the final assessment of the models. One of the advantages of the ensemble techniques which was used in this study is to address the overfitting issue. In addition, the evaluation of the models was performed on the test set of the data used in this study.

4. Results

In this study, we have explored the post-vaccination adverse event for patients with cardiovascular comorbidity indicators by targeting those who are alive or dead. We used the following metrics to assess the performance of the seven algorithms and the two stacked ensembles: Accuracy, Precision, Recall, F1, AUC, and MCC. We also plot the confusion matrix and ROC curve for the algorithms. SHAP

Table 5
Best hyperparameters for each model

Model	Best hyperparameters
LR	'C': 100, 'penalty': 'l1', 'solver': 'liblinear'
K-NN	'metric': 'manhattan', 'n_neighbors': 11, 'weights': 'distance'
SVM	'C': 10, 'gamma': 'scale', 'kernel': 'rbf'
MLP	'activation': 'relu', 'alpha': 0.0001, 'hidden_layer_sizes': (100,), 'solver': 'adam'
XGBoost	'learning_rate': 0.2, 'max_depth': 7, 'n_estimators': 200'
RF	'max_depth': 20, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 200
CatBoost	'depth': 7, 'iterations': 200, 'learning_rate': 0.2
Stack Ensemble-CatBoost	'logreg_C': 0.0033904370788804157, 'knn_n_neighbors': 15, 'svm_C': 11.240009361665773, 'mlp_alpha': 0.011592470799986722, 'mlp_hidden_size': 143, 'rf_max_depth': 15, 'rf_min_samples_leaf': 5, 'rf_min_samples_split': 6, 'rf_n_estimators': 206, 'xgb_learning_rate': 0.23525278997943233, 'xgb_max_depth': 6, 'xgb_n_estimators': 189, 'catboost_iterations': 1164, 'catboost_learning_rate': 0.011319876453809892, 'catboost_depth': 5
Stack-Ensemble-XGBoost	'logreg_C': 0.03508229819901894, 'knn_n_neighbors': 5, 'svm_C': 469.50966160024706, 'mlp_alpha': 0.09685443295968259, 'mlp_hidden_size': 78, 'rf_max_depth': 14, 'rf_min_samples_leaf': 4, 'rf_min_samples_split': 8, 'rf_n_estimators': 109, 'catboost_iterations': 1755, 'catboost_learning_rate': 0.12118414866381266, 'catboost_depth': 3, 'xgb_learning_rate': 0.03690655905977055, 'xgb_max_depth': 5, 'xgb_n_estimators': 129

was also used to determine the importance of the features and their impact on the best-performing model. The results of the seven models and two stack ensemble models are presented in Table 6.

From Table 6, CatBoost had the best performance among all the individual models that were trained and tested. As a result, CatBoost was used as a meta-model for the stacked ensemble, while the remaining six models were used as base models. However, we noticed a slight drop in the performance of Stack ensemble-CatBoost compared to the performance of CatBoost as an individual model. We went on to pick the individual model that performed next to CatBoost, which was XGBoost, as the meta-learner of the stacked ensemble, and the result showed a superior performance.

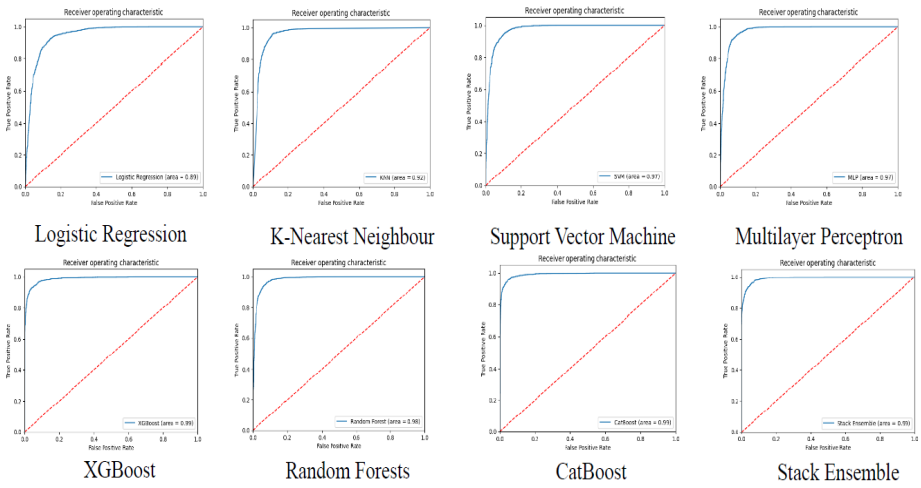

Figure 5: ROC Curve on test data

Table 6
Performance of the Models on Test Data

Model		Precision	Recall	F1-Score	Accuracy	AUC	MCC
Logistic Regression	Alive (0)	0.92	0.86	0.89	0.89	0.89	0.79
	Death (1)	0.87	0.92	0.89			
	Macro Average	0.89	0.89	0.89			
	Weighted Average	0.89	0.89	0.89			
K-Nearest Neighbour	Alive (0)	0.94	0.90	0.92	0.92	0.92	0.84
	Death (1)	0.90	0.95	0.92			
	Macro Average	0.92	0.92	0.92			
	Weighted Average	0.92	0.92	0.92			
Support Vector Machine	Alive (0)	0.95	0.89	0.92	0.92	0.92	0.85
	Death (1)	0.89	0.96	0.92			
	Macro Average	0.92	0.92	0.92			
	Weighted Average	0.92	0.92	0.92			
Multilayer Perceptron	Alive (0)	0.94	0.90	0.92	0.93	0.93	0.85
	Death (1)	0.91	0.95	0.93			
	Macro Average	0.93	0.93	0.93			
	Weighted Average	0.93	0.93	0.93			
XGBoost	Alive (0)	0.95	0.94	0.94	0.94	0.94	0.89
	Death (1)	0.94	0.95	0.94			
	Macro Average	0.94	0.94	0.94			
	Weighted Average	0.94	0.94	0.94			
Random Forest	Alive (0)	0.96	0.92	0.94	0.94	0.94	0.88
	Death (1)	0.92	0.96	0.94			
	Macro Average	0.94	0.94	0.94			
	Weighted Average	0.94	0.94	0.94			
CatBoost	Alive (0)	0.96	0.95	0.95	0.96	0.96	0.91
	Death (1)	0.95	0.96	0.96			
	Macro Average	0.96	0.96	0.96			
	Weighted Average	0.96	0.96	0.96			
Stack Ensemble-CatBoost	Alive (0)	0.95	0.95	0.94	0.95	0.95	0.90
	Death (1)	0.95	0.95	0.95			
	Macro Average	0.95	0.95	0.95			
	Weighted Average	0.95	0.95	0.95			
Stack Ensemble-XGBoost	Alive (0)	0.95	0.97	0.96	0.96	0.99	0.92
	Death (1)	0.97	0.95	0.96			
	Macro Average	0.96	0.96	0.96			
	Weighted Average	0.96	0.96	0.96			

From Figure 5, the receiver operating characteristic (ROC) curves for all the models are shown. The ROC is used to evaluate the performance of a binary diagnostic classification method. The ROC tells how much the model is capable of distinguishing between classes. All the models show outstanding performance of more than 0.9, except in the logistic regression model, which has 0.89. The ROC results show that the models can clearly distinguish between the two classes, Alive or Died, in the case of this study.

4.1. Model prediction explanation

SHAPley Additive ExPlanations (SHAP) is used to explain the output of a machine learning model. It uses game theory to assign credit for a model's prediction to each feature or feature value [38][39]. Out of all the seven (7) base models, the performance of the CatBoost model was the best. The CatBoost model's predictions on the test data were used as input to the SHAP for the model's predictions' explanations. The SHAP summary plot is presented in Figure 6.

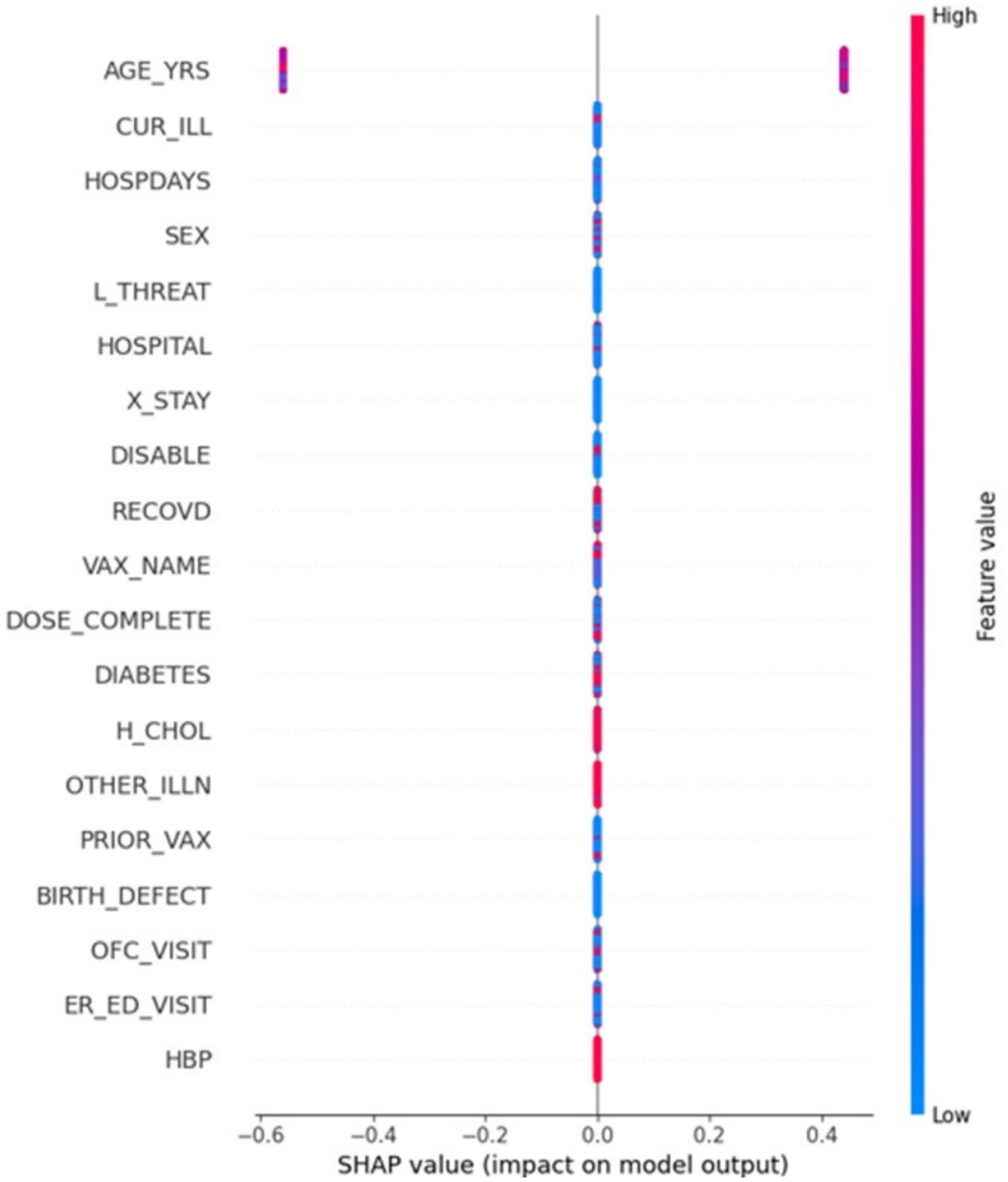

Figure 6: SHAP Summary plot for the CatBoost model's predictions

From Figure 6, the Y-axis indicates the feature names in order of importance from top to bottom. The X-axis represents the SHAP value, which indicates the degree of change in log odds. The colour of each point on the graph represents the value of the corresponding feature, with red indicating high values and blue indicating low values. Each point represents a row of data from the original dataset. From Figure 6, it is clear that the feature 'AGE_YRS' has the highest impact on the model while the feature 'HBP' has the least impact.

We identified the significant determinants of post-vaccination mortality in individuals with cardiovascular comorbidities from the SHAP summary plot. The risk of death generally rises with age and the number of days spent in the hospital; on the other hand, the likelihood of survival increases with the absence of current illness, life-threatening conditions, hospitalization, prolonged hospitalization, disability, birth defects, doctor visits, and emergency care. On the other hand, the risk of death is raised by diabetes, high blood pressure, high cholesterol, and other conditions. The results of this study add to

our knowledge of important variables to consider when managing post-vaccination adverse events for patients who also have cardiovascular disease comorbidity indicators.

5. Discussion

Authors in [40] identified the top ten factors that increase the risk of cardiovascular disease (CVD) as unhealthful nutrition, physical inactivity, dyslipidemia (high cholesterol), hyperglycemia (high blood sugar), high blood pressure (hypertension), obesity, considerations of select populations (older age, race/ethnicity, and sex differences), thrombosis/smoking, kidney dysfunction and genetics/familial hypercholesterolemia. This study investigates how COVID-19 post-vaccination adverse events affect persons with CVD risk factors, particularly focusing on the co-occurrences of dyslipidemia (high cholesterol), hyperglycemia (high blood sugar), diabetes, and high blood pressure (hypertension). Our results show that the risk of death from COVID-19 post-vaccination adverse events increases for persons with diabetes, high blood pressure, high cholesterol, and other illness conditions. This observation agrees with the results of previously reported clinical studies.

According to [41], based on a study on 187 COVID-19 patients in which 144 patients survived and 43 patients died, found that the probability of death was higher (69.44%) for those with underlying CVD and elevated TnTs (troponin T) levels. Also, patients with underlying CVD were more likely to exhibit elevation of TnT levels (54.5%) compared with the patients without CVD (13.2%). In addition, patients with elevated TnT levels had evidence of more severe respiratory dysfunction and developed more frequent complications. Also, [32], a systematic review study that captured findings from 3912 participants, found that patients with preexisting CVD had worse outcomes and increased risk of death from COVID-19. Also, CVD risk factors such as hypertension, diabetes mellitus, and obesity were associated with the severity of COVID-19 infection, intensive care unit admission and poor prognosis. Thus, the results of this study add to our knowledge of important variables to consider when managing post-vaccination adverse events for patients who also have cardiovascular disease comorbidity indicators.

6. Limitations of the study

The VAERS dataset used in this research work is user generated information and there was no validation done on the reports. The type of features in the dataset also constitutes a limitation as most of the features extracted from the dataset are categorical variables; the dataset did not have the actual values of cholesterol, blood pressure, glycemia values as numerical variables. The absence of these numerical variables limit the depth of computational analysis that is possible.

7. Conclusion and further work

This paper explores the critical determinants of mortality in post-vaccination adverse events for persons with cardiovascular disease risk factors using XAI. We did this by extracting 16657 records (from 2020 to May 2024) of COVID-19 vaccinated persons with preexisting cardiovascular disease risk factors (any two of high blood pressure, diabetes, and high cholesterol) from the VAERS public dataset. Next, we applied a predictive modelling process that has four stages, which are 1) data preprocessing; 2) model training and performance evaluation of seven ML algorithms: Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Categorical Boosting (CatBoost), Random Forest (LR), and Extreme Gradient Boosting (XGBoost); 3) modelling using two stacked ensembles consisting of six base models — Using Catboost and XGBoost as the meta-learners respectively; 4) model explainability using SHAP.

The results show that Stacked ensemble - XGBoost had the best overall performance (Acc = 0.96, F1 = 0.96, AUC = 0.99), whereas CatBoost has the best performance among the individual machine learning models (Acc = 0.96, F1 = 0.96, AUC = 0.96). We also identified the critical determinants of mortality in

persons with cardiovascular disease risk comorbidity when a post-vaccination adverse event occurs. The risk of death generally rises with age and the number of days spent in the hospital; on the other hand, the likelihood of survival increases with the absence of current illness, life-threatening conditions, hospitalization, prolonged hospitalization, disability, birth defects, doctor visits, and emergency care. Also, the risk of death is raised by diabetes, high blood pressure, high cholesterol, and other conditions. The results of this study foster an understanding of critical factors that could enable better handling of adverse events related to post-vaccination in patients with cardiovascular disease comorbidities.

In further research, we will explore the severity of COVID-19 post-vaccination adverse events in patients with cardiovascular disease risk factors. We will also try to explore real-life datasets on the same scenario proposed in this study.

Acknowledgments

The research was supported by the National Research Foundation (NRF) of South Africa, and the University of Pretoria, South Africa.

References

[1] K. Kim, Risk stratification of cardiovascular disease according to age groups in new prevention guidelines: a review, Journal of Lipid and Atherosclerosis 12(2) (2023) 96-105. https://doi.org/10.12997/jla.2023.12.2.96.

[2] F. Tian, L. Chen, Z. Qian, H. Xia, Z. Zhang, J. Zhang, C. Wang, M. G. Vaughn, M. Tabet, H. Lin, Ranking age-specific modifiable risk factors for cardiovascular disease and mortality: evidence from a population-based longitudinal study, eClinicalMedicine 64 (2023) 102230. doi: 10.1016/j.eclinm.2023.102230.

[3] WHO, Number of COVID-19 deaths reported to WHO (cumulative total) (2024). https://data.who.int/dashboards/covid19/deaths?n=c

[4] P. Hosseinzadeh, M. Zareipour, E. Baljani, M.R. Moradali, Social consequences of the COVID-19 pandemic, A systematic review. Investig. Educ. Enferm. 40 (2022).

[5] J. A. Elharake, F. Akbar, A. A. Malik, W. Gilliam, S. B. Omer, Mental health impact of COVID-19 among children and college students: a systematic review, Child Psychiatry Hum. Dev. 54(3) (2022) 913-925. https://doi.org/10.1007/s10578-021-01297-1.

[6] B. Hu, H. Guo, P. Zhou, Z. L. Shi, Characteristics of SARS-CoV-2 and COVID-19. Nat. Rev. Microbiol. 19 (2021)141.

[7] X. Yang, Y. Yu, J. Xu, H. Shu, J. Xia, H. Liu, Y. Wu, L. Zhang, Z. Yu, M. Fang, T. Yu, Y. Wang, S. Pan, X. Zou, S. Yuan, Y. Shang, Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study, Lancet Respir Med 8 (2020) 475–81. https://doi.org/10.1016/S2213-2600(20)30079-5.

[8] C. Dayaramani, J. D. Leon, A. B. Reiss, Cardiovascular Disease Complicating COVID-19 in the elderly. Medicina, 57(8) (2021) 833. https://doi.org/10.3390/medicina57080833

[9] Y. Levi, M. L. Brandeau, E. Shmueli, D. Yamin, Prediction and detection of side effects severity following COVID-19 and influenza vaccinations: utilizing smartwatches and smartphones, Scientific Reports 14(1) (2024) 6012. doi: 10.1145/1188913.1188915.

[10] N. Biswas, J. K. Mustapha, J. H. Price, The nature and extent of COVID-19 vaccination hesitancy in healthcare workers. J. Commun. Health 46 (2021) 1244.

[11] H. Azarpanah, M. Farhadloo, R. Vahidov, L. Pilote, Vaccine hesitancy: evidence from an adverse events following immunization database, and the role of cognitive biases. BMC Public Health 21 (2021) 1-13.

[12] B. K. Muhar, J. Nehira A. Malhotra S. O. Kotchoni, The race for COVID-19 vaccines: the various types and their strengths and weaknesses, J Pharm Pract. 36(4) (2023) 953-966. doi: 10.1177/089719010.1177/089719002210972480221097248.

[13] W. W. Leitner, H. Ying, N. P. Restifo, DNA and RNA-based vaccines: principles, progress and prospects, Vaccine 18(9-10) (1999) 765-777. doi: 10.1016/s0264-410x(99)00271-6.

[14] N. P. Restifo, H. Ying, L. Hwang, W. W. Leitner, The promise of nucleic acid vaccines, Gene Ther. 7(2) (2000) 89-92. doi: 10.1038/sj.gt.3301117.

[15] Centers for Disease Control and Prevention, Vaccines and immunizations: viral vector COVID-19 vaccines. Atlanta, GA: Centers for Disease Control and Prevention (2021). www.cdc.gov/vaccines/covid-19/hcp/viral-vector-vaccine-basics.html.

[16] M. H. Paknahad, F. B. Yancheshmeh, A. Soleimani, Cardiovascular complications of COVID-19 vaccines: a review of case-report and case-series studies, Heart Lung 59 (2023) 173-180. doi: 10.1016/j.hrtlng.2023.02.003.

[17] Z. Akhtar, M. Trent, A. Moa, T. C. Tan, O. Fröbert, C. R. MacIntyre, The impact of COVID-19 and COVID vaccination on cardiovascular outcomes, European Heart Journal Supplements 25(Supplement_A) (2023) A42–A49. https://doi.org/10.1093/eurheartjsupp/suac123.

[18] D. Xu, Z. Xu, Machine learning applications in preventive healthcare: A systematic literature review on predictive analytics of disease comorbidity from multiple perspectives, Artificial Intelligence in Medicine 156 (2024) 102950. https://doi.org/10.1016/j.artmed.2024.102950.

[19] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N Díaz-Rodríguez, F. Herrera, Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence. Information Fusion 99 (2023) 101805. https://doi.org/10.1016/j.inffus.2023.101805

[20] T. Kolajo, O. Daramola, Human-centric and semantics-based explainable event detection: a survey, Artificial Intelligence Review 56 (2023) 119-158. https://doi.org/10.1007/s10462-023-10525-0

[21] J. Amann, D. Vetter, S. N. Blomberg, H. C. Christensen, M. Coffee, S. Gerke S, To explain or not to explain?—Artificial intelligence explainability in clinical decision support systems. PLOS Digit Health 1(2) (2022) e0000016. https://doi.org/10.1371/journal.pdig.0000016

[22] U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J. M. F. Moura, P. Eckersley, Explainable machine learning in deployment. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency pp.648 - 657

[23] L. Farah, J. M. Murris, I. Borget, A. Guilloux, N. M. Martelli, S. I. Katsahian, Assessment of performance, interpretability, and explainability in artificial intelligence–based health technologies: what healthcare stakeholders need to know. Mayo Clinic Proceedings: Digital Health 1(2) (2023) 120-138. https://doi.org/10.1016/j.mcpdig.2023.02.004

[24] Gerdes, A. The role of explainability in AI-supported medical decision-making. Discov Artif Intell 4 (2024) 29. https://doi.org/10.1007/s44163-024-00119-2.

[25] V. Hassija, V. Chamola, A. Mahapatra, A. Singal, D. Goel, K. Huang, S. Scardapane, I. Spinelli, M. Mahmud, A. Hussain, Interpreting black-box models: a review on explainable artificial intelligence, Cogn Comput 16 (2024) 45-74. https://doi.org/10.1007/s12559-023-10179-8.

[26] T. Kirat, O. Tambou, V. Do, A. Tsoukiàs, Fairness and explainability in automatic decision-making systems. a challenge for computer science and law, EURO Journal on Decision Processes 11 (2022) 100036. https://doi.org/10.1016/j.ejdp.2023.100036.

[27] M. M. Soliman, E. Ahmed, A. Darwish, A. E. Hassanien, Artificial intelligence powered metaverse: analysis, challenges and future perspectives. Artif Intell Rev 57 (2024) 36. https://doi.org/10.1007/s10462-023-10641-x.

[28] Z. Liu, X. Gao, C. Li, Modeling COVID-19 vaccine adverse effects with a visualized knowledge graph database, Healthcare 10 (2022) 1419. https://doi.org/10.3390/healthcare10081419.

[29] H. Gupta, O. M. Verma, Vaccine hesitancy in the post-vaccination COVID-19 era: a machine learning and statistical analysis driven study, Evolutionary Intelligence 16 (2023) 739–757 https://doi.org/10.1007/s12065-022-00704-3.

[30] M. Ahamad, S. Aktar, J. Uddin, R. Al-Mahfuz, A. K. M. Azad, S. Uddin, S. A. Alyami, I. H. Sarker, A. Khan, P. Liò, J. M. W. Quinn, M. A. Moni, Adverse effects of COVID-19 vaccination: machine learning and statistical approach to identify and classify incidences of morbidity and postvaccination reactogenicity, Healthcare 11 (2023) 31. https://doi.org/10.3390/healthcare11010031.

[31] R. O. Irsheidat, S. Nakhleh, R. M. Alruosan, H. Najadat, Utilizing machine learning techniques to predict adverse reactions to COVID-19 vaccines in individuals with allergies, 2023 14th International Conference on Information and Communication Systems (ICICS). doi: 10.1109/ICICS60529.2023.10330537.

[32] S. Cheon, T. Methiyothin, I. Ahn, Analysis of COVID-19 vaccine adverse event using language model and unsupervised machine learning. PLoS ONE 18(2) (2023) e0282119. https://doi.org/10.1371/journal.pone.0282119.

[33] D. Chen, R. Zhang, COVID-19 vaccine adverse event detection based on multi-label classification with various label selection strategies, IEEE Journal of Biomedical and Health Informatics 27(9) (2023) 4192-4203.

[34] K. Wong, C. Kuo, I. T-zeng, C. Hsu, C. Wu, The COVIDTW2 study: role of COVID-19 vaccination in intubated patients with COVID-19-related acute respiratory distress syndrome in Taiwan, Journal of Infection and Chemotherapy 30(5) (2024) 393-399. https://doi.org/10.1016/j.jiac.2023.11.010.

[35] K. Faksova, D. Walsh, Y. Jiang, J. Griffin, A. Phillips, A. Gentile, J. Kwong, K. Macartney, M. Naus, Z. Grange, S. Escolano, G. Sepulveda, A. Shetty, A. Pillsbury, C. Sullivan, Z. Naveed, N. Janjua, N. Giglio, J. Perälä, . . . A. Hviid, COVID-19 vaccines and adverse events of special interest: a multinational Global Vaccine Data Network (GVDN) cohort study of 99 million vaccinated individuals, Vaccine 42(9) (2024) 2200-2211. https://doi.org/10.1016/j.vaccine.2024.01.100.

[36] H. S. Mangat, B. Rippon, N. T. Reddy, A. A. Syed, J. M. Maruthanal, S. Luedtke, J. J. Puthumana, A. Srivatsa, A. Bosman, P. Kostkova, Reported rates of all-cause serious adverse events following immunization with BNT-162b in 5-17-year-old children in the United States, PLoS One 18(2) (2023) e0281993. doi: 10.1371/journal.pone.0281993.

[37] S A. Amer, E. A. Imam, E. M. Ishteiwy, I. F. Djelleb, L. R. Abdullh, D. Ballaj, Y. A. H. R. Amer, A. M. Elshabrawy, G. Eskander, J. Shah, M. L. Raza, A. M. ALsafa, H. T. Ali, H. M. Fawzy, Exploring the reported adverse effects of COVID-19 vaccines among vaccinated Arab populations: a multi-national survey study, Scientific Reports 14(1) (2024) 1-15. https://doi.org/10.1038/s41598-024-54886-0

[38] L. S. Shapley, A value for N-person games, Contributions to the Theory of Games 2 (1953) 307-317.

[39] E. Štrumbel, I. Kononenko, An efficient explanation of individual classifications using game theory, J. Mach. Learn. Res. 11 (2010) 1-18.

[40] H. E. Bays, P. R. Taub, E. Epstein, E. D. Michos, R. A. Ferraro, A. L. Bailey, H. M. Kelli, K. C. Ferdinand, M. R. Echols, H. Weintraub, J. Bostrom, Ten things to know about ten cardiovascular disease risk factors, American Journal of Preventive Cardiology 5 (2021) 100149.

[41] T. Guo, Y. Fan, M. Chen, X. Wu, L. Zhang, T. He, H. Wang, J. Wan, X. Wang, Z. Lu, Cardiovascular implications of fatal outcomes of patients with coronavirus disease 2019 (COVID-19), JAMA cardiology 5(7) (2020) 811-818.