Predicting Response Severity to COVID-19 Vaccines

Mariam Grigoryan*, Yejin Cha*, Gordon Kong*, Claire Flynn*

Computer Science Department, Colgate University

Hamilton, NY

*These authors contributed equally to this work

Abstract

So far, eligibility for various supply-constrained COVID-19 vaccines rely on identifying and vaccinating individuals based on broad demographic categories such as age, occupation, and past medical history. These groupings, however, cannot guarantee whether certain vaccines are suitable and low-risk for the small percentage of individuals who may expect serious side effects. Here, we use data on the demographics of patients to build machine-learning detection models that predict the seriousness of a future patient's reactions and determine their suitability or risk for vaccination. While the dataset used is somewhat limited due to the tendency for more extreme reactions to being reported more often than mild effects, the accuracy of the model suggests that the features can be effective predictors. As a result, our model can predict the worst-case scenario for a patient, that is, if a patient has a reaction, the model will predict how severe it will likely be. Since the consequences of under-predicting reactions are more detrimental than the consequences of our conservative over-predicting, we find our model to be acceptable, though imperfect. We find that the models predict non-serious findings at high accuracy, but demonstrate a tradeoff between predicting serious and nonserious symptoms. SVM predicted serious symptoms the best at 29%, but only predicted nonserious symptoms at 74%. The Random Forests model predicted serious symptoms at 26% but predicted nonserious systems at a much higher rate of 87%. Our findings can be of great benefit to determining important risk factors to consider when deciding on which vaccine is the most appropriate for individuals and whether certain individuals should be more closely monitored after vaccination. Being able to better anticipate adverse reactions on a case-by-case basis is essential to maintaining a successful rollout of vaccination, the only way to bring an end to the worldwide COVID-19 pandemic.

1. Introduction

The COVID-19 pandemic continues to drastically disrupt the world with unprecedented challenges to public health, the economy, and society. These negative distributions this pandemic has caused are unlikely to be resolved without meaningful and reliable treatments and cures to this terrible disease. The situation has caused countries around the world to develop, evaluate, and approve mRNA-based vaccines, a new type of protein-based vaccine technology against infectious diseases, in under a year, which is a success in vaccine development history. Given the rapid rate of distribution and demand for the vaccines over a short timeline, data on potential adverse effects for certain individuals are just beginning to grow.

Although recently developed vaccines from companies like Moderna and Pfizer have conducted clinical trials which have shown to meet intense safety and efficacy standards under the U.S. Food and Drug Administration (FDA) and Centers for Disease Control and Prevention (CDC) guidelines, the possibility for adverse side effects in a small percentage of individuals still exists. With millions of people lining up to receive the recently developed COVID-19 vaccines to reduce the risks of contracting COVID-19, detection of possible severe effects is vital.

The ability to predict the worst reactions to COVID-19 is particularly important due to the nature of this virus and the unique political environment that we find ourselves in. This illness targets and kills the most vulnerable in society, the elderly, and those with preexisting conditions. These populations are being prioritized in vaccination rollouts and are most likely to be seriously affected by any adverse reactions. In addition, during this pandemic, the country has seen an unprecedented wave of misinformation and vaccine hesitancy. Any method to predict severe reactions could provide opportunities to monitor those patients and treat the effects more efficiently. Confidence in the efficacy and safety of COVID-19 vaccines is essential to combating vaccine hesitancy and bringing this pandemic to an end as quickly as possible. As the country has seen recently, with the pause of the Johnson and Johnson vaccine, even rare reactions, if serious enough, can halt the vaccination of hundreds of thousands of people. The ability to predict these rare reactions could be very useful to prevent serious illness while allowing those not at risk to receive a potentially life-saving shot.

The Vaccine Adverse Event Reporting System (VAERS) under the FDA and CDC receives reports of adverse events from doctors and vaccine providers and documents them in a dataset that is updated biweekly. These reports are used to monitor whether the rate of adverse events of any vaccine is higher than the expected rate. So far, many people have experienced mild symptoms such as pain, fatigue, headaches, and chills. Others have experienced more adverse effects such as loss of consciousness and tinnitus (ringing in the ears), or even death.

Since doctors have an obligation to report any medical issue a patient experiences following vaccination, the range of symptoms is incredibly broad. Predicting specific reactions would be challenging, whereas classifying some symptoms as serious versus not serious is an easier task. For instance, death or loss of consciousness is clearly serious, while fatigue and diarrhea are very common side effects and generally not serious. Therefore, we want to focus initially on the severity of the response to the vaccine and explore its predictability. If we see success with predicting severity, we would like to attempt to predict categories of symptoms. Though the range of symptoms is broad, we have been able to split them into five general groups.

Throughout an 8-week project exploring and training the dataset, multiple machine-learning models have been explored to determine which model classifies vaccine response severity most accurately. Specifically, we look at SVM, decision trees, KNN, and Random Forest models and consider which features are significant and which features can be excluded.

The SVM model has demonstrated a recall rate of 74% for negative and 29% for positive labels. The Random Forests model demonstrated a recall rate of 87% for negative and 26% for positive labels. Moreover, the Decision Tree model demonstrated a recall rate of 90% for negative and 22% for positive labels. The strength of these models is an encouraging indication that models for variations of this prediction would be effective as well. In particular, it would be valuable to develop a model to even more specifically predict the general type of reaction an individual could contract or a model that could predict the brand of vaccine that an individual should use to avoid adverse effects. The most important takeaway from the strength of the various models is that the features in the dataset are strongly correlated with the strength of adverse effects.

2. Related Work

Prior COVID-19 machine learning studies exploited emerging datasets mostly from China and some from European countries. Many of these studies generated machine learning models from large amounts of data on COVID-19 compiled by medical services. Studies like Kang et al. (2021), Lassau et al. (2021), and Zoabi et al. (2021), all process large amounts of data from medical imaging and other clinical data made available from China and European countries. Unlike past works, our work studies a more recent phenomenon of COVID-19 vaccination that is happening in the United States.

The previous COVID-19-related studies have been focusing on developing models to accurately diagnose COVID-19. For instance, Zoabi et al. (2021) have developed a model to accurately predict COVID-19 using simple features such as sex, age, known contact with an infected individual, as well as five initial clinical symptoms. Their goal was to have their framework 'assist medical staff worldwide in triaging patients, especially in the context of limited resources' [Zuabi et al. (2021)].

Some other studies more similar to our study focused on COVID-19 severity classification. For instance, Kang et al. (2021) developed a COVID-19 severity classifier in people who tested positive. They established criteria for separating out symptoms in terms of mild, moderate, severe, and critically ill severity. Another study done by Yao et al. (2020) built a different severity detection model, basing their model on blood and urine test results. Additionally, Lassau et al. (2021) trained a deep learning model based on chest computerized tomography scans and predicted a severity score using other clinical and biological variables. Similar to Kang's study of severity detection, our study also outlines and separates symptoms based on severity.

Estiri et al. (2021)'s study on predicting COVID-19 mortality involves training age-stratified generalized linear models based on more widely available past medical information in electronic health records to understand the differences in risk factors across various age groups. The motivation of that study relates the most to our study of COVID-19 vaccination, as one of our goals is to help efficiently allocate resources such as vaccination, to the general public.

An abundance of machine-learning studies already exists on classifying the severity of COVID-19 or detecting COVID-19. By focusing on COVID-19 vaccines, we delve into a more recent and barely explored field of COVID-19 that is just as important in the present context. In this paper, we propose a machine-learning model that accurately classifies the patients' responses to the COVID-19 vaccines as serious or not serious.

3. Methods

3.0 Data Description and Analysis

The dataset we are working with comes from the VAERS database under the FDA and CDC. The data comprises 33,194 identification numbers for individuals across the US from the period January 1st, 2021 through March 19th, 2021. These identification numbers match individual demographic data such as age, sex, state to an encoded list of symptoms recorded by doctors and other vaccine providers. Data on the vaccine received is also matched. The dataset consists of three different datasets, the first one including the demographics information(2021VAERSDATA.csv), the second including the symptoms(2021VAERSSYMPTOMS.csv), and the third one including the vaccination data such as the vaccine type, manufacturer, the dose series, and vaccination site (2021VAERSVAX.csv).

The following list displays each of the dataset's features we included in the model:

A. Basic demographic data
    a. Age
    b. Sex
B. Vaccine information
    a. Vaccine manufacturers
    b. Vaccine dose series
C. Other information:
    a. Other medications
    b. History
    c. Allergies

Table 1. Characteristics of the dataset used by the model in this study

Feature	Total n = 34174		Not serious n = 24238		Serious n = 9936	
	n	%	n	%	n	%
Sex						
Male	8768	24.66	5756	23.75	3012	30.31
Female	24567	71.89	17785	73.38	6782	68.26
Unknown	839	2.46	697	2.88	142	1.43
Other Meds						
True	17512	51.24	11772	48.57	5740	57.77
False	16662	48.76	12466	51.43	4196	42.23
Med History						
True	17236	50.44	11203	46.22	6033	60.72
False	16938	49.56	13035	53.78	3903	39.28
Allergies						
True	13271	38.83	8656	35.71	4615	46.45
False	20903	61.17	15582	64.29	5321	53.55
Vaccine Manufacturers						
Pfizer	16534	48.38	11302	46.63	5232	52.66
Moderna	16513	48.32	11978	49.42	4535	45.64
Janssen	1109	3.25	950	3.92	159	1.60
Unknown	18	0.53	8	0.03	10	0.10
Vaccine Doses						
0	5255	15.38	3982	16.43	1273	12.81
1	22776	66.65	16310	67.29	6466	65.07
2+	6143	17.98	3946	16.28	2197	22.11

* Nominal features are not shown in the table

Below we include a few figures describing the merged full data.

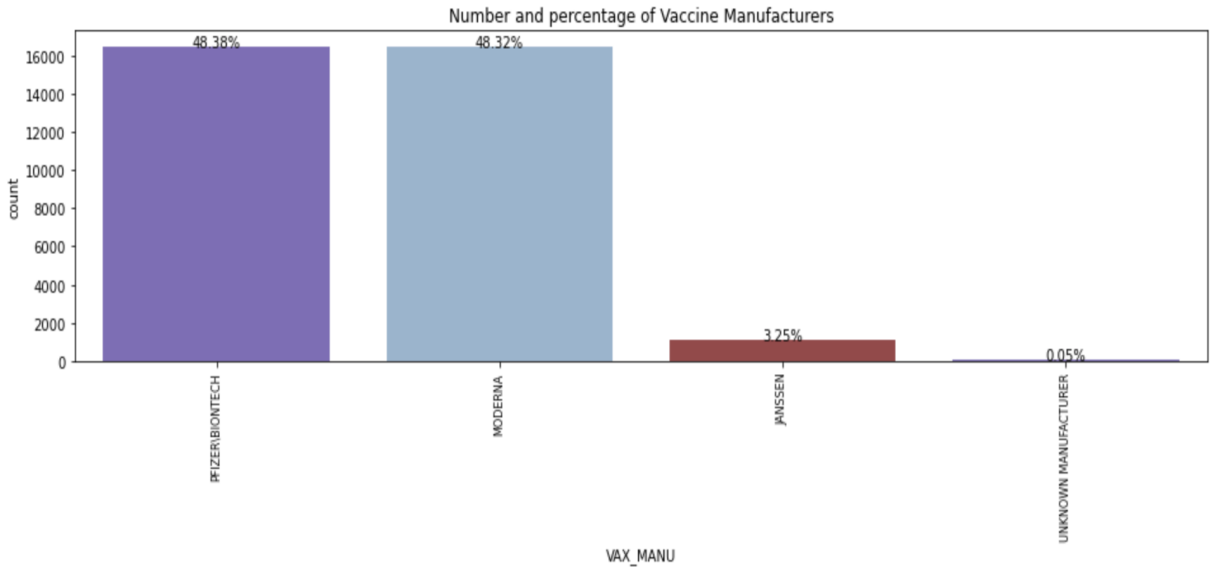

Figure 1: Distribution of vaccine manufacturers

Figure 1 shows the distribution of vaccine manufacturers in the dataset. Not surprisingly, the majority of the reported symptoms were from patients who received Pfizer-BioNTech or Moderna, which comprises the majority of cases and we can observe an almost equal distribution of those two vaccines.

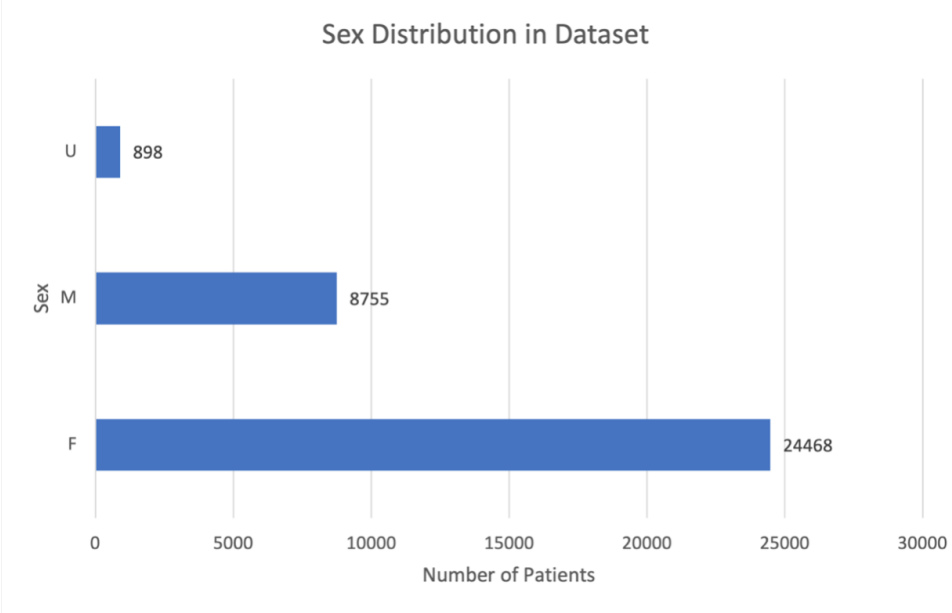

Figure 2: Gender Distribution in Reported Cases

Figure 2 shows the gender distribution in the reported cases in the dataset. Interestingly, we can notice that there was more than twice the number of female cases reported than male cases. It is important to note that this can make out data biased.

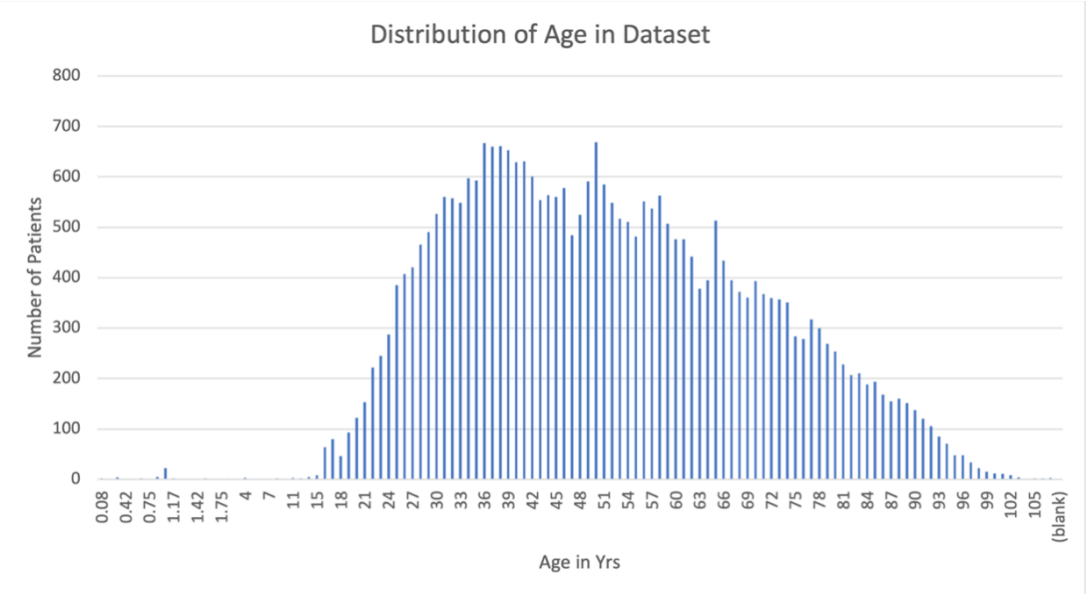

Figure 3: Age Distribution in Reported Cases

Figure 3 shows the age distribution in the reported cases in the dataset. We can notice a relatively normal distribution of ages, although there is a little bit of skew as well.

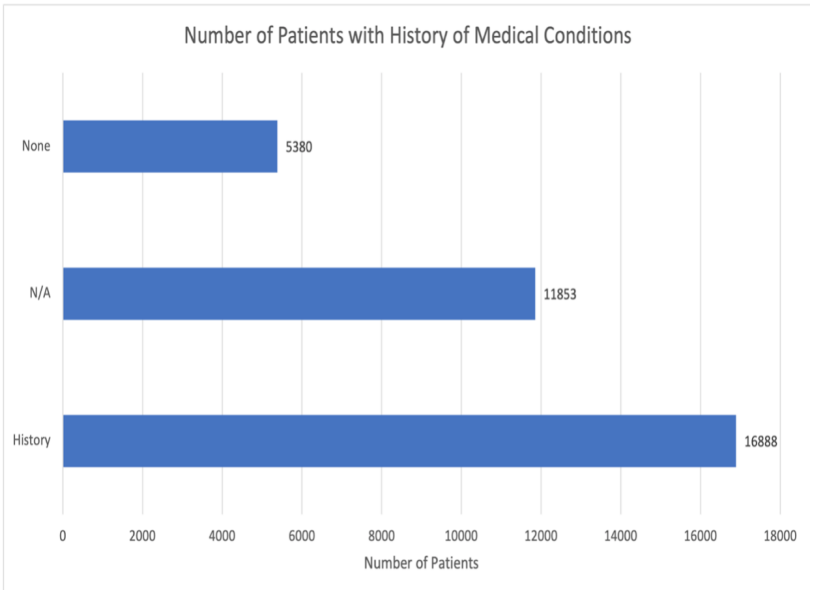

Figure 4: Number of patients with a history of medical conditions

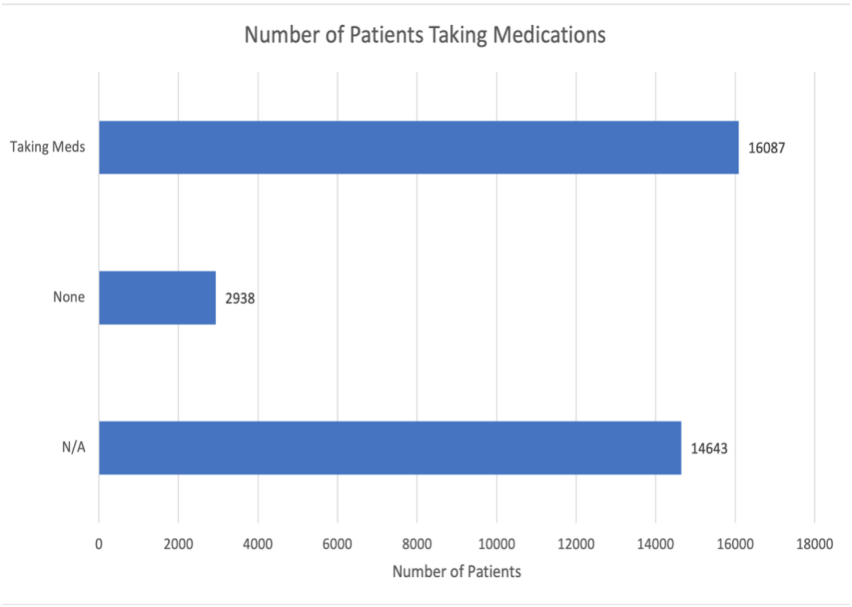

Figure 5: Number of patients taking medications

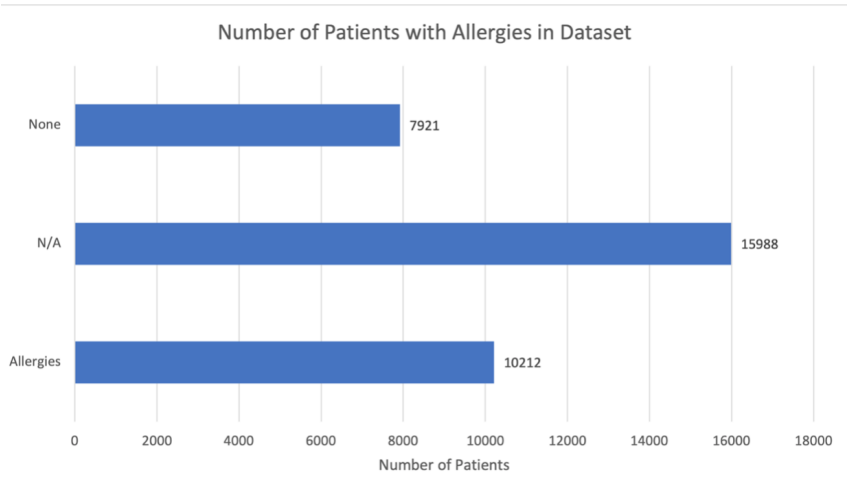

Figure 6: Number of patients with allergies in the dataset

Figures 4-6 show the distribution of patients with a history of medical conditions, patients taking medications, and patients with allergies, respectively. From Figures 5 and 6 we can notice an equal distribution of patients with histories of medical conditions versus not reporting any, as well as equal distribution of patients on medications versus not taking any. Figure 6 shows more than twice more people with allergies as without.

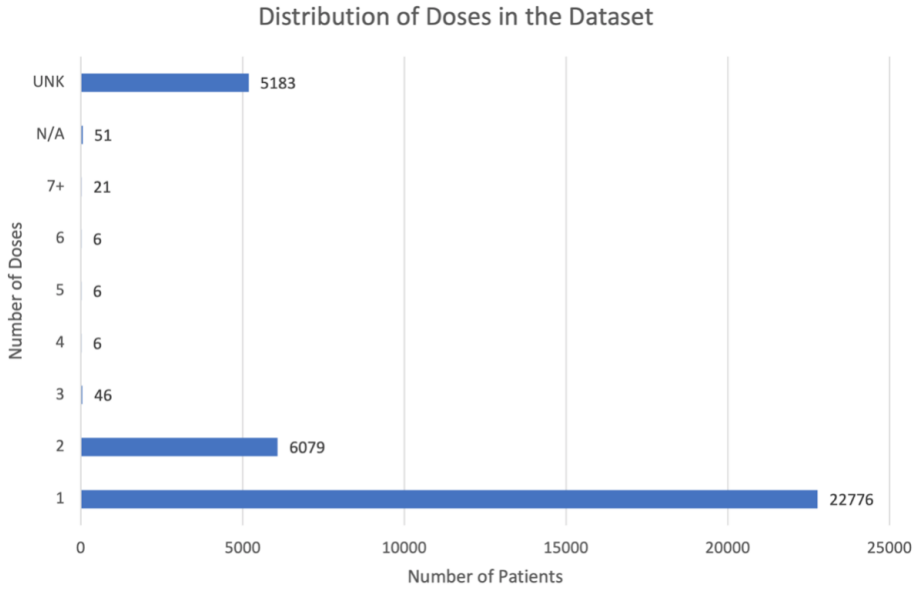

Figure 7: Distribution of the vaccine doses in the dataset

Figure 7 shows the distribution of the vaccine doses in the dataset. The majority of the distribution consists of patients reporting after receiving the first dose, followed by the second dose reports. The data also consists of 5183 unknown values and some unusual doses above 2.

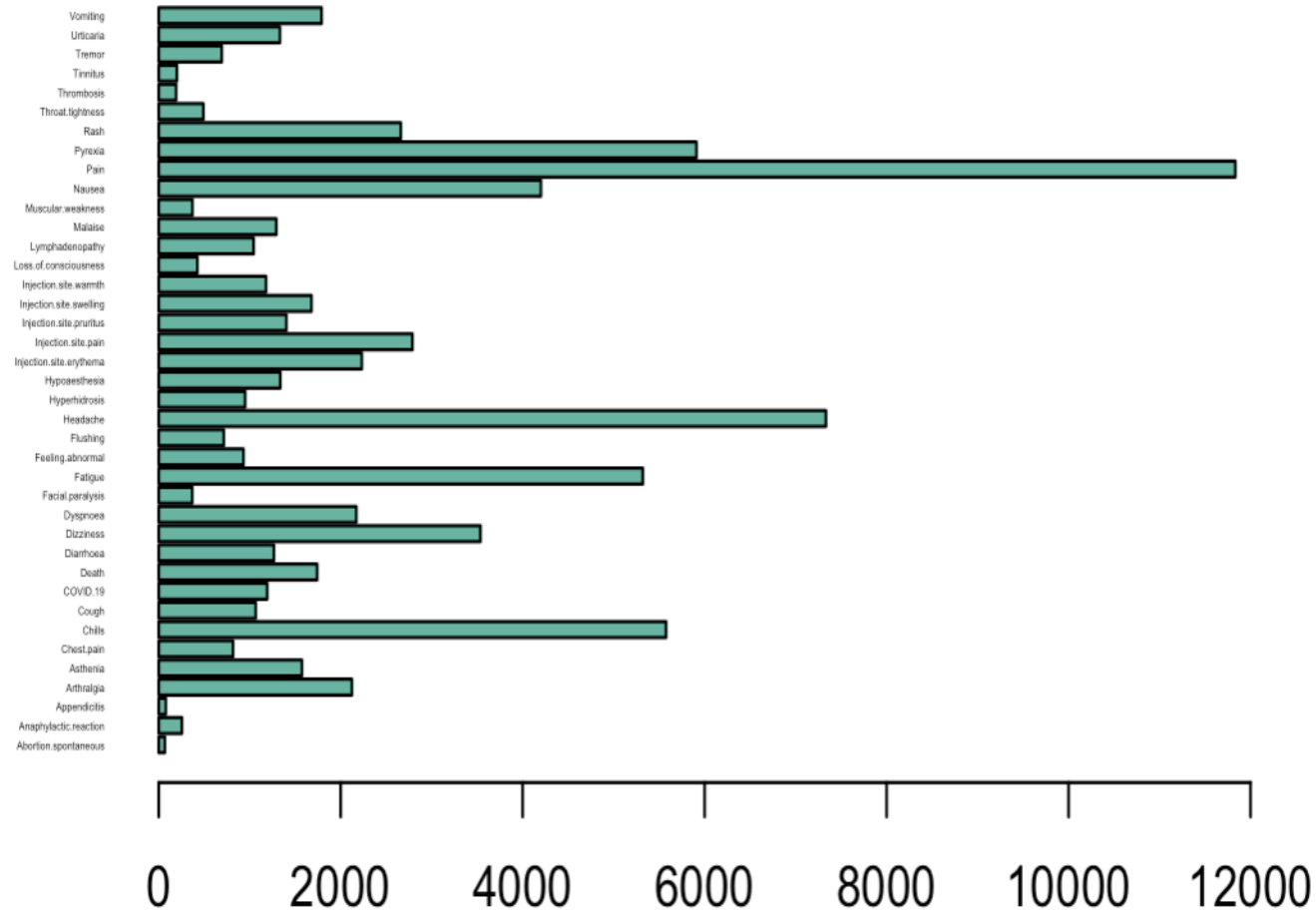

Figure 8: Occurrence of Symptoms

Figure 8 shows the distribution of symptoms in our dataset. Up to 90% of the data comes from reports concerning relatively minor symptoms, such as pain at the injection area, fatigue, or fever. The remainder describes serious symptoms that are life-threatening. The most-reported symptom was pain, followed by headache, pyrexia, chills, fatigue, dizziness, injection site pain, rash, vomiting, arthralgia (joint pain), nausea, and some other least frequent symptoms such as facial paralysis, flushing, tremor, asthenia, spontaneous abortion, chest pain, anaphylactic reaction, appendicitis, arthralgia, malaise, tinnitus, hypoaesthesia, etc.

3.1 Data Preprocessing

First, cases not attributed to COVID-19 vaccines were filtered. This includes around 3% of the

reported vaccinations that were related to flu.

Then, the missing entries were deleted or made false depending on the category. NaN values for categorical variables such as emergency room visits or birth defects were assumed to be false.

Next, redundancies in the dataset were removed. For instance, the dataset included three different columns of age calculated in three different ways, which were removed.

Since many patients never reported their age, it was necessary to assign all those patients the dataset's mean age. In addition, three categorical components of the demographic information were essential to our models. The patients' medical histories, medications, and allergies were evidence of pre-existing conditions that may influence the likelihood of severe reactions. Therefore, for these features, if the patients recorded 'none', 'no', or 'n/a', we changed the values to 0, and, if patients recorded any conditions, allergies, or medications, we changed the values to 1. Another important feature from the demographic subset was the number of vaccination doses the patient had received. While the vast majority of the subjects reported 1 or 2 doses, some have reported more doses of the COVID-19 vaccines than doctors are supposed to administer or have received an unknown number. For those who did not report a dosage, we assigned 0 doses. While, for subjects who reported more than 2 doses, we assumed that they received the maximum number of doses possible and assigned them as 2 doses. In order to better organize for our first model and prepare for further models, the five different columns of symptoms, which were written in different ways, were merged and combined into the final features listed in Figure 4.

The dataset included 5 different columns of symptoms, which oftentimes phrases exactly however the doctor wrote them down. This resulted in some symptoms being written in different ways. Figure 9 shows the symptoms reported in the dataset before any preprocessing is done. To tackle this, similar symptoms recorded in slightly different wording were modified to be under the same category to remove redundancies. The five different columns of symptoms written in different ways were merged and combined into the final features listed in Figure 10. An example of this includes combining the phrases including the common word 'pain' all into the same category of 'pain'. This resulted in us extracting 40 different categories of symptoms, which we then included each of those in a separate boolean column specifying whether the patients exhibited those symptoms or not.

VAERS_ID	SYMPTOM1	SYMPTOM2	SYMPTOM3	SYMPTOM4	SYMPTOM5
916600	Dysphagia	Epiglottitis			
916601	Anxiety	Dyspnoea			
916602	Chest discomfort	Dysphagia	Pain in extremity	Visual impairment	
916603	Dizziness	Fatigue	Mobility decreased		
916604	Injection site erythema	Injection site pruritus	Injection site swelling	Injection site warmth	
916606	Pharyngeal swelling				
916607	Abdominal pain	Chills	Sleep disorder		
916608	Diarrhoea	Nasal congestion			
916609	Vaccination site erythema	Vaccination site pruritus	Vaccination site swelling		
916610	Rash	Urticaria			
916611	Blood pressure decreased	Chest pain	Chills	Confusional state	Decreased appetite
916611	Dyspnoea	Fatigue	Feeling abnormal	Head discomfort	Headache
916611	Heart rate decreased	Heart rate increased	Hypertension	Injection site pain	Musculoskeletal chest pain
916611	Nausea	Pain	Pain in extremity	Paraesthesia oral	Pyrexia
916611	SARS-CoV-2 antibody test	SARS-CoV-2 test negative			

Figure 9: Symptoms data before data preprocessing

```
> symptomsOfInterest
 [1] "Abortion spontaneous"    "Anaphylactic reaction"   "Appendicitis"            "Arthralgia"
 [5] "Asthenia"                "Chest pain"              "Chills"                  "Cough"
 [9] "COVID-19"                "Death"                   "Diarrhoea"               "Dizziness"
[13] "Dyspnoea"                "Facial paralysis"        "Fatigue"                 "Feeling abnormal"
[17] "Flushing"                "Headache"                "Hyperhidrosis"           "Hypoaesthesia"
[21] "Injection site erythema" "Injection site pain"     "Injection site pruritus" "Injection site swelling"
[25] "Injection site warmth"   "Loss of consciousness"   "Lymphadenopathy"         "Malaise"
[29] "Muscular weakness"       "Nausea"                  "Pain"                    "Pyrexia"
[33] "Rash"                    "Throat tightness"        "Thrombosis"              "Tinnitus"
[37] "Tremor"                  "Urticaria"               "Vomiting"
```

Figure 10: All reported symptoms

Next, adverse symptoms were combined and one-hot encoded for binary classification. In addition, we transformed the columns including pre-existing conditions, other medications, and allergies into a binary column specifying whether the patient reported that or not. Next, we modified the vaccine dose series to filter out unusual values. The doses 0,1, and 2 were left as is; the missing values were converted to 0, and any dosage reported above 2 was converted to 2.

Furthermore, the VAERS dataset does not provide severity categories or labels, which we needed for supervised learning. Therefore, a column of severity labels was created based on the features that we decided should be considered serious, such as the ones assumed to be life-threatening. The features listed below were assumed to be serious:

- Abortion spontaneous
- Anaphylactic reaction
- Appendicitis
- Death
- Dyspnea (shortness of breath)
- Facial paralysis
- Loss of consciousness
- Lymphadenopathy
- Thrombosis
- Tinnitus (a condition affecting hearing)
- ER Visit

3.2 Model Building

Predictions were generated using four different models. The first model we used is Support Vector Machines (SVM), which are very effective in high dimensional spaces. The downside of it is that the cross-validations can be expensive. The second model we used was Decision Trees. The advantage of using Decision Trees is that it performs classification decently with little data preparation because the model is able to handle both numerical and categorical forms. One downside is that the model might create trees that do not generalize well. The third model we consider using is k-Nearest Neighbors (KNN), which are generally effective with large data sets and are robust to noise. A major con, however, is that there is a high computation cost associated with finding the optimal k, the

number of nearest neighbors to include in the model. The last model we consider is Random Forest, which is associated with a reduction in over-fitting but is a slow and complex prediction process.

The following features were used in training our models:

- Age
- Sex
- Other medications
- History of pre-existing conditions
- Allergies
- Vaccine manufacturer
- Vaccine dose

Symptoms and ER Visit information were not included in our features list because otherwise, that would make our models trivial since we already used those two columns in creating the labels.

3.2.1 Support Vector Machine (SVM)

To enumerate model selection and training options available for SVM and find the best combination for our data, we performed a grid search using GridSearchCV with a 5 fold cross-validation. We considered multiclass types, SVM model types, and one of the SVM hyperparameters, margin violation weight 'c'. Multiclass types we tested include one-vs-one(ovo) and one-vs-rest(ovr). We tested linear SVM and kernel SVM types, such as polynomial kernel, radial basis function kernel, and sigmoid kernel. Lastly, we tested 0.1, 1, 10, 100, and 1000 for our 'c' value.

The resulting dataframe displays various combinations of our testing hyperparameters along with their average recall rates. Using our best parameters, we retrained the model on the full training dataset and conducted a test set prediction.

3.2.2 Decision Trees

Using Scikit-Learn's built-in GridSearchCV class, we performed a grid search with 10-fold cross-validation for the Decision Tree model.

We chose the following hyperparameters to test: max depth, criterion, and min samples split. Max depth is the maximum number of child nodes that can grow out of the decision tree before the tree gets cut off. Here, we set the numbers to 5, 10, and 20. Criterion is used to measure the quality of a split. Default is set as "gini," and we added "entropy" as a second option. "Gini" measures Gini impurity and "entropy" measures information gain. Min samples split is the minimum number of samples needed to split an internal node. Default is 2, and we added 5 and 10 as additional options.

3.2.3 KNN

Next, we developed a K-nearest neighbors model using a GridSearchCV class with 10-fold cross-validation on the number of neighbors and weights. In the GridSearchCV, we set the number of neighbors to a range between 1 to 50. The weight was decided between the uniform or distance weight options.

3.2.4 Random Forest

The last model we created is the Random Forest Model. Finding the optimal hyperparameters was conducted through a RandomizedSearchCV on three features: the number of estimators, max features, and max depth. Here, we searched the n_estimators starting at 200 and stopping at 2000. The max_features considered were 'auto' and 'sqrt'. Max depth values considered ranged from 100 to 500.

3.3 Model Evaluation

The models were evaluated based on the test set classification accuracy metric. To compare our models, we considered accuracy, precision, and recall as our metrics. Although accuracy might be the metric most understandable to the general public, it is actually not the best measure to use in our models, since we have an imbalance classification problem, meaning that the positive class is greatly outnumbered by the negative class. Hence, evaluating precision and recall metrics could add more information about the model performance. Precision, which is the number of true positives divided by the sum of true positives and false positives, tells us the ability of our model to return only relevant instances. In addition, recall, which is the number of true positives divided by the sum of true positives and false negatives, is the metric we seek to maximize since we want our model to detect all the serious cases for certain. In addition, confusion matrices were utilized to see where the errors are happening.

4. Results

4.1 Model Building

For each model, we divided our pre-processed data into train and test sets using an 80/20 train test set ratio. After splitting the data, we ended up with 27,339 examples in the training set and 6,835 in the test set. Of the total symptoms, about 29% of the symptoms are serious. Then, we standardized it based on the training set using StandardScaler to ensure that the data is internally consistent for comparison purposes.

4.1.1 Support Vector Machine (SVM)

Our SVM model took approximately 5 hours to complete with 5 fold cross-validation. Based on the results, one-vs-one, sigmoid kernel, and c>=0.1 is the best combination of hyperparameters. As shown in the figure, our model has a recall rate of 74% for negative cases and 29% for positive cases. Moreover, the average accuracy of such a combination is 0.61.

	precision	recall	f1-score	support
0	0.71	0.74	0.73	4785
1	0.33	0.29	0.31	2050
accuracy			0.61	6835
macro avg	0.52	0.52	0.52	6835
weighted avg	0.60	0.61	0.60	6835

Figure 11a: Summary of SVM Classification Results

Figure shows the summary of the SVM classification using the best hyperparameter combination.

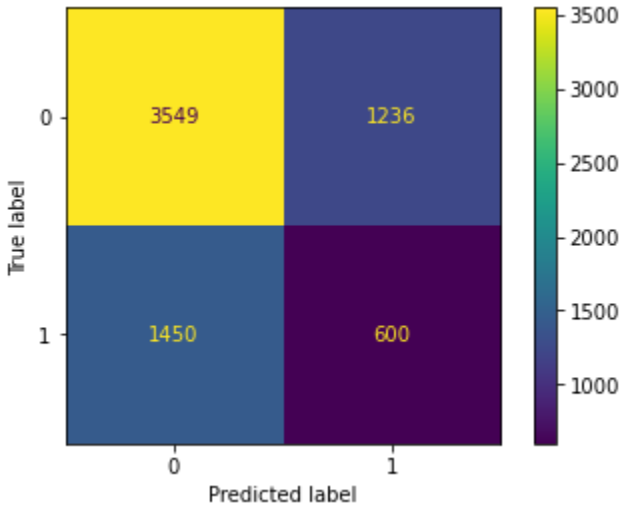

Figure 11b: Confusion matrix of SVM classification model

4.1.2 Decision Trees

Based on the results, using the maximum depth of 20 and minimum samples split of 10 returned the best recall value of 22%. Both 'gini' and 'entropy' functions return the same recall values, so we know that criterion parameter does not have an effect on our result. The overall accuracy of our model was 70%.

	precision	recall	f1-score	support
0	0.74	0.90	0.81	4871
1	0.47	0.22	0.30	1964
accuracy			0.70	6835
macro avg	0.60	0.56	0.55	6835
weighted avg	0.66	0.70	0.66	6835

Figure 12a: Summary of Decision Tree model results

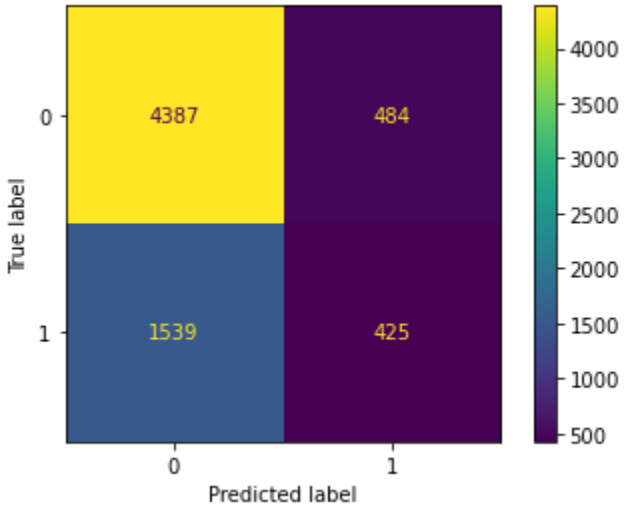

Figure 12b: Confusion matrix of Decision Tree classification model

4.1.3 KNN

	precision	recall	f1-score	support
0	0.72	0.99	0.83	4871
1	0.57	0.03	0.05	1964
accuracy			0.71	6835
macro avg	0.64	0.51	0.44	6835
weighted avg	0.68	0.71	0.61	6835

Figure 13a: Summary of KNN Classification Results

The optimal hyperparameters for the KNN model based on recall scoring ended up being 1 number of neighbors with uniform weights. The accuracy metric was 0.71 or 71%. The recall score for negative cases was 99%, while the recall score for positive cases was 3%.

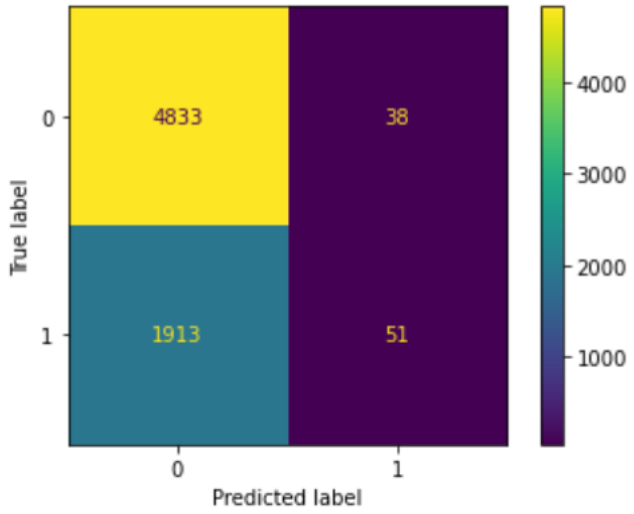

Figure 13b: Confusion matrix of KNN classification model

4.1.4 Random Forest

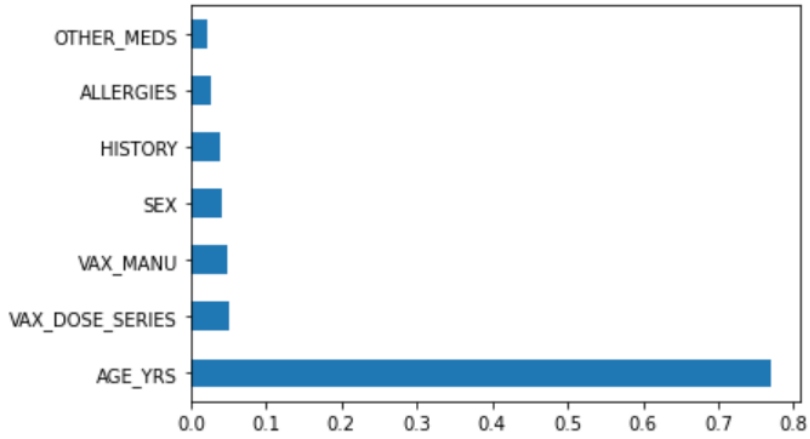

Figure 14: Feature importance for Random Forest

Figure shows the feature importance for the Random Forest model. We can see that age is the dominant feature.

```
=== Classification Report ===
              precision    recall  f1-score   support

           0       0.75      0.87      0.80      4871
           1       0.45      0.26      0.33      1964

    accuracy                           0.70      6835
   macro avg       0.60      0.57      0.57      6835
weighted avg       0.66      0.70      0.67      6835
```

Figure 15a: Summary of Random Forest Classification Results

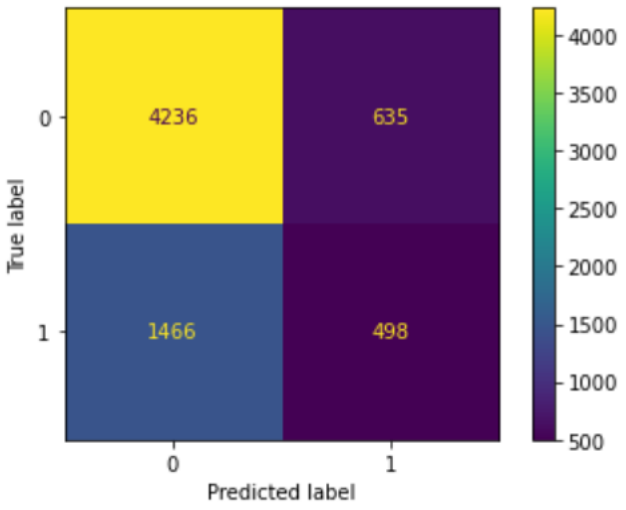

Figure 15b: Confusion matrix of Random Forest Classification model

From the RandomizedSearchCV, the most optimal hyperparameters are the following: Number of estimators: 1400; Max features: 'auto'; and Max depth: 460. Figure shows that the accuracy score is 70%, and the recall score for negative cases and positive cases is 87% and 26%. respectively.

4.1.5 Model Comparisons

Overall, the SVM model predicted serious symptoms the best at 29%. Random forest had slightly lower results for predicting serious symptoms at 26%. Random forest predicted non-serious symptoms more accurately at 87%, compared to SVMs score at 76%. Though the KNN model predicted non-serious systems at 99%, prediction for serious systems was 3%. Decision Tree predicted serious and non-serious symptoms at 22% and 90% respectively. In general, there seems to be a tradeoff between predicting serious symptoms and non-serious systems.

5. Discussion

There were several challenges faced throughout our project, the solutions to which had important ethical and technical consequences for our finished model.

The challenges were largely encountered during the data-preprocessing, which took the majority of our time and required some distinct skills and thoughtful decision-making. Initially, we were trying to predict the seriousness of COVID-19 vaccine response by using both the features we ultimately trained with and the symptom categories we developed. Not surprisingly, we achieved 99% accuracy in training the simplest SVM models. However, our group realised that we had made the predictions trivial by including the symptoms as features, since we already used some of those symptoms while creating the severity labels. Hence, we removed all the symptoms from our final feature list and achieved more reasonable accuracy scores.

In addition, during preprocessing, we had to make various conceptual decisions throughout this project in order to make it manageable and time-efficient, but also useful. For example, one of the decisions we made in the beginning of the project was switching from predicting 'severity' of the vaccine response to 'seriousness'. Since most symptoms can themselves be classified as mild, moderate or severe, it would be very difficult to interpret which case was reported by the patient, whereas very few symptoms incorporate the 'seriousness' in them, making creating the labels a little easier.

Due to the complexity of the unsupervised learning approach, the large size of our dataset, as well as the time constraint, we opted in for using a supervised learning approach. This itself created a major problem, since the VAERS dataset does not provide 'seriousness' categories, and thus we had to create those labels ourselves. After some medical research, we created a list of symptoms that we considered most serious. If a patient exhibited any of those symptoms, the 'seriousness' label would be 1 and if the patient did not, the label would be 0. While our symptom selections were based on our educated guesses, it is important to note that **no members of our group are certified medical professionals.** Therefore, it would be problematic and incredibly unethical, if our model were to be used in real life in order to decide which patients should receive more attention. Our project can be used as evidence for the viability of similar models, but those future models should consult with appropriate medical specialists.

Another potentially contentious decision we made was to include ER visits as a determination of 'seriousness' in addition to the symptoms described above. In many cases, a patient receiving care at the ER is an indication of the most severe and serious medical emergencies. For this reason, we decided that all ER visits be assigned a 1 for the 'seriousness' label. However, it is important to note that not all patients who end up in the ER are in the midst of a medical emergency. Due to the nature of the US healthcare system, there are many people living in the US without sufficient medical insurance. For those who do not have a general physician or cannot afford regular medical care, the ER may be the only option. This is an imperfect indicator of 'seriousness', but, to avoid excluding the worst of the reactions, we have decided it better to utilize it than to not.

In addition, dealing with missing values in our features was challenging since we had to decide whether we wanted to remove those examples or replace them with reasonable values. For instance, 3036 patients did not report their age in the dataset, which was about 9% of the 34174 total patients. Since 9% is a decent amount, we decided to replace these missing age values with the mean age values. On the other hand, some patients did not report any allergies or previous history of diseases, which we replaced with False instead. Another interesting problem we faced was during filtering out the vaccine dosage column. A few patients had reported more than two doses of the COVID-19 vaccines, some as many as seven, even though doctors only recommend two shots. Since such reckless over vaccination seems unlikely, we assumed these inputs were probably due to human error during reporting. We decided to assume that these subjects had received a second dose since they were reported to have received more than one.

The model training process, while more straightforward than preprocessing, was incredibly time consuming. First, we had to decide what models we wanted to train and pick the ones with the best potential, which required some careful consideration of pros and cons of each type of the model. However, even after analyzing the options, it was necessary to experiment with multiple methods and multiple hyperparameters. Due to the size of our dataset and the nature of the algorithms, especially SVM due to its expensive cross-validations, each model took considerable time to run.

It is important to acknowledge that, even ignoring the preprocessing and training decisions made by our group, the effectiveness of our model is dependent on the data reported to the government. Though there is a wide range of ages represented in the dataset, the majority of reports are from female patients. This bias definitely affects the behavior of our model. However, it is not yet clear if female subjects are more likely to experience reactions to the COVID-19 vaccines or if they are just more likely to report them. Though there is research that suggests both, there is currently no scientific consensus. In addition to the obvious sex bias that we see in the VAERS database, there is also a more subtle issue to consider. While doctors are ethically obligated to report any medical conditions that develop after a patient receives vaccination, as we discussed previously, not all people who need medical attention have access to it. This unequal access to treatment, and therefore reporting, creates an inherent bias in the dataset that our model is trained on. Though there is no easy, feasible solution to this issue, it is still important to be aware of when considering its effectiveness.

6. Conclusion

In conclusion, based on CDC and FDA Vaccine Adverse Event Reporting System data, we developed multiple models for predicting response severity to COVID-19 vaccines by using basic demographic and previous medical history features. Our framework can be of benefit to determining important risk factors to consider when deciding on which vaccine is the most appropriate for individuals and whether certain individuals should be more closely monitored after vaccination. In addition, the methodology presented in this study may benefit the health system response to different COVID_19 vaccine distributions.

There are various extensions to our project that can be done in the future. First, we can revisit the unequal gender distribution in the dataset, evaluate the bias and run the models separately on the two genders reported to see if there is any correlation between gender and the seriousness of the response to the vaccines. Next, we could test the two main vaccines, Pfizer and Moderna, reported in the dataset and try to find any differences between predicting seriousness of the vaccine responses. Furthermore, as the data is being updated biweekly, we could further incorporate the data associated with the Johnson & Johnson vaccine into our model training and testing, as more people start reporting their symptoms after getting that vaccine. Lastly, we could shift our focus on the first versus second doses separately and see if there are any significant differences in the seriousness of the responses reported.

7. References

Estiri, H., Strasser, Z. H., Klann, J. G., Naseri, P., Wagholikar, K. B., & Murphy, S. N. (2021). Predicting COVID-19 mortality with electronic medical records. *NPJ digital medicine*, *4*(1), 1-10.

Kang, J., Chen, T., Luo, H., Luo, Y., Du, G., & Jiming-Yang, M. (2021). Machine learning predictive model for severe COVID-19. *Infection, Genetics and Evolution*, *90*, 104737.

Lassau, N., Ammari, S., Chouzenoux, E., Gortais, H., Herent, P., Devilder, M., ... & Blum, M. G. (2021). Integrating deep learning CT-scan model, biological and clinical variables to predict severity of COVID-19 patients. *Nature communications*, *12*(1), 1-11.

Yao, H., Zhang, N., Zhang, R., Duan, M., Xie, T., Pan, J., ... & Wang, G. (2020). Severity detection for the coronavirus disease 2019 (covid-19) patients using a machine learning model based on the blood and urine tests. *Frontiers in cell and developmental biology*, *8*, 683.

Zoabi, Y., Deri-Rozov, S., & Shomron, N. (2021). Machine learning-based prediction of COVID-19 diagnosis based on symptoms. *npj Digital Medicine*, *4*(1), 1-5.