# 1. Problem Statement / Motivation

N-linked Glycosylation is the attachment of Oligosaccharide (can be referred also as glycan) to an asparagine molecule denoted by N in the Protein Sequence. N-linked glycosylation in virus is an effective phenomenon that requires phenomenal focus due to emerging viruses such as Ebola, Corona etc for which the cure is not yet certain. Some of the motivations for achieving our goal are given below:

- The viruses on undergoing n-linked glycosylation can do functions such as entry into host cells, proteolytic processing and protein trafficking virus. Predicting these sites can help us in incubating these functions.
- N-linked Glycosylation is done conventionally through mass spectrometry and other experimental procedures which are costly, but machine learning models take lesser cost and requirement
- There is a need to do glycosylation specifically for viruses

# 2. Background Literature

There are two important papers that this project was based up on which I took into reference for designing this project. I am going to summarize papers intents and the information provided by them.

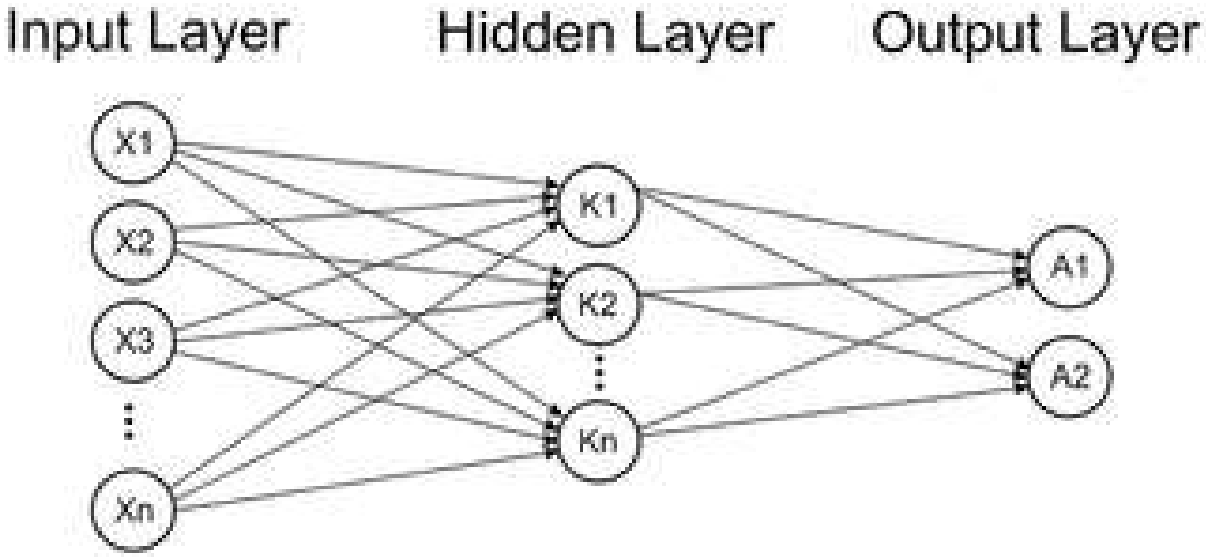
1) *Vigerust DJ, Shepherd VL. Virus glycosylation: role in virulence and immune interactions. Trends Microbiol. 2007 May;15(5):211-8. doi: 10.1016/j.tim.2007.03.003. Epub 2007 Mar 29. PMID: 17398101; PMCID: PMC7127133.*

This paper explains about the role of glycosylation in increasing virulence and viral immunity using medical in-depth terminology. Below are mentioned some important viruses and the abilities that get increased

| | |
|---|---|
| HIV1 | survival and immune evasion |
| West Nile Virus | replication and maturation |
| Influenza | receptor binding, infectivity, virus release and neurovirulence |
| Ebola | infectivity |

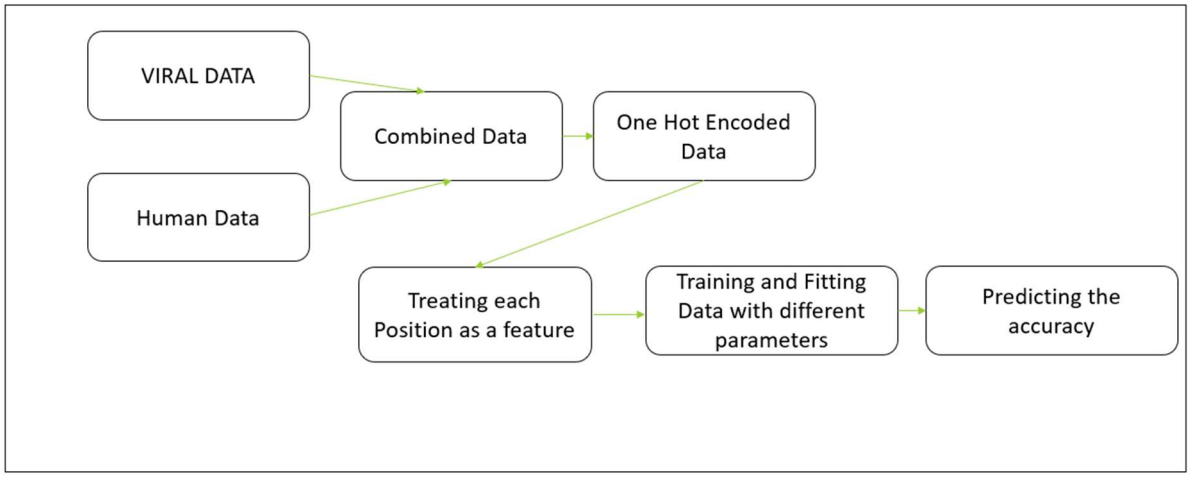
2) *Prediction of N-linked glycosylation sites using position relative features and statistical moments (Muhammad Aizaz Akmal, Nouman Rasool, Yaser Daanial Khan)*

This paper uses the input a site vicinity vector and statistical moments. The site vicinity vector is a vector of protein sequences surrounding the molecule that we are predicting if it is going to be glycosylated or not. The Statistical moments are a quantitative measure used for describing a collection of data. Mathematicians and statisticians have formed various moments based on certain well-known polynomials and distribution functions. The moments used in order to elucidate the proposed problem are raw, central and Hahn moments. They use this data and do data cleaning, feature extraction, and other data pre processing techniques. After this data is fed into an Artificial Neural Network.

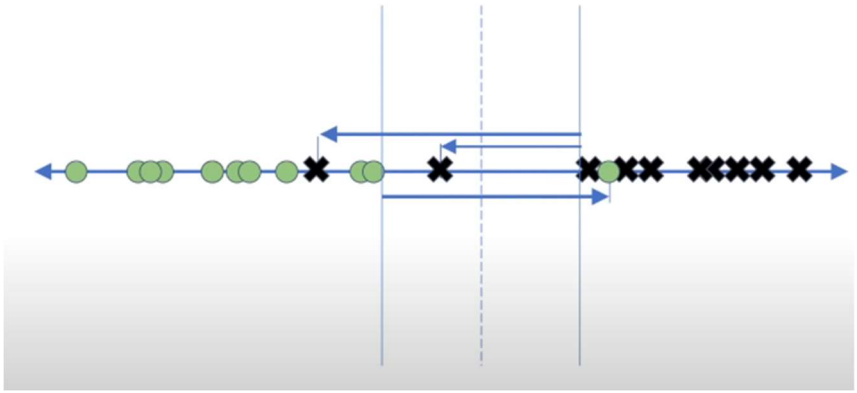



The artificial neural network has an Input Layer with n number of neurons where n is the number of inputs in the vector. On training input values are passed from each of the input layer to hidden layer and a weight will be assigned to each of the inputs. This weight will be multiplied with the input and added to bias function in the hidden layer neurons. If the entire sum is greater than a particular threshold then the neuron gets activated and passes its value to next neuron. In the hidden layer there may be more than one layer. It will pass through all these layers and finally reach the output layer. This is called forward propagation. Since even the output training data will be passed along with the input. If a misclassification happens the weights are revised in the opposite direction. This process is known as backward propagation. On training continuously, we get weights which can predict the values for glycosylation.

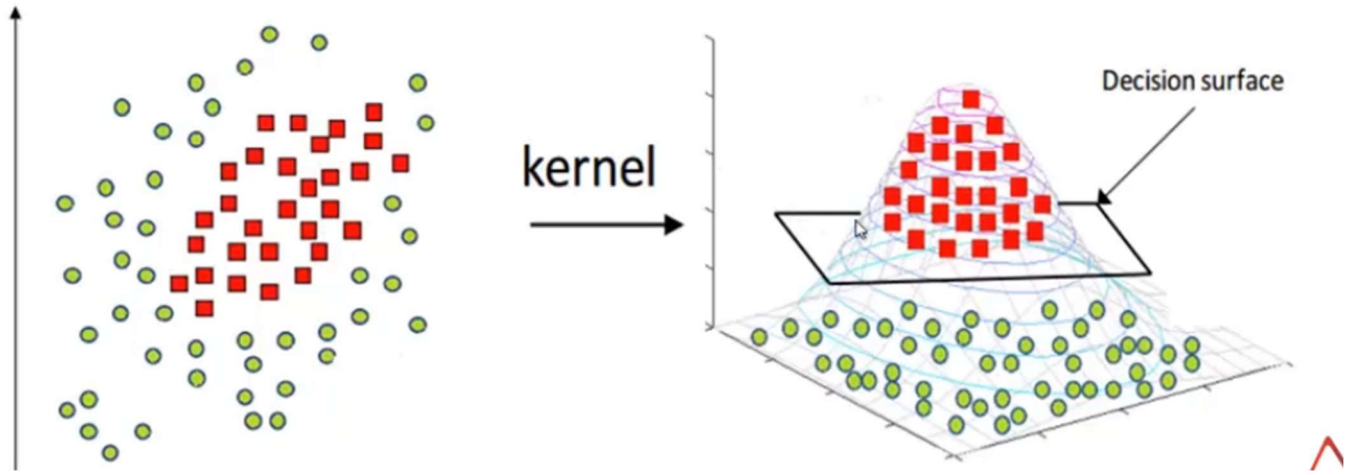
# 3. Methodology




The problem with viral data was there was no presence of data that can be used for N-linked Glycosylation, so we use transfer learning to obtain data from species humans for which data is abundant. On getting this data then we proceed to encode the data using one hot encoding. One hot encoding is a mechanism of representing alphanumerical characters to number. We use it for our protein sequences. We used this data and feed into a machine learning model. The machine learning model that I choose to implement is Support vector Machine.



A simple SVM model aims to separate data belonging to different class by fitting a line with the least penalty possible and the widest margin. In the above diagram we can see that the data is linearly separable. So, it draws a line such that it gets the widest margin and there is a least number of misclassifications as possible. This type of separation which can be done by using linear kernel.




The image represented above shows the gaussian kernel. Gaussian Kernel is used when the data is not linearly separable. Given a 3d space it tries to cut using a 2d plane. The dimensional slicing depends often on the number of features you feed to the model. For our data I took 5 proteins on either side of the Asparagine protein as an input and fed into one hot encoding function which returns a vector of size 253. So, each of the bits in these 253 vectors are given as a feature to the support vector machine. So, our model is cut by 252-dimensional slice to separate all the data. This is the effective process implemented. The data and code related to this implementation using pythons inbuilt sk learn is given below.

https://github.com/BhargavVasudevaVunnam/ViralN-LinkedGlycosylation

# 4.Results

10 Fold Validation Test :

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Training | Training | Training | Training | Training | Training | Training | Training | Training | Test |
| Training | Training | Training | Training | Training | Training | Training | Training | Test | Training |
| . | | | | | | | | | |
| . | | | | | | | | | |
| Test | Training | Training | Training | Training | Training | Training | Training | Training | Training |

In this type of test the data is divided into 10 sub parts. The test of blocks size 1 is changed each time to from $10^{th}$ block to $1^{st}$ block. Each time before changing the test from 1 block to

other we train and test and get the score of each of the test. On averaging overall scores the resultant score is the 10 fold test score.

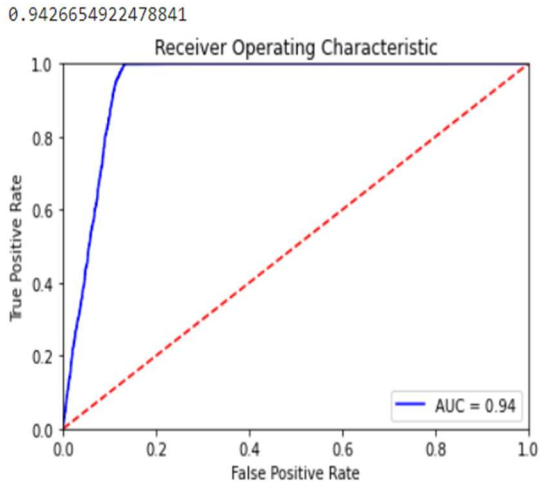
Achieved Score = 93 %

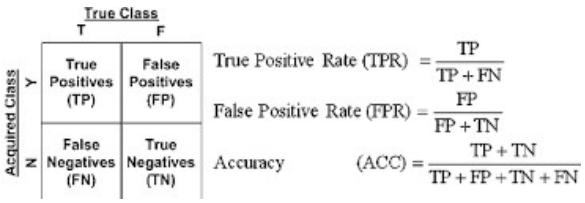
Confusion Matrix:

| | **Positive** | **Negative** |
|---|---|---|
| Positive | 3457 | 570 |
| Negative | 3 | 4302 |

The confusion matrix for our model is given above. A confusion matrix explains in each row the number of true positives, true negatives, false positives, false negatives

Roc Curve:

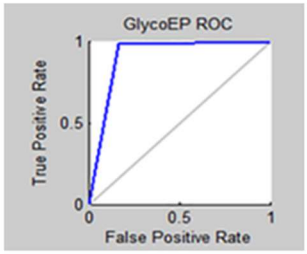


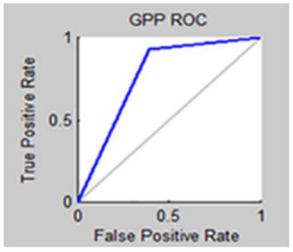



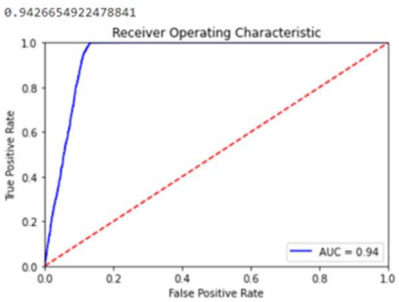

The ROC curve is plotted using the values given in the confusion matrix. If true positive rate is 1.0 for all values than that is the best value possible. The area under this curve indicates the area under the curve value.

Achieved Score: 94%

# 5. Conclusion and Discussion




Using transfer learning we can reach an accuracy of about 93 Percent and we get the area under curve value to be 94.2 percent. The given target is reached which is fairly accurate and can be compared to some values for servers like GlycoEP and GPP found on the internet.

Given the results we can further try to increase the accuracy close to 98. By web scraping we can get more data relating to virus as the amount of data present regarding humans and other species is more than that of virus.

# 6. References

1) Vigerust DJ, Shepherd VL. Virus glycosylation: role in virulence and immune interactions. Trends Microbiol. 2007 May;15(5):211-8. doi: 10.1016/j.tim.2007.03.003. Epub 2007 Mar 29. PMID: 17398101; PMCID: PMC7127133.

2) Prediction of N-linked glycosylation sites using position relative features and statistical moments (Muhammad Aizaz Akmal, Nouman Rasool, Yaser Daanial Khan)

3) Sun, S., Hu, Y., Ao, M. et al. N-GlycositeAtlas: a database resource for mass spectrometry-based human N-linked glycoprotein and glycosylation site mapping. Clin Proteom 16, 35 (2019). https://doi.org/10.1186/s12014-019-9254-0

4) Scikit-learn: Machine Learning in Python

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay; 12(85):2825−2830, 2011.

5) Olsen JV, Mann M. Status of large-scale analysis of post-translational modifications by mass spectrometry. Mol Cell Proteomics. 2013. https://doi.org/10.1074/mcp.O113.034181.

6) Wilhelm M, Schlegl J, Hahne H, Gholami AM, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H. Mass-spectrometry-based draft of the human proteome. Nature. 2014;509(7502):582–7.

7) Shi X, Brauburger K, Elliott RM. Role of N-linked glycans on Bunyamwera virus glycoproteins in intracellular trafficking, protein folding, and virus infectivity. Journal of virology. 2005 Nov 1;79(21):13725–34. pmid:16227292

8) Steen PV, Rudd PM, Dwek RA, Opdenakker G. Concepts and principles of O-linked glycosylation. Critical reviews in biochemistry and molecular biology. 1998 Jan 1;33(3):151–208. pmid:9673446

9) Aebi M. N-linked protein glycosylation in the ER. Biochimica et Biophysica Acta (BBA)-Molecular Cell Research. 2013 Nov 30;1833(11):2430–7.