Fending off the flu: Using machine learning to diagnose and manage outbreaks

Project Objective

The purpose of this research is two-fold:

1) To develop a machine learning program that can accurately and rapidly diagnose flu in adults.
2) To optimize the program so that it can be used as a surveillance system for outbreak management.

To meet the objectives, this study aims to engineer a machine-learning based diagnostic program based on previous influenza symptoms and cases. The accuracy of this diagnostic will also be studied, and based on this information, the program will be optimized.

Project Background and Significance

The flu, also known as influenza, is an acute respiratory illness induced by different strains of the Influenza viruses (World Health Organization, 2025). It is an enveloped, RNA-virus that infects human epithelial cells. There are four types of influenza viruses A, B, C, and D, each with different characteristics in terms of infectivity, severity, and structures (Kayser et al., 2021). The two well-known influenza types are A and B, and both have the capability to instigate an epidemic. Type A also has the potential to cause a global pandemic as evidenced by the 1918 Spanish flu, the deadliest global pandemic in history. While there are treatment options (e.g., antiviral) for Flu, there is no cure. (Geisinger, 2024). Such characteristics accentuates the need for proper preparation and action plan in case an outbreak occurs.

One of the most efficient ways to ensure that an outbreak does not evolve into an epidemic or pandemic is early diagnostics and confinement. Boehme et al. (2019) emphasized that early diagnostics can greatly decrease attack rate. For example, analysis into the Ebola epidemic in West Africa from 2013 to 2016 suggests that diagnosing 3/5th of the patients within one day compared to 5 days can drastically decrease the attack rate from 80 % to almost 0%. However, as mentioned in their paper, it can be difficult making diagnostic tests available to everyone due to financial and logistical reasoning. Therefore, a rapid diagnostic test that is user friendly and widely available to the public may be a plausible solution in preventing disease outbreak. (Boehme et al., 2019)

Most rapid diagnostic tests include reagent-based point-of-care tests, which are conducted in a clinic or an at -home test-kit. Recently, however, there have been cases in the use of machine-learning algorithms to develop a more accessible rapid diagnostic test (Zeng et al., 2023). Machine learning is an algorithm-based technology which utilizes the information provided by the researcher to deduce patterns and relationships among the variables. The learned information can then be used by the algorithm to make predictions. Due to its ability to handle a large amount of data and make predictions, it is an attractive alternative for rapid diagnostic tests. In contrast to other diagnostic tests/kits, machine learning based diagnostic tests will not need a lab kit nor a lab facility.

The significance of this project is two-fold. First, the developed program can allow for rapid-diagnosis, and, in-turn, a timely outbreak management if necessary. Second, machine-learning based diagnostic tests increase accessibility. This is especially relevant in the United States, where healthcare costs can be expensive. Individuals can use this program first to determine if they must visit the hospital for a final confirmation.

Research Methods

Data Collection

First, to accurately represent the adult-population, we will be collecting data from people 18-65 years old from the Florida government to train the machine learning model. The Florida's Department of Health keeps records on the number of flu cases in the database called Florida Influenza Surveillance Reports. An open record request will be made to the department of health, which will allow access to more specific data (age, symptoms, flu diagnostics etc.) without violating any HIPPA regulations. Then, this information will be used to train the machine learning program. Once trained, we will then allow campus members to input their symptoms for the prediction portion. After training, in order to take user input, we will design a website so that users can easily enter their symptoms into a form. After we conduct our research, we will also ask those who visited the hospital for a final confirmation for flu to fill out a second form to measure the accuracy of the program.

Machine Learning

Second, after collecting the data, we will then clean and process the data using the programming language Python and its built in data processing libraries Numpy and Pandas, to make sure that the model and effectively train and learn. This includes making sure that the data is accurate, free of outliers, and organized in a way that allows the model to be the most effective. Then, we will use a logistic regression model, gradient boosting model, and potentially even a neural network in order to make predictions based on the users given symptoms. Despite only needing one model, we will compare and contrast each model to see which one gives us the best accuracy for our software. After training and validating the best model using the Scikit-Learn library, we will then have the website feed the model the data from the user website form to make predictions on whether or not the users have the flu.

Data Analysis

Finally, after making the predictions, we can determine the accuracy of the program by comparing the diagnosis made by the machine learning program to that from the hospital. Statistical analysis will be performed to determine to what degree of confidence level the machine-learning program can accurately diagnose users for flu based on the symptoms provided. From this we can complete our research objective of managing outbreaks based on the number of confirmed cases.

Expected Outcome

From our research, we expect to develop a website that uses a machine learning model to diagnose a user for flu based on their symptoms. If successful, this website will serve as a rapid and accurate diagnostic

test for flu. It may even have the capability to replace the traditional rapid diagnostic tests (e.g., at home test kit) if the accuracy of the machine learning program is on-par with the current commonly used rapid diagnostic test kit. In addition, it can be used as a tool for a surveillance system for the flu. Therefore, it can also provide a new tool to manage outbreak more efficiently.

We aim to publish all the methodology and statistical information to a peer-reviewed journal to disseminate the result from this project. By publishing this information, the scientific community can benefit from new methods of machine-learning based diagnostic tests. If widely applicable, the methodology may be expanded for diagnostics of ailment outside of infectious disease (e.g., cancer). Given that machine-learning based rapid diagnostic tests are a relatively new idea, this project can serve as another step in advancing diagnostic tests toward higher accessibility.

UCF can greatly benefit from this product as it will be used to manage flu outbreaks on-campus. Every year, the number of flu cases on-campus fluctuates. By having a rapid diagnostic test that is highly accessible, UCF will be able to better monitor the number of cases live. This will allow the UCF Student Health Service to better prepare in-case of an outbreak. In addition, as the software allows for easier data recording, the UCF student health service can use the data compiled to estimate how many flu vaccines may be needed in the upcoming years. The information about the machine learning program will be presented to UCF's Student Health Service via a meeting to ensure that physicians will have the chance to ask questions about the model as needed.

Literature Review

Bi, Q., Goodman, K. E., Kaminsky, J., & Lessler, J. (2019). What is Machine Learning ? A Primer for the Epidemiologist. American Journal of Epidemiology, 188(12), 2222–2239. https://doi.org/10.1093/aje/kwz189

Chang, C., Lin, C., & Lane, H. (2021) Machine Learning and Novel Biomarkers for Alzheimer's Disease. International Journal of Molecular Sciences. 22, 2761. https://doi.org/10.3390/ijms22052761

Deb, S., Warule, P., Nair, A., Sultan, H., Dash, R., & Krajewski, J. (2023) Detection of Common Cold from Speech Signals using Deep Neural Network Circuits, Systems and Signal Processing, 42, 1707-1722. https://doi.org/10.1007/s00034-022-02189-y

Flu remedies: What to do if you have the flu. Geisinger. (2024, November 20). https://www.geisinger.org/health-and-wellness/wellness-articles/2024/11/06/16/46/flu-remedies-at-home

Kelly-Cirino, C. D., Nkengasong, J., Kettler, H., Tongio, I., Gay-Andrieu, F., Escadafal, C., Piot, P., Peeling, R. W., Gadde, R., & Boehme, C. (2019) Importance of diagnostics in

epidemic and pandemic preparedness. BMJ Global Health, 4(2), 001179. https://doi.org/10.1136/bmjgh-2018-001179

Mostafa, K. (2025) Influenza (seasonal). World Health Organization. https://www.who.int/news-room/fact-sheets/detail/influenza-(seasonal)

Nuwarda, R. F., Alharbi, A.A, & Kayser, V. (2021) An Overview of Influenza Viruses and Vaccines. Vaccines, 9(9), 1032. https://doi.org/10.3390/vaccines9091032

World Health Organization. (2025, February 28). Influenza (seasonal). World Health Organization. https://www.who.int/news-room/fact-sheets/detail/influenza-(seasonal)

Zhao, T., Xiao, X., Wu, Y., Ma, J., Li, Y., Lu, C., Shokoohi, C., Xu, Y., Zhang, X., Zhang, Y., Ge, G., Zhang, G., Chen, J., & Zeng, Y. (2023). Tracing the Flu Symptom Progression via a Smart Face Mask. Nano Letters, 23(19), 8960–8969. https://doi.org/10.1021/acs.nanolett.3c02492

Preliminary Work and Experience

We all have completed at a base level, Biology 1 and Statistics 1, to have a base level understanding about what we are referencing here and interpreting data. This will allow us to use the appropriate base level hypothesis testing in contributing to this research. Austin completed Machine Learning and Software Engineering classes. Austin is a Computer Science major and with his coursework, he will help contribute to understanding how machine learning will work. Ethan has had experience in classes consisting of Statistics 2 and Data Analysis classes. Ethan is a Data Science major and with his coursework, he plans to contribute to the data analytics that go into the machine learning model. Jae has experience in Microbiology and Molecular Biology classes. In addition, he has experience with biochemistry research. Jae is a Biomedical Science major and with his coursework, he plans to contribute to the research, a more medical scientific understanding of symptoms for our research.

IRB/IACUC Statement

Our research will require IRB approval, as we plan to use medical records to train the machine learning model, and then surveys with subjects' symptoms. We will fairly pick and select research subjects to allow the research to have the expected outcome.

Budget

In order to host our software, we will use Render and MongoDB Atlas. Render allows us to deploy the website so that the users can access the website from anywhere. MongoDB Atlas is a cloud based database service that allows the website to use a database without depending on a local machine. For one month of the Render professional plan it will cost $19 and for one month of the MongoDB Atlas dedicated plan it will cost $58. We will also need a domain name in order to have a forefront for a user to see the survey and complete it, this will cost us about $13 dollars a year for a ".net" ending domain. Bringing our entire total to $90, this will be our operations chunk of the budget. We also plan to try to get UCF students involved in our research; therefore, $1,000 will be allocated to incentivize students to part-take in surveys.

Render (professional plan)	$19
MongoDB Atlas	$58
Website Domain	$13
Survey Incentive	$1000
Total	**$1090**.00