MSBD5012 Machine Learning

Imploring the Adverse Events of Covid-19 Vaccines using the VAERS dataset and discerning widespread false information on related vaccinations

Chan Yiu Chung (ycchanau@connect.ust.hk, 20430665)
Lam Chun Ting Jeff (ctjlam@connect.ust.hk, 12222973)
Mak Chun Wai Michael (cwmakah@connect.ust.hk, 20801333)
Ngerng Sherilynn Siew Fong (ssfngerng@connect.ust.hk, 20786961)

1 Introduction

Vaccination has become the primary proposition towards surmounting the wide spread Covid-19 pandemic. However, with the influence of rampant information sharing over the internet, the realisation of herd immunity through mass vaccination is restricted as false information towards complications after vaccine administrations have been pervasive. Henceforth, the US government, namely the Centers for Disease Control and Prevention (CDC) and the Food and Drug Administration (FDA), have released publicly accessible data on the adverse effects after vaccinations have been administered. Therefore, the objective of this project is to gain hands-on insight towards the adverse events from the three mRNA-based vaccines, which are the Pfizer-BioNTech, Janssen and Moderna vaccines and perhaps contribute to the efforts on combating the widespread of false information from volatile online sources.

To gain better insights for deduction between true information and commonly spread false information, the following objectives are proposed to analyse the VAERS dataset:

1. The onset time of adverse events is studied and prediction models are carried out to identify certain high risk populations and ascertain key medical developments after vaccinations have been administered.
2. Identify predominant traits on hospitalization symptoms and detect trends on proposed treatment towards prominent adverse events after vaccination administration.

2 Related Work

Various research work has been published online towards implementing machine learning to predict outcomes related to vaccination and its adverse events. A general procedure for predicting immunity rates and reactogenicity upon vaccination is established by Gonzalez-Dia et al.[1] through data processing, feature selection, algorithm selection and model testing. Another study is proposed by Ahmad et al. [2] on using SVM, gradient boost, decision tree and random forest models to classify significant features that contribute to death a

fter vaccination. The researchers encountered a significant problem towards a sparse symptom input feature and discussed that the machine learning model’ s performances were impacted due to this.

Hence, after much research into handling sparse input features, we have decided on applying a logistic regre ssion model with sparse PCA [3] to summarize the sparse input features. To compare this model’ s perfor mance, we also implement a sparse Naive Bayes model to pre-process the sparse features. [4]

3 Dataset

The Vaccine Adverse Event Reporting System (VAERS), which was established by the U.S. Department of Heal th and Human Services (DHHS) and is currently managed by the Food and Drug Administration (FDA) and Cent ers for Disease Control and Prevention (CDC). The systems will collect the record of all suspected adverse event s related to any U.S. licensed vaccine product. This system aims at detecting any possible side-effects of vaccine s that are missing during the testing stage. Health care providers and vaccine manufacturers are obligated to repo rt all the adverse events to this system. And the entire dataset is open to the public and the annual data can be do wnloaded on the official website.

In our project, based on VAERS, we collect the adverse events that happened from 1 Jan 2021 to 8 October 2021 which are related to three licensed COVID19 vaccines, COVID19 (MODERNA), COVID19 (PFIZER-BIONT ECH) and COVID19 (JANSSEN). Overall, there are 563,653 records.

4 Method

4.1 Data Interpretation

Information is outlined within the United States in terms of frequency and distribution of Covid-19 vaccinat ion receipts. The quantity of adverse events was calculated by dividing the total number of vaccinations liab le with VAERS with the number of vaccinations carried out on certain days. The values within the subseque nt statistics were imputed to fill in for any missing records. Histograms and line graphs are used to point out the frequency and distribution of events.

4.2 Onset Time Prediction

In order to predict the time of the onset event after a person being vaccinated, the date time of the first even t is marked with the person being vaccinated, and the time interval as the onset data minus the marked date. We considered the interval within 28 days to do the prediction, which contains 457,762 records. For trainin g, data splitting is performed such that the dataset will be splitted with 80% training set, and 20% testing se t.

In the prediction, the input feature size is 143, which includes Age, Sex, Birth Defect, Manufacturor, Prior Vaccinated, Other Medication, Current Illness, History, and Allergies.

For Prior Vaccinated, Other Medications, Current Illness, History, and Allergies, since they are text fields, we need some extra pre-processing steps to vectorize them. The text is first pre-processed by steps like tok enization, removal of stop word and lemmatization. Unigram and 2-gram are then selected as features based on their frequency and some domain knowledge (some tokens are grouped together since they are the same medically). In the following parts, we will use short forms to indicate which category a feature comes from.

Category	Number of Features Extracted	Short Forms
Other Medication	34	(O)
Current Illness	28	(C)
History	44	(H)
Allergies	31	(A)

Table 1. Number of category data features and the respective short form

There are a number of methods on performing the prediction, which are some conventional non-neural-netw ork methods such as Linear Regression, Lasso Ridge, and Random Forest, and neural network methodology. Since the prediction is used for healthcare aspects, it is critical to understand how our models make their de cision. Therefore, besides building the models for regression, we try to explain how these machine learning methods are working by estimating the importance of each feature in these models.

4.2.1 Non-Neural-Network Method

- Linear Regression: a linear approach that separates different classes. It aims to minimize the residu al sum of squares between the labeled dataset, which is the train data, and use that to predict new d ata. The importance of features is approximated by the magnitude of the coefficients.

- Lasso: a linear approach as well, and it is optimized with a L1 regularization. Different from linear regression, it hopes to have a better fit in prediction data with the help of regularization terms. Simi lar to linear regression, the importance of features is approximated by the magnitude of the coeffici ents.

- Ridge: a linear approach as well, and it is optimized with a L2 regularization. Different from linear regression, it hopes to have a better fit in prediction data with the help of regularization terms. Simi lar to linear regression, the importance of features is approximated by the magnitude of the coeffici ents.

- Random Forest: an ensemble approach that generates a number of decision tree regressors to fit the data. 400 of estimators have been set, with the 15 maximum depth. The importance of a feature is

approximated by the Gini importance, which is the (normalized) total reduction of the MSE brough t by that feature.

4.2.2 Ordinary Neural Network And Interpretability

Neural network (NN) is a powerful method which outperforms many conventional machine learning techniq ues. It mimics the biological neural system with layers of neurons. Each artificial neuron is connected to ne urons in the next layer with specific weights, and its value is activated before passing forward. However, on e obvious limitation of NN is the interpretability. The entire model works like a black-box, and it is hard to directly understand why the network makes a specific decision.

To predict the onset date of adverse events, we have built a NN with an input layer with 143 dimensions, an d with 1 dimension output, which indicates the numerical result. There are two hidden layers with size 858 a nd 858. The Swish function is used as the activation function and the dropout is 0.5. There are a total of 86 1,433 trainable parameters in the network. L1 loss and Adam optimizer with learning rate 0.001 is used for t raining.

To measure the importance of each feature in the NN, we have tried three approaches, feature perturbation, integrated gradient and DeepLift. For each feature, the average of the importance score in all validation sam ples is used as an estimation of its overall importance.

Feature Perturbation

Feature perturbation, also known as feature ablation, is a simple perturbation-based algorithm to approximat e the attribution of each input feature. By default, each input feature in the input vector will be replaced wit h a reference/baseline value independently, and the importance score is indicated by the difference in outpu t. This method can be applied in many cases, for example, when dealing with images, a small region is occl uded to approximate its attribution. In our task, since our input is just a 1-d vector. We replace each scalar v alue in it with 0, the absolute value of the difference in output is used to estimate the importance.

Integrated Gradient

Integrated Gradient (IG) is a gradient-based algorithm for explainable AI, which was first proposed in 2017 [5]. Compared with conventional gradient-based methods which compute the attribution by directly using th e gradient with respect to input feature, the importance of an input feature is determined by the integral of g radient with respect to the value along the path from baseline to the input feature value.

$$IntegratedGradients_i(x) ::= (x_i - x'_i) \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} d\alpha$$

where:

$_i$ = feature
x = input
x' = baseline
α = interpolation constant to perturb features by

Since it is hard to directly compute the integral of gradient, the integral will be approximated by Riemann s um.

$$IntegratedGrads_i^{approx}(x) ::= (x_i - x'_i) \times \sum_{k=1}^{m} \frac{\partial F(x' + \frac{k}{m} \times (x - x'))}{\partial x_i} \times \frac{1}{m}$$

where:

$_i$ = feature (individual pixel)
x = input (image tensor)
x' = baseline (image tensor)
k = scaled feature perturbation constant
m = number of steps in the Riemann sum approximation of the integral

Since IG can be applied to any differentiable neural network and is easy to compute, it is a widely-used inte rpretability method. In our task, we set the baseline to 0 for every feature value and the steps for approximat ion of integral to 50, and the integral is used to indicate the feature importance.

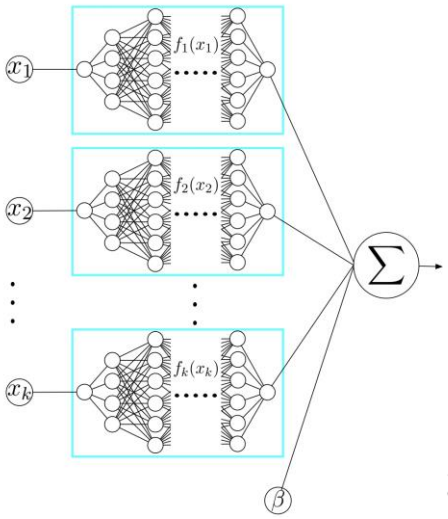
DeepLift

DeepLift [6] is an interpretability method which evaluates the contribution score of each feature by the com parison between the difference of the output from some baselines and the difference of the input feature val ues from some baselines. Since DeepLift doesn’ t exploit gradients, it can still work when the gradient is z ero. Also, it can reveal dependencies that are overlooked by other approaches such as IG. In addition, the co mputation can be carried out efficiently within a single backward pass. In our cases, we take the absolute va lue of the contribution score as the feature importance.

4.2.3 Explainable Neural Network --- Neural Additive Model (NAM)

Although the previously mentioned interpretability methods can give us some basic understanding of the im portance of each feature, the entire model is still a black-box. It is crucial to have an inherently explainable NN. Therefore, we adopt the Neural Additive Model, an inherently interpretable network [7] to build a glass -blox NN for our task.

NAM is derived from the family of Generalized Additive Models. In NAM, each input feature is first transf ormed by an independent neural network, and the output is calculated by the linear combination of the trans formed input feature value. These feature neural networks are trained jointly by backpropagation and can le arn arbitrary complex transform functions. Since the impact of an input feature on the final outcome is inde pendent of other input features , it is easy to understand the model.



$$y = \beta + f_1(x_1) + f_2(x_2) + \ldots + f_k(x_k)$$

Figure 1.The Structure Of NAM

In our case, the neural network for a single input feature has two hidden layers with 64 and 32 neurons and Leaky Relu is used as an activation function. There are a total of 320,464 trainable parameters in the networ k. L1 loss and Adam optimizer with learning rate 0.001 is used for training.

To estimate the importance of an input feature, we exploit two simple approaches. In the first approach, we use the absolute values of the transformed feature values to approximate the feature importance scores. Simi lar to previous methods, for each feature, the average of the importance score in all validation samples is us ed as an estimation of its overall importance. Another approach is that, for each feature, we consider the abs olute value of the Pearson correlation coefficients between the transformed feature value and onset time as i ts importance.

4.3 Prediction on Hospitalization Rates in event of Adverse effects

The need for hospitalization was widely understood as the pandemic broke out and hospitals were quickly o verburdened. Hence, the regard for hospitalization after an individual has been vaccinated is studied to unde rstand if medical attention should be reserved as mass vaccinations are being rolled out.

To achieve the prediction of hospitalization rates, machine learning approaches are implemented to classify the adverse effects experienced by a vaccinated patient. The symptoms leading to adverse effects recorded i n the VAERS dataset are identified as the input features and sparsely encoded according to international me dical terminologies (MedDRA).

The problem arises as there are 5 reported symptoms on average for each patient. Hence, leading to a high dimensional and increasingly sparse input feature matrix. Moreover, through preliminary analysis, it is discovered that the majority of the patients that reported adverse events ultimately do not require hospitalizations. This finding holds true for over 90% of the predicted outcomes as most patients often experience less severe symptoms that are self-manageable.

Essentially, the issue of circumventing a high dimensional and highly sparse dataset is duly noted for this research study.

4.3.1 Using SparsePCA to reduce dimensions of the dataset

We approached the input feature matrix that is highly dimensional and sparse by implementing the sparse PCA method to reduce its dimensionality of 5487 features to 5. However, we later ran into the issue whereby the label is now top-heavy. In this regard, the principal components model was set to extract only symptoms that entailed necessary hospitalizations. Following this, the low dimensional features of all examples from the training set were processed with logistic regression using balanced class weights to predict hospitalization risk.

4.3.2 Naive Bayes Sparse Feature Selection

As the need for hospitalization resembles a Bernoulli distribution model, thus, we applied a Naive Bayes classifier with sparse constraint and used Laplace smoothing to leverage the presence of zero probabilities due to the highly sparse feature matrix. The Laplace smoothing is done by adding a small constant value, K, for each possible transition and then recalculating the transition matrix and adding 1 event to the 0-transition (ignoring other probabilities).

5 Result

5.1 Data Interpretation

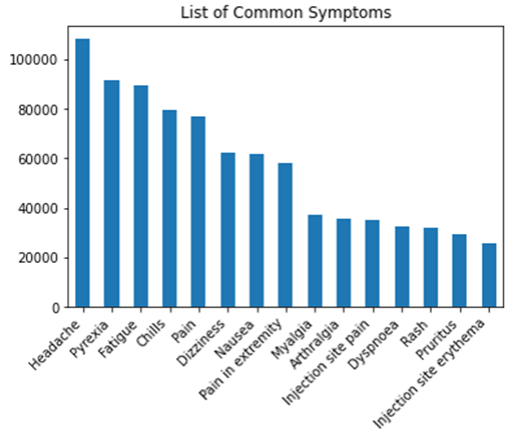


Figure 2: List of Common Symptoms

There are 15 common symptoms experienced during an adverse event, with the top 5 being the patient experiencing headache, pyrexia, fatigue, chills and pain. Pyrexia relates to experiencing a fever as there will be abnormal elevation in the patient's body temperature. Although the top 5 symptoms have been unpleasant, it is not as deadly as leading to some discomfort.

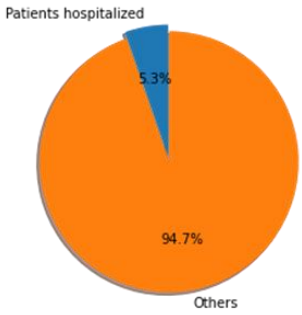


Figure 3: Distribution of patients hospitalized

As observed, a majority by 94.7% of the individuals do not require hospitalization even though they experienced adverse effects after receiving their vaccinations. As discussed from the 15 common symptoms, most of the symptoms experienced provide discomfort but are non-deadly.

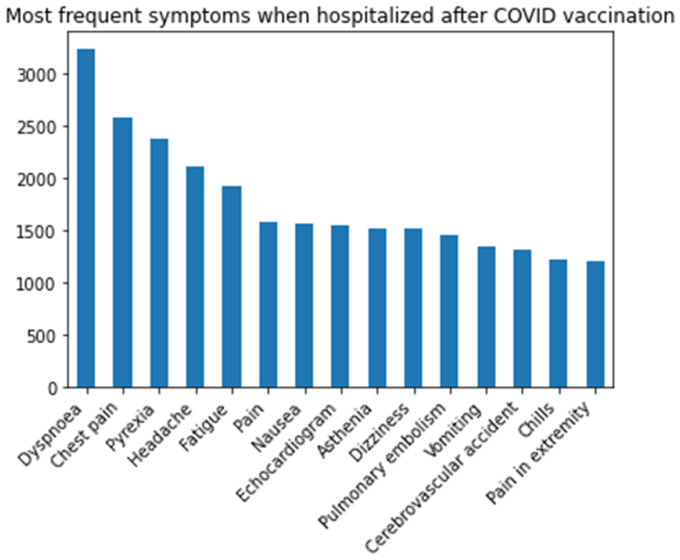


Figure 4: Most frequent symptoms when hospitalized after vaccination

Figure 3 focuses on the breakdown of common symptoms experienced by vaccinated patients after they are hospitalized. Essentially, the patients are checked into the hospital mainly for more concerning symptoms s uch as dyspnoea, chest pain and pyrexia. Most of these symptoms remain non-concerning as most of the pat ients associated with such symptoms are hospitalized for a few days only.

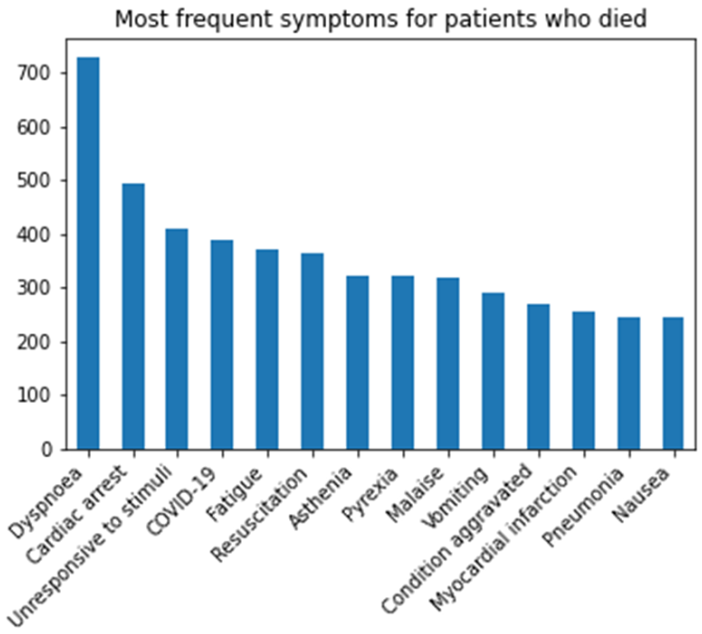


Figure 5: Most frequent symptoms for patients who died

By analyzing the common symptoms experienced by patients who died, the symptoms are severe, such as c ardiac arrest, unresponsiveness to stimuli or even being infected with Covid-19 after vaccination. Dyspnoea is often mistaken for snoring during sleep periods and hence, may often be overlooked as it can be deemed a less alarming symptom. It is noted that a significant portion of deaths are contributed by infecting Covid-1

9 after vaccination, however, it should also be put into contrast that these rates are studied based on 1.2% pa tients who died after vaccination. To put into perspective, the magnitude of contracting Covid-19 upon vacc ination is much smaller.

5.2 Onset Time Prediction

This section will show the performance difference in predicting the onset time.

5.2.1 Non-Neural-Network Method Result

The top 10 important features for predicting the onset time are Age, Birth_Defect, Prior Vaccinated, Covid (H), Covid(C), Birth Control(O), Prenatal(O), Cancer(C), Depression(C), and Hydrocodone(A). The differe nce between Lasso and Ridge with Linear Regression are Vitamin, and Obesity. By applying random fores t, the mean absolute error can be reduced by 0.1. Compared with linear methods, random forest considers Manufacturer Biontech, Manufacturer Moderna, Hypertension(H), Seasonal Allergies(C), Levothyroxine(O), and Asthma(H) more important for the prediction.

Method	Mean Absolute Error	Top 10 Important Feature
Linear Regression	3.446	Age, Birth_Defect, Prior Vaccinated, Covid(H), Covid(C), Birth Control(O), Prenatal(O), Cancer(C), Depression(C), Hydrocodone(A)
Lasso	3.448	Age, Birth_Defect, Prior Vaccinated, Covid(H), Birth Control(O), Covid(C), Prenatal(O), Vitamin(O), Obesity(H), Hypertension(C)
Ridge	3.446	Age, Birth_Defect, Prior Vaccinated, Covid(H), Covid(C), Birth Control(O), Prenatal(O),

		Cancer(C), Depression(C), Hydrocodone(A)
Random Forest	3.336	Age, Vitamin(O), Prior Vaccinated, Sex Male, Manufacturer Biontech, Manufacturer Moderna, Hypertension(H), Seasonal Allergies(C), Levothyroxine(O), Asthma(H)

Table 2: Non neural network's mean absolute error and the result

5.2.2 Ordinary Neural Network Result

The MAE of the neural network is 2.68 which is better than all of the non-NN methods mentioned in the pr evious parts. From Table X, we can see the important features figured out by the three approaches are quite similar. They all includes Age, Vitamin(O), Sex Male, Penicillin(A), Prior Vaccinated and Hypertension(H). Actually, some of these common features are also important features in non-NN methods.

Method	Feature Perturbation	Integrated Gradients	DeepLift
Top 10 Impo rtant Feature	Age, Vitamin(O), Sex Male, Manufacturer Biontech, Penicillin(A), Asthma(H), Prior Vaccinated, Sulfa Drugs(A), Hypertension(H), Birth Control(O)	Age, Vitamin(O), Levothyroxine(O) Penicillin(A), Sulfa Drugs(A), Hypertension(H), Prior Vaccinated, Sex Male, Birth Control(O), Lisipril(O)	Age, Sex Male, Manufacturer Biontech, Prior V accinated, Asthma(H), Manufacturer Moderna, Vitamin(O), Penicillin(A), Hypertension(H), Anxiety(H)

Table3: Neural network's mean absolute error and the result

5.2.3 Neural Additive Model

The MAE of the NAM is 2.772 which is slightly higher than the NN but still better than those non-NN met hods. One possible reason is that, in NAM, owing to the network structure, features cannot interact with ea ch other.

It seems that NAM considers different features from other methods but we still can find some common imp ortant features. From Table X, when we consider the absolute value of the transformed feature value as the f eature importance, we can see Birth Control(O), Covid(C) and Hydrocodone(A), which also exist in previou

s parts. When we consider the absolute value of the Pearson correlation coefficients between the transforme d feature value and onset time, we can see Vitamin(O), Hypertension(H), Prior Vaccinated, Levothyroxine (O), Lisipril(O), which also exist in previous parts.

Approach	Top 10 Important Feature
Absolute value of the transformed feature value	Bee(A), Birth Control(O), Irritable Bowel Syndrome(H), GERD(C), Omeprazole(O), Codeine(A), Aspirin(A), Covid(C), Ceclor(A), Hydrocodone(A)
Absolute value of the Pearson correlation coefficients between the transformed feature value and onset time	Vitamin(O), Hypertension(H), Prior Vaccinated, Atorvastatin(O), Aspirin(O), Calcium(O), Metoprolol(O), Levothyroxine(O), Hyperlipidemia(H), Lisipril(O)

Table 4: Neural additive model's result

5.3 Hospitalization Prediction

Based on the preliminary analysis discussed in 4.1 Data Interpretation, the top 15 symptoms that contributed to the hospitalization rates for vaccinated patients were extracted through **logistic regression with sparse PCA** and **sparse Naive Bayes**.

There existed 11 overlapping symptoms based on the two methods, which are: Pain in extremity, Nausea, Pyrexia, Chills, Fatigue, Pain, Arthralgia, Dizziness, Vomiting, Myalgia and Asthenia. The listed overlapping symptoms are more severe and can lead to the need for hospitalization.

In comparison of the two methods, the sparse Naive Bayes method was able to summarize the common symptoms but failed to determine symptoms that lead to the hospitalization of vaccinated patients. On the other hand, the logistic regression with sparse PCA method fared better by obtaining the principal components for hospitalization needs. It was also efficient in describing more severe symptoms that lead to hospitalizations. Predictions on the need for hospitalization is summarized in the following table:

	Logistic Regression with SparsePCA	Sparse Naive Bayes
Optimal probability threshold	0.46	0.03
AUC	0.7625	0.5327
Training set sensitivity	0.5595	0.1581
Training set specificity	0.8699	0.9753
Validation set sensitivity	0.5570	0.4638
Validation set specificity	0.8651	0.9289

Table 5: Summary of performance for Logistic Regression with SparsePCA and Sparse Naive Bayes models

Essentially, the logistic regression with sparse PCA model was much suited in predicting hospitalization rates for the vaccinated patients.

6 Conclusion

In this research study, we mainly discussed the implementation of various machine learning models to tackle a sparse dataset that is tediously detailed such as VAERS.

The two methods were primarily adopted to predict the onset of adverse events: conventional non-neural network methods and neural network-based methods. This is followed by Gradient Integral and DeepLift to approximate the importance of each feature leading to the outcome. To better predict the onset time, we attempted to build a Glass-Box neural network alongside the Neural Additive model.

To summarize the performances of these models, the classic neural network model performed the best in predicting features that contribute to onset of adverse events. This is followed by the performance of the Neural Addictive model we built. Both models outperformed the non-neural network models.

As for predicting the hospitalization rates, the logistic regression with sparse PCA method was able to tackle the sparse input feature and significantly outperformed the sparse Naive Bayes model. Through this model, we were able to successfully compress 5,487 features into just 5 and extract common symptoms that lead to hospitalization as well as specific in describing the severe symptoms.

The methods discussed can be employed by future researchers towards the VAERS dataset.

7 Discussion

For future reference, more resourceful findings can be sought out by combining the feature findings from th e VAERS dataset alongside various healthcare datasets to improve the healthcare system.

In this case, if we have the spatial temporal data of the adverse events, we could certainly predict where and when the adverse events are happening with probability. Location and time of hospitalization due to adverse events could be predicted so that capacity planning and scheduling in hospitals could be done in advance.

8 Reference

[1] Gonzalez-Dias, P., Lee, E. K., Sorgi, S., de Lima, D. S., Urbanski, A. H., Silveira, E. L., & Nakaya, H. I. (2020). M ethods for predicting vaccine immunogenicity and reactogenicity. *Human vaccines & immunotherapeutics*, *16*(2), 269 – 276. https://doi.org/10.1080/21645515.2019.1697110

[2] Ahamad, M. M., Aktar, S., Uddin, M. J., Rashed-Al-Mahfuz, M., Azad, A. K. M., Uddin, S., Alyami, S. A., Sarke r, I. H., Liò, P., Quinn, J. M. W., & Moni, M. A. (2021). Adverse effects of covid-19 vaccination: Machine learning an d statistical approach to identify and classify incidences of morbidity and post-vaccination reactogenicity. https://doi.or g/10.1101/2021.04.16.21255618

[3] Jenatton, R., Obozinski, G. & Bach, F.. (2010). Structured Sparse Principal Component Analysis. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, in *Proceedings of Machine Learning Research* 9:366-373 Available from https://proceedings.mlr.press/v9/jenatton10a.html.

[4] Askari, A., d' Aspremont, A. & Ghaoui, L.E.. (2020). Naive Feature Selection: Sparsity in Naive Bayes. *Pro ceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, in *Proceedings of Mach ine Learning Research* 108:1813-1822 Available from https://proceedings.mlr.press/v108/askari20a.html.

[5] Sundararajan, M., Taly, A., & Yan, Q. (n.d.). *Axiomatic attribution for Deep Networks - arXiv*. Retrieved Novemb er 28, 2021, from https://arxiv.org/pdf/1703.01365.

[6] Shrikumar, A., Greenside, P., & Kundaje, A. (2019, October 12). *Learning Important Features Through Propagatin g Activation Differences*. Retrieved November 28, 2021, from https://arxiv.org/pdf/1704.02685.pdf.

[7] Agarwal, R., Melnick, L., Frosst, N., Zhang, X., Lengerich, B., Caruana, R., & Hinton, G. E. (2021, October 24). *Neural Additive Models: Interpretable Machine Learning with Neural Nets*. Retrieved November 28, 2021, from https: //arxiv.org/pdf/2004.13912.pdf.

[8] F.P. Polack, S.J. Thomas, N. Kitchin, J. Absalon, A. Gurtman, S. Lockhart, et al. (2020, December 12). *Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine.* Retrieved November 28, 2021, from https://www.nejm.org/doi/1 0.1056/NEJMoa2034577.

[9] P.T. Heath, E.P. Galiza, D.N. Baxter, M. Boffito, D. Browne, F. Burns, D.R. Chadwick, R. Clark, C. Cosgrove, J. Galloway (2021). *Safety and Efficacy of NVX-CoV2373 Covid-19 Vaccine.* Retrieved November 28, 2021, from https://www.nejm.org/doi/full/10.1056/NEJMoa2107659.