Lowrey Peyton
15 December 2021

BIOF509 Final Project Report

I. INTRODUCTION

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the causative agent behind Coronavirus disease 2019 (COVID-19), which was declared a global pandemic in March 2020 by the World Health Organization. Just two months after the original SARS-CoV-2 genomic sequence was published in January 2020, phase I clinical trials for vaccines against the virus began in humans. This was an unprecedented feat in science. In late 2020, the phase III results of the mRNA vaccine trials were published, and these novel vaccines, the BNT162b2 from Pfizer-BioNTech and the mRNA-1273 from Moderna, proved extremely efficacious: over 90 percent. These vaccines were approved for Emergency Use Authorization in the United States by the Food and Drug Administration (Carvalho et al., 2021).

The Pfizer-BioNTech BNT162b2 vaccine consists of a two-dose (now three-dose, with the additional booster shot) regimen of the RNA encoding the prefusion-stabilized spike protein of SARS-CoV-2. With 95% efficacy in the phase II/III clinical trial (Polack et al., 2020), this vaccine has been instrumental in the global response to the pandemic. However, because this is the first time mRNA vaccines have been rolled out on a global scale, much is still unknown about their long-term effectiveness. Studies from Israel have detected a decreased immune response a few months post-inoculation (Goldberg et al., 2021; Levin et al., 2021), which means that a person's susceptibility to COVID-19 might increase over time without additional booster doses.

Additionally, with variants such as Delta and Omicron, there is increased transmissibility and decreased protection with the vaccine (Katella, 2021). Determining the long-term immune response and effectiveness of the COVID-19 mRNA vaccines is crucial for global public health.

II. ABOUT THE DATASET

The dataset, from Goldberg et al., consists of information about different demographic groups of people in Israel and the rate at which they tested positive for COVID-19 in the period of July 11, 2021, to July 31, 2021. It includes information of the gender, age group, religious sector, vaccination time period (by month), how many previous PCR tests they received, and when they tested positive during the study period. This dataset characterizes the vaccination status and positive test rate of approximately 4 million people. The dataset contains 1,135 samples with eight features for each sample.

This data is publicly available from Github, as indicated in the Python code: (https://raw.githubusercontent.com/yairgoldy/BNT162b2_waning_immunity/main/pos_data_days11-31_7.csv). The original study of this data (Goldberg et al., 2021) was reported in the New England Journal of Medicine after a surge in COVID-19 infections in Israel in mid-2021. The study utilized Poisson regression models and determined that most of the people contracting severe disease were elderly patients who had been vaccinated the earliest. The suspected reasons

behind the surge include decreased efficacy of the vaccine against the Delta variant and diminished immune responses to infection 6 months after vaccination. For the study, the group removed samples of adolescent-aged people because many were unvaccinated and people who had recently traveled abroad. They also analyzed disease severity in their model, but this information was not included in my neural network.

III. METHODS

I created a classification artificial neural network (ANN) for the analysis of the data. I selected this architecture because my goal was to determine if a neural network can accurately predict the rate of COVID-19 infection of a given demographic group when provided certain demographic features.

To preprocess the data, I converted the variables to ordinal sequential data or used one hot encoding. The Vaccine Period column consisted of the month and time frame of vaccination, either A or B (ex. JanB, FebA for late January and early February, respectively). Thus, I converted the earliest timepoint (JanB) to be 0 and sequentially up through May. For the age group, I similarly converted the youngest age group to 0, the middle age group to 1, and the oldest to 2. The Epi Week (when the positive test occurred) was converted from the week to 0, 1, and 2 from earliest to latest.

For gender and sector (religion), I used one hot encoding, so for gender, this was encoded in 2 columns with a 1 and a 0. For the religion, this was encoded in 3 columns with a 0, 0, and 1. I used one hot encoding for these columns because, with the other features, there was an inherent sequence to the data (early vs late timepoint, younger vs older in age), but there is no inherent order for gender or religious affiliation, so one hot encoding was best suited. Finally, I normalized each feature such that each feature had the same weight in the neural network. With the one-hot-encoded columns, this created 9 input features for each sample. For the output data, I created 5 bins with various ranges for the positive rate per 1000 people (see Table 1).

Bin range	Frequency
0	428
(0, 1]	416
(1, 2]	204
(2, 5]	79
(5, infinity)	7

Table 1. The bin ranges and frequencies of the five output classes of the ANN.

I constructed an ANN with 9 input features, two hidden layers with 6 neurons each, and 5 output classes. I selected 6 neurons for the hidden layers because this falls between 5 and 9, the number of output classes and input features, respectively. I trained using 80% of the data and

tested using 20% because this is standard (Willemink et al., 2020). The ANN used a batch size of 16, based on the Pytorch-TrainTest.ipynb example from the class materials.

IV. RESULTS

First, I plotted the accuracy of the epochs to determine how many epochs would be appropriate. I discovered that 4 epochs was ideal because the accuracy plateaued after this and avoided overfitting (see Figure 1). The ANN ran 4 epochs because, in my evaluation of the model, I discovered that more epochs led to overfitting (i.e. epoch accuracy of about 70% but testing accuracy of only 40%). Additionally, the accuracy of the epochs would plateau close to 60%, so adding more epochs did not increase the training accuracy very much. The model yielded an accuracy of 57.3%.

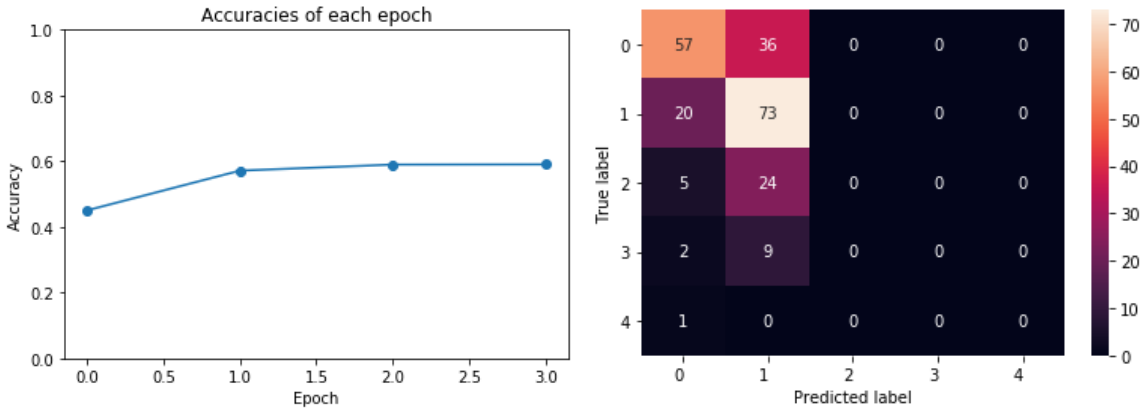

Figure 1. Left: the accuracy of each epoch. Right: a confusion matrix of the results of the model testing.

V. CONCLUSIONS AND FUTURE WORK

This artificial neural network was not effective at predicting the COVID-19 infection rate of fully-vaccinated Israeli residents. As seen in Figure 1, the model was not effective at predicting the labels of higher infection rates. This is likely due to the low sample counts from each of those bins (only 7 samples falling into label "4," which had infection rate of greater than 5 per 1000 people).

For future directions of this project, it would be interesting to incorporate the disease severity information to determine if a neural network is able to predict the disease outcomes, rather than simply the infection rates within demographic groups. Additionally, the Israel Ministry of Health has published more data related to this, including more samples, so those data could also be analyzed with machine learning techniques. Lastly, I developed a classification ANN for this dataset, but it would be valuable to create a multilayer perceptron to utilize the continuous variable of the rate of infection per 1,000 people.

References

Carvalho, T., Krammer, F., & Iwasaki, A. (2021). The first 12 months of COVID-19: a timeline of immunological insights. *Nature Reviews Immunology*, *21*(4), 245–256.

Goldberg, Y., Mandel, M., Bar-On, Y. M., Bodenheimer, O., Freedman, L., Haas, E. J., Milo, R., Alroy-Preis, S., Ash, N., & Huppert, A. (2021). Waning Immunity after the BNT162b2 Vaccine in Israel. *New England Journal of Medicine*, *85*(1), 1–10.

Katella, K. (2021). Omicron, Delta, Alpha, and More: What To Know About the Coronavirus Variants. *Yale Medicine: Family Health*. https://www.yalemedicine.org/news/covid-19-variants-of-concern-omicron

Levin, E. G., Lustig, Y., Cohen, C., Fluss, R., Indenbaum, V., Amit, S., Doolman, R., Asraf, K., Mendelson, E., Ziv, A., Rubin, C., Freedman, L., Kreiss, Y., & Regev-Yochay, G. (2021). Waning Immune Humoral Response to BNT162b2 Covid-19 Vaccine over 6 Months. *New England Journal of Medicine*, *84*(1), 1–11.

Polack, F. P., Thomas, S. J., Kitchin, N., Absalon, J., Gurtman, A., Lockhart, S., Perez, J. L., Pérez Marc, G., Moreira, E. D., Zerbini, C., Bailey, R., Swanson, K. A., Roychoudhury, S., Koury, K., Li, P., Kalina, W. V., Cooper, D., Frenck, R. W., Hammitt, L. L., … Gruber, W. C. (2020). Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine. *New England Journal of Medicine*, *383*(27), 2603–2615.

Willemink, M. J., Koszek, W. A., Hardell, C., Wu, J., Fleischmann, D., Harvey, H., Folio, L. R., Summers, R. M., Rubin, D. L., & Lungren, M. P. (2020). Preparing Medical Imaging Data for Machine Learning. *Radiology*, *295*(1), 4–15.