Applying Machine Learning to Study Influenza Virus Behavior

Rohan Koodli, Lynbrook High School, San Jose, CA 95129
rovik05@gmail.com

1. Introduction

Influenza is a virus that mutates and changes yearly. It infects millions of people worldwide, causing 500,000 deaths yearly. Millions of dollars are spent each year finding a vaccine to possibly prevent the flu. In recent years, people have made attempts at predicting how the influenza virus has changed. For example, every year, world health officials try to predict what drugs to include in the coming year's vaccine. However, many of the methods used have been called "questionable" [8] [9]. Many other methods have been used, such as creating formulas to predict the flu, which have had a little more success. However, no method has utilized phylogenetics and machine learning. Machine learning is the ability for the computer itself to locate trends within data and then predict new data based on those trends. Phylogenetics is the study of evolutionary history and the relationships between individuals on a phylogenetic tree.

A key insight is to train a machine learning model on parent-child sequences present in a phylogenetic tree. Our algorithm studies how the influenza *hemagglutinin* (HA) and *neuraminidase* (NA) glycoproteins have changed in previous years, and builds decision trees based on the patterns it recognizes. When given a new HA or NA sequence, it predicts the offspring's HA and NA sequences using the decision trees the algorithm has constructed. This is performed for a set of three machine learning algorithms, namely, Decision Trees, Random Forests and Extremely Randomized (Extra) Trees.

Performance analysis indicates that the Random Forests is able to capture the underlying shifts in flu mutations with an accuracy of 93% for H1N1 and 84% for H3N2 flu subtypes. This is better than the existing literature work based on formula-based prediction which suffer from their inability to adapt as the mutations vary over a period of time.

2. Goals

The goal of this project is to predict future sequences of the *hemagglutinin* (HA) and *neuraminidase* (NA) glycoproteins of the influenza virus A/H1N1 and A/H3N2 subtypes using machine learning models.

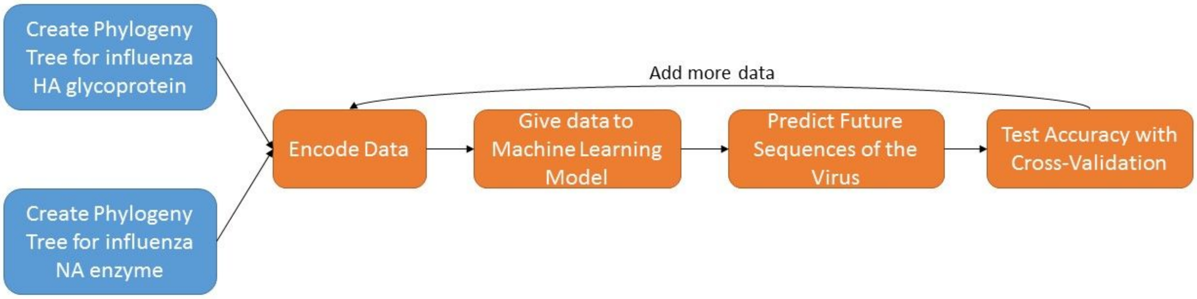
Create Phylogeny Tree for influenza HA glycoprotein
Create Phylogeny Tree for influenza NA enzyme
Add more data
Encode Data
Give data to Machine Learning Model
Predict Future Sequences of the Virus
Test Accuracy with Cross-Validation

Figure 1: An Overview of Flu Strain Prediction Algorithm

3. Model Design

We trained our model using H1N1 and H3N2 sequences from the Influenza Research Database (IRD). For any supervised learning training set, there is a known input and known output. In our case the input is the parent flu strain; the known output is the child flu strain. Our model is then trained to see patterns between the parent and child strains. Then, we compute the accuracy using a technique called cross-validation. Cross-validation is an effective method of evaluating a machine learning model's performance. Cross-validation involves splitting the model's data into two parts: a training set and a predicting set. The model trains on the first part, then predicts the second part. The model then compares what it predicted with the actual outcome of the second part, and compares what the model predicted accurately. See Figure 2.

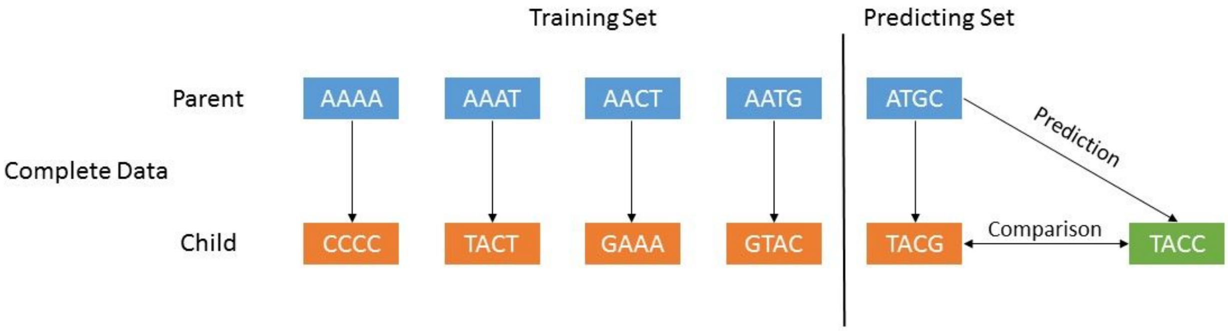

Figure 2: Machine Learning Algorithm Design

4. Results

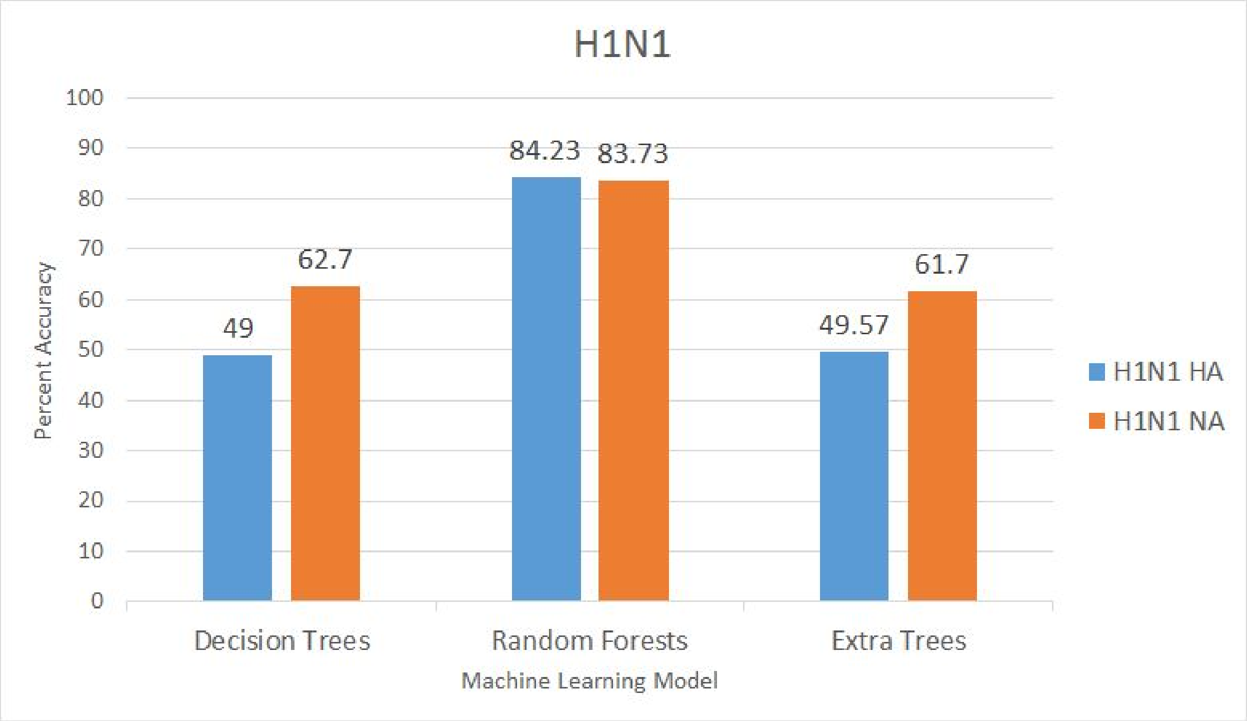

Figure 3: Accuracy of Machine Learning Models on H1N1 HA and NA proteins

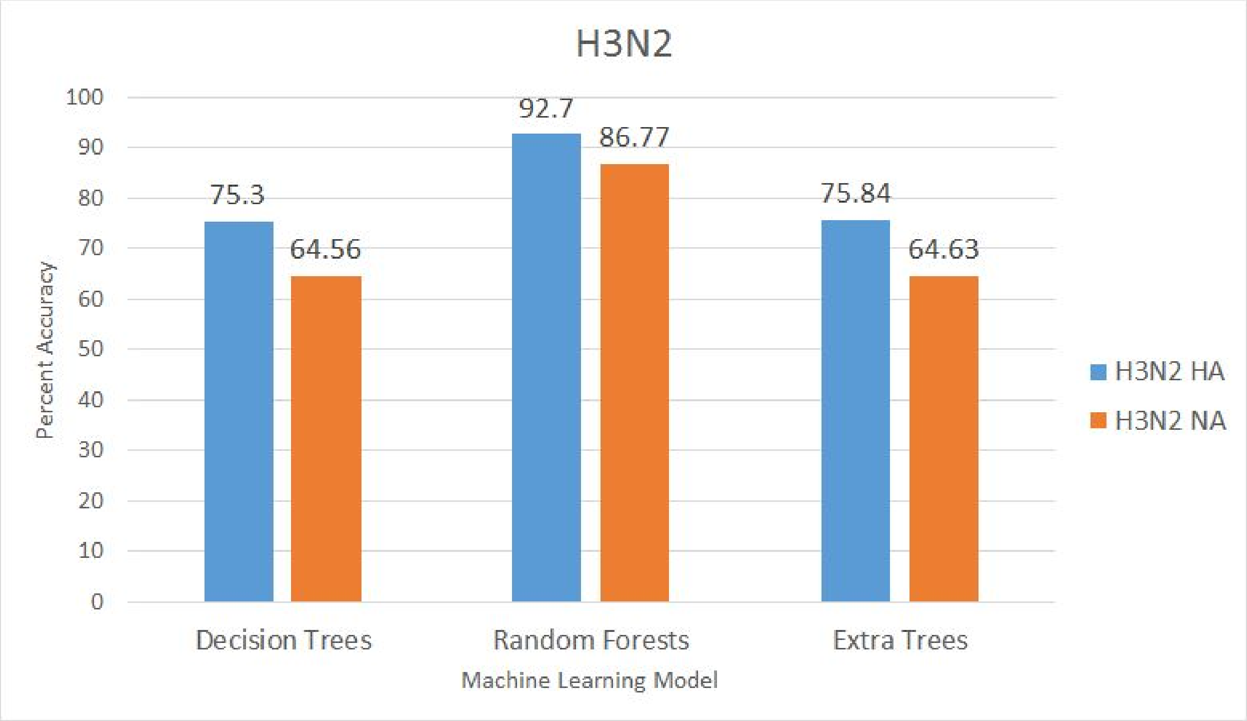

Figure 4: Accuracy of Machine Learning Models on H3N2 HA and NA protein

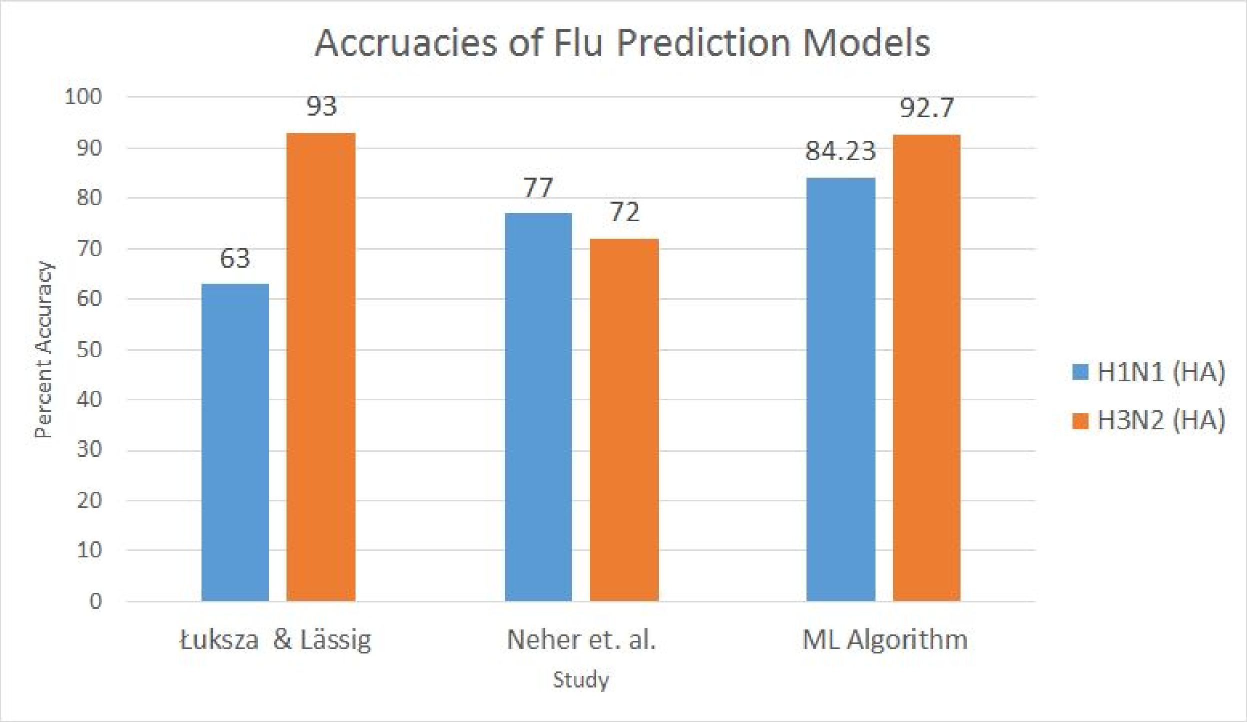

Figure 5: Accuracies of Different Flu Prediction Models

5. Analysis

We tested three algorithms on the HA and NA glycoproteins of two flu subtypes (H1N1 and H3N2) to investigate which had the greatest accuracy.

The Random Forests algorithm had the best average overall accuracy (84% for H1N1, 90% for H3N2). The Extra Trees and Decision Tree algorithms performance was nearly identical (56% for H1N1, 70% for H3N2 for both algorithms).

There are two types of flu mutations: *antigenic shift* and *antigenic drift*. Antigenic Shifts are drastic changes in the flu that can happen suddenly. These mutations occur occasionally, and can cause pandemics like the 2009 H1N1 outbreak [6]. Antigenic Drifts are gradual changes that happen slowly, but can accumulate over time. As a result, in some years, drift mutations could be small, and in other years, can be much more profound [6]. For any machine learning algorithm, it is important that it does not assign overweight values to the large antigenic drift changes, as these decisions could add unnecessary decision nodes to the decision tree and lower the accuracy of the predictions.

The Decision Tree algorithm was able to locate small antigenic drift changes in flu mutations, but the greater antigenic drift changes caused it to over-complicate its tree. *Decision Tree* found that asking certain questions which led to large antigenic drift mutations had a low entropy, so it added unnecessary nodes to its decision tree, thus lowering the overall accuracy. *Random Forests*, on the other hand, randomly modified conditions on its decision nodes, which allowed it to locate random antigenic drift mutations. Instead of selecting nodes which had the least entropy, Random Forests generated many different trees, each different from one another due to the randomization. Because it created many such trees without trying to optimize for minimum entropy, Random Forests was able to reduce the importance of the large drift changes by averaging the randomized trees. This increased the accuracy of the model. Extra Trees attempted even more randomization of nodes on its trees than Random Forests, allowing it to locate large antigenic drift mutations. However, this increased randomness caused it to give too little importance to the small antigenic drift mutations, making it often predict those mutations incorrectly. So, the "sweet spot" for prediction lies between optimizing for minimum entropy and complete randomization. An algorithm which captures a good balance between these two comes closer to obtaining the best accuracy, which is illustrated in Figure 8 below.

We compared these results with other previous attempts to study how the flu changes from year to year. In a study by *Neher, Richard A. et al [3],* scientists developed a model involving studying different concentrations of antibodies and how viruses would react to them. Their model predicts on two flu subtypes: H1N1and H3N2. For H1N1, the model has an accuracy of 77%, and for H3N2, the model has an accuracy of 72%. Also, a study by *Marta Łuksza and Michael Lässig* at the University of Cologne in Germany [2] predicts the frequency of appearances of new strains in future years, based on how clades (large branches of a phylogenetic tree) of the flu have evolved from a common ancestor. The accuracies were 63%, and 93%, respectively. Our algorithm involving machine learning has 84% accuracy when predicting H1N1 flu strains, and 93% accuracy when predicting H3N2 strains. Our model performs better when predicting H1N1, and predicts as well as Łuksza and Lässig's model with H3N2.

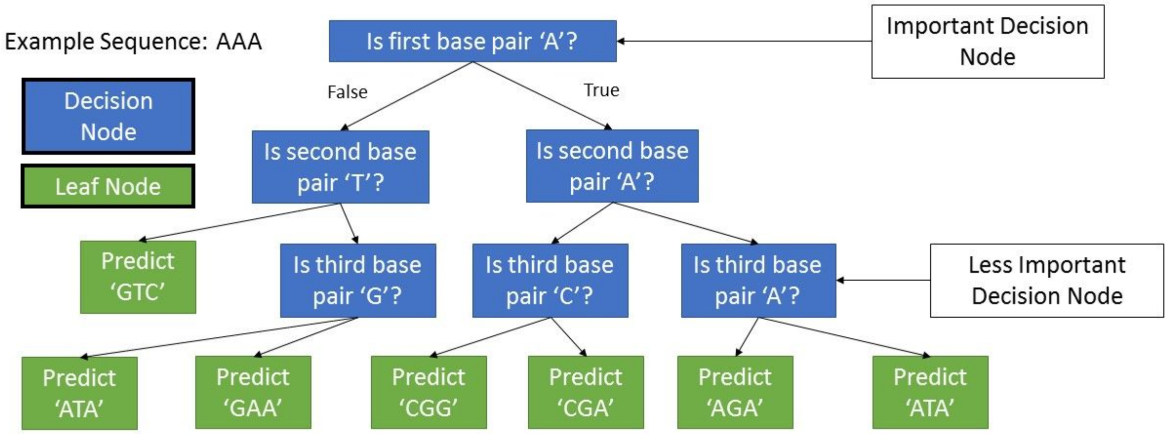

Figure 6: A Decision Tree

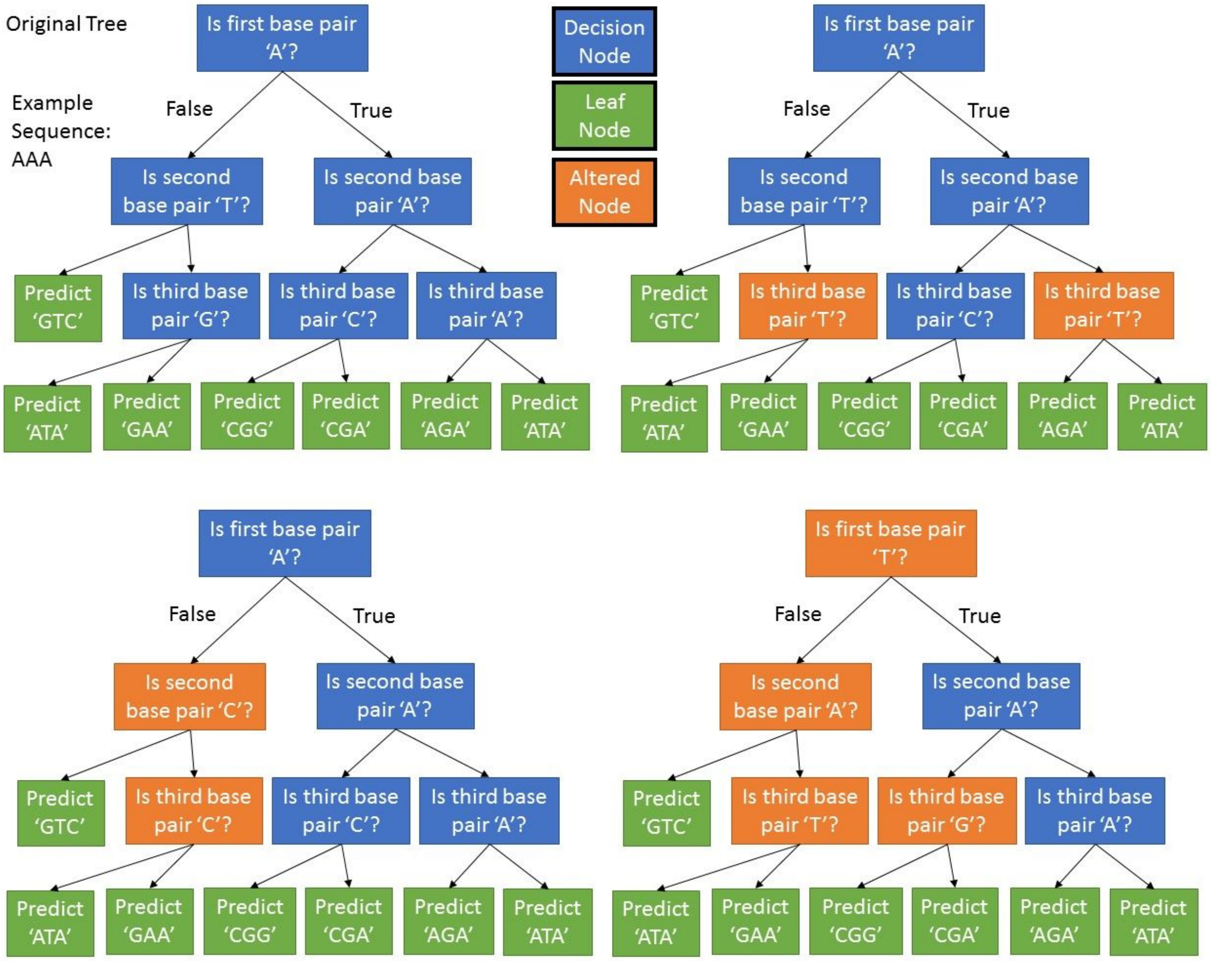

Figure 7: A Random Forest

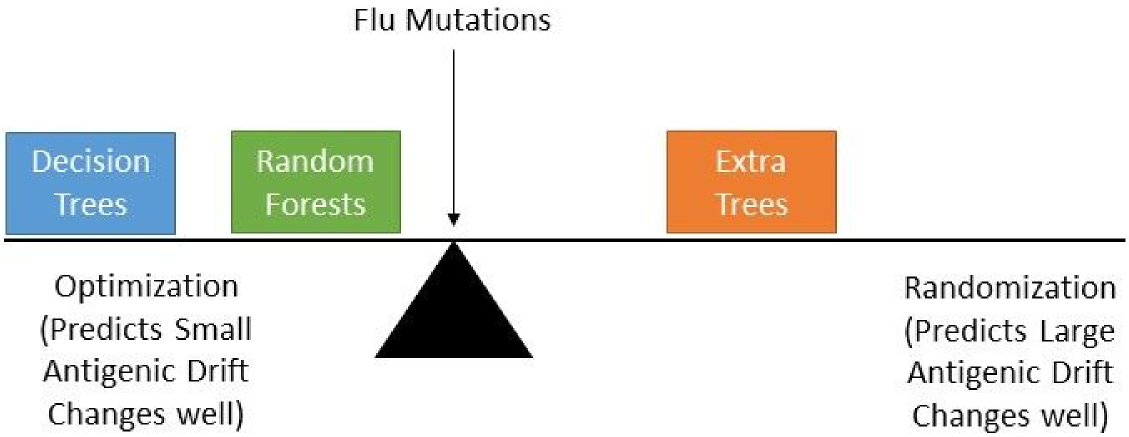
Flu Mutations
Decision Trees
Random Forests
Extra Trees
Optimization (Predicts Small Antigenic Drift Changes well)
Randomization (Predicts Large Antigenic Drift Changes well)

Figure 8: Factors which have Effects on Flu Strain Prediction

6. Conclusion

In the end, our engineering goals were met. We were able to come up with an algorithm which could effectively predict a next flu season's flu strains. We were able to encode our data, turning it from a string of base pairs into a list of floating-point numbers. We then trained with three different machine learning models: Decision Tree, Random Forests, and Extra Trees. The Random Forests provided the best accuracy, and outperformed two similar studies when predicting future strains of both H1N1 and H3N2 flu subtypes. This shows that machine learning is more effective at predicting future sequences of the influenza compared the formula-based predictions referenced earlier. With the improved accuracy, our model can be implemented in the real world, and can help save countless lives each year.

7. Future Work

Our next steps would be to create a "super algorithm", which would combine the strengths of Decision Tree, Random Forests, and Extra Trees, and minimize the weaknesses. The algorithm could learn where to create nodes with least entropy and when to create nodes randomly. In addition to predicting the flu strain, the algorithm could predict phenotypic data of that specific strain (such as drug resistances and age of person infected). Doing so would further increase the accuracy of our model, and would better predict flu mutations in the coming years.

8. Acknowledgements

We would like to thank Arjun Athreya for his guidance and his help in the early steps of this project. We would also like to thank our mentor, Mr. Iams, for his advice, and our parents, for their support and guidance.

9. Bibliography

1. Stamatakis, Alexandros. *Phylogenetics: Applications, Software, and Challenges. Cancer Genomics and Proteomics.* Foundation for Research and Technology-Hellas, Institute of Computer Science, Crete, Greece. 31 Aug 2005. Print. 02 Jan 2016.
2. Łuksza, Marta, and Michael Lässig. "A Predictive Fitness Model for Influenza." *Nature.com*. Nature, 06 Mar. 2014. Web. 24 Feb. 2016
3. Neher, Richard A., Trevor Bedford, Rodney S. Daniels, Colin A. Russell, and Boris A. Shraiman. "Prediction, Dynamics, and Visualization of Antigenic Phenotypes of Seasonal Influenza Viruses." (n.d.): n. pag. Max Planck Institute for Developmental Biology, 26 Oct. 2015. Web. 15 Feb. 2016.
4. Grus, Joel. *Data Science from Scratch*. 1st ed. Sebastopol: O'Reilly, 2015. Print.
5. *Influenza Research Database*. 15 Jan 2016. <www.fludb.org>.
6. "How the Flu Virus Can Change: “Drift” and “Shift”." *Centers for Disease Control and Prevention*. Centers for Disease Control and Prevention, 19 Aug. 2014. Web. 10 Mar. 2016.
7. Lemey, Philippe, Marco Salemi, and Anne-Mieke Vandamme. *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing*. 2nd ed. Cambridge: Cambridge UP, 2009. Print. 22 Dec 2015.
8. Saxena, Shailendra, Rosaiah Kotikalapudi, Sneham Tiwari, and Charuvaka Muvva. "Influenza A(H1N1)pdm09 Virus." MedScape. MedScape, 10 July 2012. Web. 16 Feb. 2016.
9. Ramos, AP, et al. "Molecular and Phylogenetic Analysis of Influenza A H1N1 Pandemic Viruses in Cuba, May 2009 to August 2010." NCBI. U.S. National Library of Medicine, 17 July 2013. Web. 16 Feb. 2016.
10. Mohan, Geoffrey. "Can a Genetic Model Predict next Year's Flu Strain?" *Los Angeles Times*. Los Angeles Times, 26 Feb. 2014. Web. 24 Feb. 2016.
11. Herzum, I., T. Lutz, F. Koch, R. Geisel, and A. Gehrt. "Diagnostic Performance of Rapid Influenza Antigen Assays in Patients Infected with the New Influenza A (H1N1) Virus." *National Center for Biotechnology Information*. U.S. National Library of Medicine, 2010. Web. 24 Feb. 2016.