Analysis and Prediction of Adverse Events Following Immunization (AEFIs) of COVID-19 Vaccines

ABSTRACT

There were lot of challenges that the global population have been facing since the pandemic arose. Fortunately, everyone was able to witness the advancement of the vaccine technology when mRNA-based vaccines were ready for their administration within a year of the pandemic. But majority of the population had their concerns about the vaccine safety and reliability. It is important to constantly monitor the status of the patients who are administered with the immunization and perform mass-study based on the data generated through these studies. We through this project wanted to contribute and learn more about AEFIs reported post SARS-CoV-2 related vaccines. We performed exploratory data analysis to understand the nuances of the data patterns. Then we defined two objectives for our study. We wanted to predict the onset time of the adverse effect using regression based, tree based, neural network models. Also, we wanted to study the possibility of hospitalization based on the symptoms that are observed. But we have major challenges of imbalance data and sparsity. We explored random under sampling and SMOTE technique to handle the imbalance data. We tried to handle sparsity of the data using dimensionality reduction techniques like SparsePCA, Sparse Naïve Base Feature Selection, SVD and used classification models like logistic regression, linear SVM, XGBoost, Random Forests to predict the need for hospitalization. We used RMSE as a model metric for the first objective and ROC-AUC, Specificity, Sensitivity scores for the second objective.

I. INTRODUCTION

An adverse event following immunization (AEFI) is any health-related issue that occurs during the intake of vaccine or post the intake. The adverse event is a temporal association without any causal relation to the vaccine intake. The Council for International Organizations of Medical Sciences (CIOMS) defines an AEFI as "*any untoward medical occurrence which follows immunization, and which does not necessarily have a causal relationship with the usage of the vaccine. The adverse event may be any unfavorable or unintended sign, abnormal laboratory finding, symptom or disease*". There is a difference between adverse event and adverse reaction/adverse effect where latter suggests that some scientific evidence supports the causation with such a reaction.

Vaccine Adverse Event Reporting System (VAERS)[11] is a national early warning system established to identify these unprecedented patterns of health events post the vaccine intake. We have sourced the data for this project from their portal for multiple vaccines.

II. RELATED WORK

Since the pandemic started and when emergency approval of the vaccines is provided, there has been lot of work related to the vaccine outcome and its adverse effects post the intake. For example, Md. Martuza Ahamad et al [1] have built models based on categorizing the data into 3 sets with status as "death status", the second one is "SARS-CoV-2 test status" and the third one is patient "hospital admission status". They have extracted 56 most reactions and used for training. However, the sparsity of the data is not handled here which we are very keen to study more through this project.

Also, a similar study was done by Ma'mon M Hatmal et al [2] based on the data from Jordan vaccination drive. They ran different Random Forest, MLP, XGBoost, K-Star algorithm to predict if there are any side effects seen on the individuals post the vaccine intake.

We haven't come across much research or study about how quick or late the reaction can be seen post the vaccine intake. So, we chose one of the objectives to predict the onset time of the adverse effects post the vaccine intake. Also, we are trying to handle the sparse data using the techniques like SparsePCA, Sparse Naïve Bayes Feature Selection method[3], SVD for dimensionality reduction methodologies.

III. DATASET AND FEATURES

The VAERS data is verified and checked for reliability by Center of Disease Control and Prevention, the national public health agency of the U.S. The dataset was extracted from VAERS website for a period from December 2020 up to October 2021. Currently, the data was filtered appropriately to extract data associated to COVID-19 since first we want to focus on studying the potential adverse effects of the U.S. licensed COVID-19 vaccines. VAERS data is a set of 3 open-source raw data files.

Each patient's reported details are assigned with a unique identification number 'VAERS ID' with actual identity of the patient hidden. The VAERS ID serves as a link to join all the 3 datasets. Some of the features mentioned below are of major interest for this project and will be encoded (standardized) and converted to categorical variables accordingly. Particularly text fields like History i.e., pre-existing ailments and symptoms will be utilized extensively for the study.

A. VOLUMES

Number of records - 774,192
Number of raw features - 52

B. FEATURES

The downloadable VAERS public data set consists of three separate data files. The public data set is updated periodically, and the date of the update is referenced on the website [r]. Below we highlighted some of the key features from each of the datasets which we used for our analysis.

VAERSDATA.csv – The dataset consists of 35 columns with few useful features as listed below related to the patient basic details, medical history, dates related to the timeline of events, any ongoing health conditions.

TABLE I
VAERSDATA.csv

Header	Type	Description
VAERS_ID	Int	VAERS identification number
AGE_YRS	Float	Age in Years
Sex	String	Sex
SYMPTOM_TEXT	String	Symptoms observed during reporting
DIED	String	Is the patient dead/alive?
HOSPDAYS	Float	Number of days hospitalized
DISABLE	String	Is the patient suffering from disability?
VAX_DATE	Date	Date of vaccine administration
ONSET_DATE	Date	Adverse event onset date
HISTORY	String	Chronic or long-standing health conditions or Pre-existing ailments
ALLERGIES	String	Allergies to medications, food, or other products
CUR_ILL	String	Illnesses at time of vaccination

VAERSVAX.csv – The dataset specifies which vaccine has been administered to a particular patient.

TABLE II
VAERSVAX.csv

Header	Type	Description
VAERS_ID	Int	VAERS identification number
VAX_NAME	String	Vaccination name

VAERSSYMPTOMS.csv – The dataset provides information average 5 symptoms reported by a particular patient. The type of symptoms is very high and hence this feature is likely to be of high dimension and very sparse. The symptoms are aligned with the MedDRA dictionary, and their version numbers are available in the dataset.

TABLE III
VAERSSYMPTOMS.csv

Header	Type	Description
VAERS_ID	Int	VAERS identification number
SYMPTOM1	String	Adverse event term 1
SYMPTOM2	String	Adverse event term 2
SYMPTOM3	String	Adverse event term 3
SYMPTOM4	String	Adverse event term 4
SYMPTOM5	String	Adverse event term 5

IV. EXPLORATORY DATA ANALYSIS

Figure 1 shows that most of the adverse events are reported from the following States - CA, FL, TX, NY.

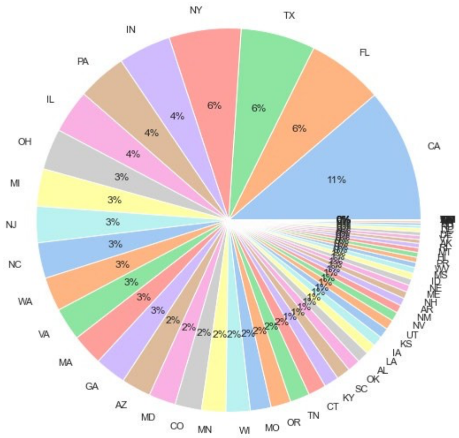

FIGURE 1. Pie Chart of number of AEFI events State wise

As per Figure 2 Headache, Pyrexia, Fatigue, Chills and Pain are the top 5 most seen symptoms in patients after undertaking vaccine.

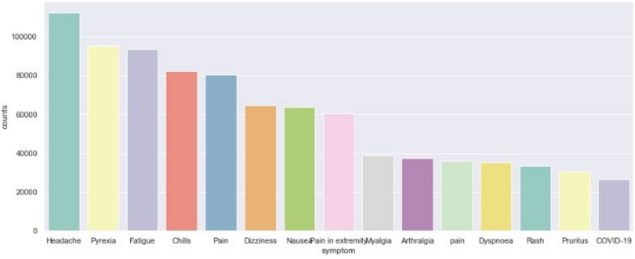

FIGURE 2. Top reported symptoms post intake of Vaccines

Figure 3 depicts the rate of the adverse event reported cases over time for all the vaccine types. Patients who had undertaken Pfizer and Janssen manufactured vaccines seem to have reported most and least adverse cases in the recent past respectively.

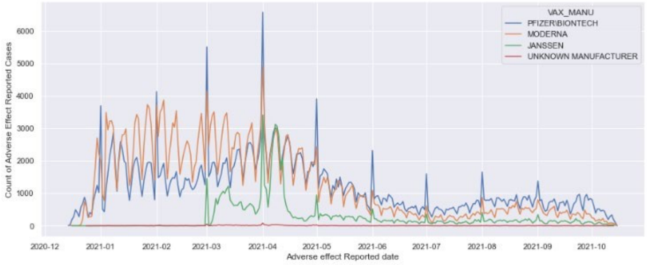

FIGURE 3. Events reported split by Vaccine Manufacturer from December 20 to October 21.

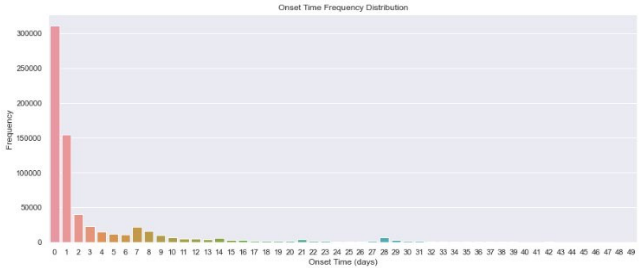

FIGURE 4. Onset Time Frequency Distribution for the events reported

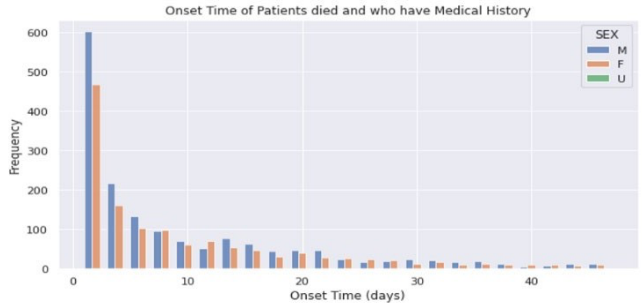

FIGURE 5. Onset Time Frequency Distribution of patients who died and had medical history

Figures 4 and 5 show that the most frequent onset times are either 0 or 1 and on average more men who have had medical history died.

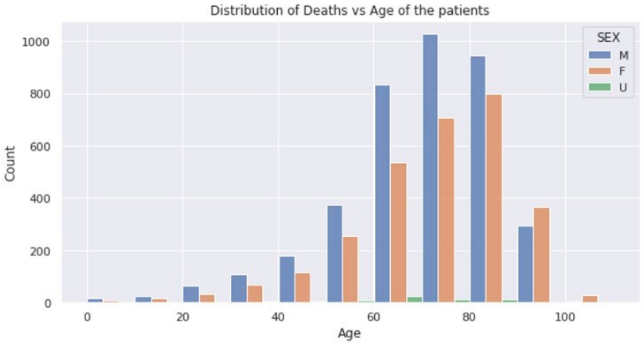

FIGURE 6. Distribution of Number of Deaths w.r.t Age of the patients

As expected, it seen that Number of deaths is positively correlated with the Age of the patients (Figure 6).

V. METHODOLOGY

A. DATA PREPARATION

Some key steps of data preparation are highlighted below which acted as a base for the model building.

- Join the 3 datasets based on the *VAERS_ID*. The volume might increase from the base data as in the *VAERSSYMPTOMS* file can have multiple records for a single *VAERS_ID*. This is because, a single patient can have more than 5 symptoms which are inserted as multiple entries.
- Categorical encoding is done on some of the variables like other meds, current illness, disabilities, sex, Vaccine manufacturer.
- Date *imputing* is done using forward imputing and backward imputing rules as part of the data cleansing step. Records with irrelevant date are dropped. This step is done for *Vaccine Intake* date and *Event Reported* date.
- 17 Baseline features have been identified based on the medical history column and they are encoded using dummies.
- Records which are outliers for *NUMDAYS* are removed.

B. PREDICTION OF ONSET TIME

The purpose is to predict the first adverse event shown by a patient after vaccine intake. The onset time in days is calculated based on the date of vaccine intake and the date on which the first adverse event was reported. 7 predictors related to AGE, SEX, Vaccine Type + 17 baseline medical history checks features have been picked to predict the onset time. The records in which *Age* is less than 11, *Sex* is Unknown, *Vaccine Manufacturer* is Unknown Manufacturer and having NAN values have been dropped. Finally, the dataset is split in the ratio of 8:2 to obtain the training and testing data samples. Multiple regression algorithms have been trained and tested using the pre-processed data. The root mean squared error has been used to evaluate all the models. The features that most affected the onset time prediction have been found for Random Forest Regression and XG Boost Regression models.

Models Used:

- **Ordinary Least Square:** The OLS has been used as a baseline model to predict onset time.
- **Regularized Regression (Lasso & Ridge):** By hyper parameter tuning using 10-fold cross-validation, we found an optimal value for alpha.
- **Random Forest Regressor:** The number of estimators used in the model was set to 500 with maximum depth of the tree equal to 100. The number of predictors used to sample at each node was set to (Number of features / 3) 8. The feature importance has been evaluated using the Mean decrease in impurity metric.
- **XG Boost Regressor:** 300 hidden layers have been used with maximum number of iterations allowed

set to 500. The feature importance has been evaluated using the average gain. The Gain implies the relative contribution of the corresponding feature to the model calculated by taking each feature's contribution for each tree in the model.

- **Multi-Layer Perceptron Regressor:** The MLP Regressor has been used to predict the onset time using a neural network-based algorithm.

C. PREDICTION FOR THE NEED OF HOSPITALIZATION

The idea of this goal to prevent deaths by predicting the need for hospitalization based on the symptoms seen during the time of reporting adverse events. We generated a sparse symptoms matrix encoded based on the MedDRA terms. Using supervised machine learning techniques, we tried to see which symptoms are contributing more to the need of hospitalization and predict whether a patient needs any hospitalization.

To generate the sparse matrix, we generated a dictionary with key coded for each distinct symptom. And in our data horizon, we saw that there are 9473 distinct symptoms. So, for each VAERS_ID, we have 9473 number of predictors and hospitalization tag as (Y/N). This is the base dataset for our model.

Random Under sampling for handling imbalanced dataset[4][5]:
We used random under sampling reduce the number of samples in the majority class. Random under sampling randomly selects samples of majority class from the feature space and then deletes them from the feature space.

SMOTE (Synthetic Minority Oversampling Technique) for handling imbalanced dataset [6][7]:
We used SMOTE to oversample the minority class of symptoms for which hospitalization was required. SMOTE oversamples the minority class by choosing samples in feature space which are close to each other. Here, SMOTE is used with the combination of random under sampling to balance the class distribution.

Sparse Naïve Bayes Feature Selection [3]:
We used Sparse Naïve Bayes Feature Selection method as the dataset has most the values of the columns as zeroes. We used the model with hyperparameter sparsity level set as 5. As the data is in the form of a Bernoulli distribution, we used the Naïve Bayes model with sparsity level set as 5 for Bernoulli distribution model.

Sparse PCA for Dimensionality Reduction [8][9]:
We used sparse PCA to solve the problem of high dimensionality space. First, we used sparse PCA to reduce the feature space from 9481 to 5, it wasn't producing convincing results as the data is highly imbalanced with having most of the patients who are not hospitalized. To solve this issue, we used the records of the patients who are hospitalized to obtain the principal components.

Singular Value Decomposition for Dimensionality Reduction [10]:
We used SVD for the same purpose as sparse PCA to reduce the feature space. The problem with sparse PCA is that it requires a lot of computation power, and it was time consuming, this problem is solved with Singular Value Decomposition.

VI. RESULTS

Goal 1: Prediction of Onset Time

The performance of all the Regression algorithms that have been used in this project are listed down in Table 1. The linear regression-based models show that predictors like **Hyperlipidemia, Arthritis, Anemia, CVD, Cancer** are the most important features which are used for predictions.

Whereas in case of Random Forest Regression model – the top 5 predictors that influence the predictions the most are **AGE_YRS, Allergies, other meds, Moderna vaccine, Pfizer vaccine**. The XGBoost model produced the best performance which produced an error of around 7 days. On observing the Fig 10 its evident that **AGE_YRS** is by far the most important feature as it influences the predictions in almost all the models.

Model	Training RMSE, Testing RMSE	Top Predictors
OLS	8.00, 7.97	Hyperlipidemia, Anemia, Arthritis, Cancer, CVD
Lasso	8.01, 7.98	Hyperlipidemia, Arthritis, Anemia, CVD, Cancer
Ridge	8.00, 7.96	Hyperlipidemia, Arthritis, Anemia, CVD, Cancer
Random Forest	6.23, 6.64	Age, Allergies, other meds, Moderna vaccine, Pfizer vaccine
XGBoost	3.40, 3.87	Pfizer vaccine, Age, Allergies, other meds, Sex_F
MLP Regressor	5.17, 5.43	

Table 1. Performance of Regression models in prediction of Onset Time

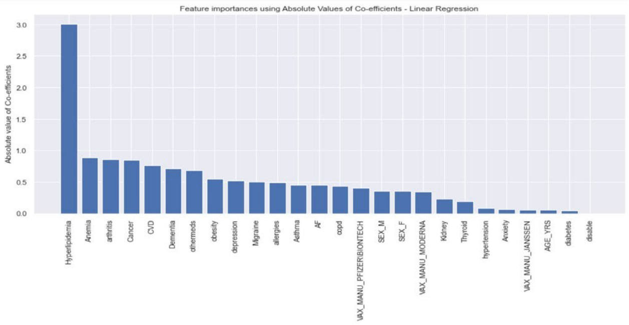

FIGURE 7. Feature Importance obtained using the Linear Regression model based on the absolute value of co-efficients

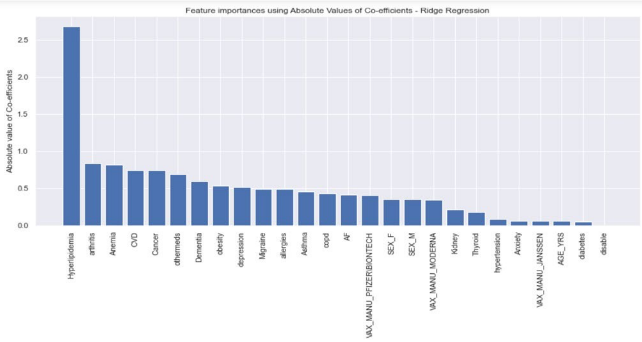

FIGURE 8. Feature Importance obtained using the Ridge Regression model based on the absolute value of co-efficients

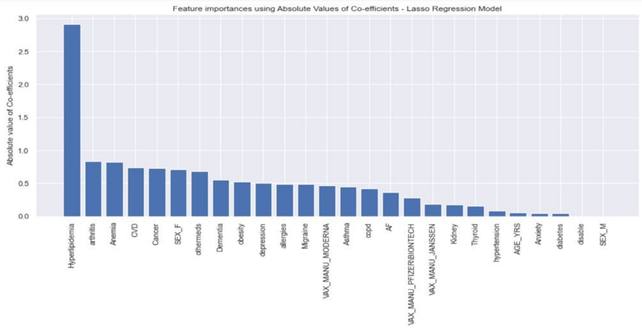

FIGURE 9. Feature Importance obtained using the Lasso Regression model based on the absolute value of co-efficients

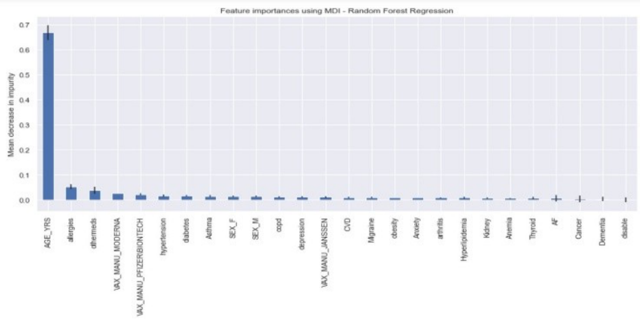

FIGURE 10. Feature Importance obtained using the Random Forest Regressor model based on the mean decrease in impurity

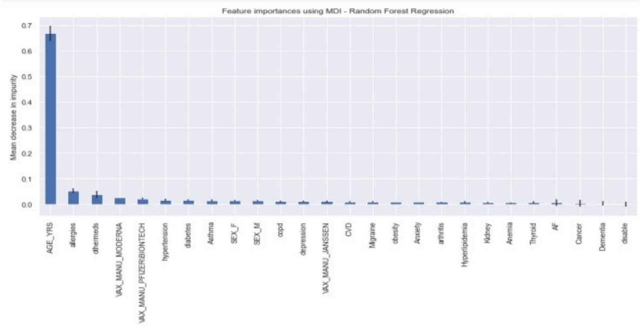

FIGURE 11. Feature Importance obtained using the XG Boost Regressor model based on average gain of the tree

Goal 2: Need for hospitalization prediction

Sparse Naïve Bayes Feature Selection

Metric	Logistic Regression	Random Forest	Linear SVM	XGBoostClassifier
Train Specificity	0.99	0.99	0.99	0.99
Train Sensitivity	0.20	0.23	0.20	0.23
ROC-AUC Score	0.64	0.64	0.64	0.64
Test Specificity	0.98	0.98	0.98	0.98
Test Sensitivity	0.20	0.20	0.20	0.20

Sparse PCA

Metric	Logistic Regression	Random Forest	Linear SVM	XGBoostClassifier
Train Specificity	0.88	0.96	0.89	0.88
Train Sensitivity	0.53	0.90	0.51	0.85
ROC-AUC Score	0.76	0.89	0.75	0.88
Test Specificity	0.88	0.87	0.89	0.86
Test Sensitivity	0.55	0.78	0.53	0.78

SVD

Metric	Logistic Regression	Random Forest	Linear SVM	XGBoostClassifier
Train Specificity	0.83	0.98	0.85	0.85
Train Sensitivity	0.65	0.98	0.62	0.87
ROC-AUC Score	0.83	0.91	0.83	0.91
Test Specificity	0.83	0.84	0.85	0.83
Test Sensitivity	0.68	0.84	0.65	0.84

VII. CONCLUSIONS

Based on our study objectives to work on both classification and regression methods, we came up with these two objectives for which we explored different models and their performance. We have seen that almost all the models have performed same for the onset time prediction. For the hospitalization, sparsity and imbalanced data were a challenge which we tackled using the sampling like SMOTE and Under Sampling and dimensionality reduction techniques like Sparse PCA, SVD and Spare Naïve Bayes

Feature Selection. We have seen that SVD is performing better than other models.

VIII. FUTURE WORK

- We can use some cloud-based hardware to run on larger datasets and more complex models.
- To handle the sparsity of data, more novel techniques in literature can be explored.
- Hyper parameter tuning for better accuracy.
- Introduce external datasets for better model building of prediction of onset time for adverse effects.
- We want to explore the model performance on other vaccines and fine-tune based on the observations seen on those datasets.

IX. REFERENCES

[1] Adverse effects of COVID-19 vaccination: machine learning and statistical approach to identify and classify incidences of morbidity and post-vaccination reactogenicity Md. Martuza Ahamad, Sakifa Aktar, Md. Jamal Uddin, Md. Rashed-Al-Mahfuz, AKM Azad, Shahadat Uddin, Salem A. Alyami, Iqbal H. Sarker, Pietro Liò, Julian M.W. Quinn, Mohammad Ali Moni
medRxiv 2021.04.16.21255618; doi: https://doi.org/10.1101/2021.04.16.21255618

[2] Mahdavi M, Choubdar H, Zabeh E, Rieder M, Safavi-Naeini S, Jobbagy Z, et al. (2021) A machine learning based exploration of COVID-19 mortality risk. PLoS ONE 16(7): e0252384. https://doi.org/10.1371/journal.pone.0252384

[3] Askari, A., d'Aspremont, A., & El Ghaoui, L. (2020, June). Naive feature selection: Sparsity in naive bayes. In International Conference on Artificial Intelligence and Statistics (pp. 1813-1822). PMLR.

[4] https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/

[5] Branco, P., Torgo, L., & Ribeiro, R. P. (2016). A survey of predictive modeling on imbalanced domains. ACM Computing Surveys (CSUR), 49(2), 1-50.

[6] https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/

[7] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

[8] Guerra-Urzola, R., Van Deun, K., Vera, J.C. et al. A Guide for Sparse PCA: Model Comparison and Applications. Psychometrika 86, 893–919 (2021). https://doi.org/10.1007/s11336-021-09773-2

[9] https://www.datatechnotes.com/2021/01/sparsepca-projection-example-in-python.html

[10] https://machinelearningmastery.com/singular-value-decomposition-for-dimensionality-reduction-in-python/

[11] https://vaers.hhs.gov/data/datasets.html?