Group Members: Hrishikesh Bagalkote
Project Title: COVID-19 Risk Neural Network

Research Plan for Engineering Fair Project

Check out the grading rubric to understand exactly what your teacher is looking for.

Rationale: Include a brief synopsis of the background that supports your research problem and explain why this research is important and if applicable, explain any societal impact of your research.

Existing COVID-19 risk forecasting methods are mostly stasticical. This is done by making certain assumptions about existing circumstances in a geographical area and computing an approximate percentage risk in infections (CDC, 2020). This method of modeling is not entirely optimal as it takes longer to observe the trends in data and determine which factors were accurately weighted in the forecasting model. By optimizing a neural network to make inferences on geographic and demographic data, conclusions can be drawn about COVID-19 concentration and risk in targeted areas such as school districts and neighborhoods without the need for large amounts of processing power (CPUs and GPGPUs) and fast storage (Enterprise-grade NVME/PCIe SSDs).

Engineering Goal(s) and Expected Outcomes: Include IV, DV, and controls. States prediction as to what will be improved, manipulated, or adapted, which determines the IV (how students will vary their multiple designs/approaches). Clear success criteria for the project should determine the DV(s). Prediction is based on and firmly supported by specific information from the research. How is this based on the rationale described above? For Engineering, controls are comparisons of the project against how the problem is currently solved.

Independent Variables: Independent variables for this experiment are the number of epochs (cycles in the network) trained and the different combinations of factors trained on and incorporated into the model. Approaches to training can be varied by experimentation with the incorporation of different input factors. The fewer and more relevant the factors are to a specific area, the greater the accuracy of the predictions will be.

Dependent Variables: The general dependent variable is the accuracy of the model's predictions, which can be measured by the percentage risk difference from the actual output value in the dataset.

Control: The control in this experiment is the existing statistical models that mathematically forecast COVID-19 cases and related deaths in a target geographic area over time. We will compare the accuracy and reasonability of our neural network predictions and predictions from existing statistical models.

Prediction / Expected outcomes: It is hypothesized that a greater amount of training epochs will result in a higher prediction accuracy of the network. Approaches to training can be varied by experimentation with the incorporation of different input factors. The prediction technique is being improved on (when compared to mathematical statistical disease models), as the utilization of a neural network allows us to quantify the relative impact of each chosen factor on COVID-19 deaths and risk in a selected geographical area. Predictive statistical models can be useful but should not be overinterpreted (Jewell et al, 2020, 1). Supervised machine learning has seen huge adoptions in recent years in applications such as computer vision, but unsupervised neural networks are able to learn from training data surprisingly effectively also. (Metz et al., 2016, 1).

Engineering Goal: The end goal of this project is to create a superior method of disease forecasting when compared to statistical modeling in utilizing a neural network. The deployed neural network and model must not be resource intensive, as they should be able to run optimally on weaker hardware without the use of GPU arrays.

Success Criteria: The produced model from the neural network has reasonable weights corresponding to each input factor and outputs a reasonable prediction that is insightful about a targeted area and low margin of error when compared to real out values in the dataset.

Procedures: Detail all procedures and experimental design including methods for data collection. How will you show that the design works as intended? Engineering design will be less detailed in the initial project plan and will change through the course of research. If such changes occur, a project summary that explains what was done is required and can be appended to the original research plan.

Materials List:

- Laptop or desktop with at least 4 GB of ram and 100 MB of storage to store the Java Runtime Environment, CSV datasets, and the network itself.
- Windows, Linux, and macOS will all work with this project, as the JRE is cross platform.
- GPUs are not required as the network training algorithm is optimized for the CPU.

Procedure:

1) Obtain CSV formatted dataset with selected input factors along with an output variable of recorded COVID-19 deaths.

   Example:

Factor 1	Factor 2	Factor n	Total COVID-19 deaths in area
x...	y…	z....	5000

2) Train and save the model locally using the CLI tool, which is a wrapper for the neural network itself.

   This is achieved by inputting a training dataset and desired number of epochs (training cycles) into the CLI wrapper.

3) Enter the prediction mode of the CLI tool, input desired dataset to predict on, along with the produced model from before.

   NOTE: The prediction dataset must be formatted in the same way as the training dataset because of the way weights are stored for each factor in the model. This means that the prediction dataset CSV must have the same number of fields as the training dataset corresponding to the same numeric input factor value.

4) The model's prediction for each line will be shown through the CLI wrapper.
5) Test the same dataset using a different number of epochs and compare the resulting model's accuracy after predictions.
6) Change the factors inputted into the training dataset and observe changes in the behavior of the model. Note that factors changed in the training dataset must also be changed in the testing dataset for the prediction algorithm and model parser to work properly.
7) Record epochs trained, factors included, predictions, and actual values for each trial. This data will be used for analysis on which training method is the most effective later.

Risk and Safety: Identify any potential risks and safety precautions needed.

N/A

Data and Analysis: Describe the procedures you will use to analyze the data/results. This is not the same as data collection. This is where you talk about statistics.

Datasets used to train the neural network and produce a model will contain numerical input factors such as average age of a geographic area, population density, poverty rate, income statistics, air quality, total confirmed COVID-19 cases, etc. Datasets will also have an output variable, which is the current number of COVID-19 recorded deaths in the specified area (state, city, couty, etc) updated periodically. The neural network will find the relationship between each factor in the dataset and the output variable of resulting COVID-19 deaths. Once a model has been produced with computed weights for each dataset factor, the network will be tested against a training dataset so that predicted outputs of forecasted deaths can be compared with the real values. A percent error measure will be the easiest approach to measuring this difference. The number of epochs (training cycles) that are optimal and reasonable for a statistically correct model should be found. Too few training epochs will result in a low accuracy, too many epochs can potentially result in unnecessarily long training runtimes and overcorrection for error. Another experiment that needs to be conducted is the combination of factors used to train the network. If irrelevant factors are removed from the datasets, accuracy of predictions will increase, lowering the error margin. Below are examples of data tables we can use to organize our data.

Epochs	Average Percent Error of Prediction
10000	15%
50000	10%
etc...	x%

Factors Trained in Model	Average Percent Error of Prediction

Population density, total confirmed COVID cases	15%
Population density, total confirmed COVID cases, median age	12%
etc...	x%

Graphs:

Scatter plot with epochs on the x axis and average percent error on the y axis can be used along with a line of best fit to analyze the effect of more training cycles over time.

A bar graph can be used with factors trained on the x axis and average percent error on the y axis to understand which factors and input variables are irrelevant to the accuracy of the model.

Conclusions:

By analyzing the data collected from the training processes of the neural network along with inferred weights from produced models from different trials, we can conclude the following:

- Which methods of training a neural network are most effective.
- Which algorithms within the neuron of the network are best to optimally produce models with short training runtimes.
- Which geographical and demographic factors have the most impact on COVID-19 deaths, transmissibility, and risk.
- Which geographical and demographic factors are irrelevant or have little impact on COVID-19 deaths, transmissibility, and risk.

Bibliography: Using APA format, list major references (e.g. science journal articles, books, internet sites) from your literature review.

Al-Masri, A. (2019, January 19). *How Does Back-Propagation in Artificial Neural Networks Work?* Towards Data Science. Retrieved November 12, 2020, from https://towardsdatascience.com/how-does-back-propagation-in-artificial-neural-networks-work-c7cad873ea7

Baltazar, G. (2018, September 13). *GPU vs CPU in Machine Learning*. Oracle. Retrieved November 13, 2020, from https://blogs.oracle.com/datascience/cpu-vs-gpu-in-machine-learning#:~:text=As%20a%20general%20rule%2C%20GPUs,be%20carried%20out%20in%20parallel.

CDC. (2020, August 10). *Assessing Risk Factors for Severe COVID-19 Illness*. CDC. Retrieved November 13, 2020, from https://www.cdc.gov/coronavirus/2019-ncov/covid-data/investigations-discovery/assessing-risk-factors.html

CDC. (2020, November 12). *COVID-19 Forecasts: Deaths*. CDC. Retrieved November 13, 2020, from https://www.cdc.gov/coronavirus/2019-ncov/covid-data/forecasting-us.html

Jewell, N. P., Lewnard, J. A., & Jewell, B. L. (2020, April 16). Predictive Mathematical Models of the COVID-19 Pandemic Underlying Principles and Value of Projections. *JAMA*, *323*(19), 2. doi:10.1001/jama.2020.6585

Jiang, H., & Nachum, O. (2020). Identifying and Correcting Label Bias in Machine Learning. *Proceedings of Machine Learning Research*, *108*, 10. http://proceedings.mlr.press/v108/jiang20a.html

Kilpatrick, S. (2019, January 22). *How to Prevent Machine Bias in AI*. Logikk. Retrieved November 13, 2020, from https://www.logikk.com/articles/prevent-machine-bias-in-ai/

Lipton, Z. C. (2015, January). *The High Cost of Maintaining Machine Learning Systems*. KDNuggets. Retrieved November 13, 2020, from https://www.kdnuggets.com/2015/01/high-cost-machine-learning-technical-debt.html

Maragakis, L. (2020, June 25). *Coronavirus and COVID-19: Who is at higher risk?* Hopkins Medicine. Retrieved November 13, 2020, from https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus/coronavirus-and-covid19-who-is-at-higher-risk

Mayo, M. (2017, October). *Neural Network Foundations, Explained: Updating Weights with Gradient Descent & Backpropagation*. KDNuggets. Retrieved November 12, 2020, from https://www.kdnuggets.com/2017/10/neural-network-foundations-explained-gradient-descent.html

Metz, L., Radford, A., & Chintala, S. (2016). Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv*, 16. https://arxiv.org/abs/1511.06434

Meyer, H. (2020, August 14). *COVID-19 data often slowed by paper forms and fax machines, thanks to creaky US public health data system*. USA Today. Retrieved November 12, 2020, from https://www.usatoday.com/story/news/health/2020/08/14/covid-19-data-slowed-us-lack-national-health-data-network/3369859001/

Rosebrock, A. (2019, February 4). *Keras: Multiple Inputs and Mixed Data*. Py Image Search. Retrieved November 13, 2020, from https://www.pyimagesearch.com/2019/02/04/keras-multiple-inputs-and-mixed-data/

Sharma, S. (2017, September 23). *Epoch vs Batch Size vs Iterations*. Towards Data Science. Retrieved November 13, 2020, from https://towardsdatascience.com/epoch-vs-iterations-vs-batch-size-4dfb9c7ce9c9

Shukla, S. (2018, June 12). *How Does Unsupervised Machine Learning Work?* Upgrad. Retrieved November 12, 2020, from https://www.upgrad.com/blog/how-does-unsupervised-machine-learning-work/

Vestal, C. (2020, August 3). *Lack of Public Data Hampers COVID-19 Fight*. Pew. Retrieved November 12, 2020, from https://www.pewtrusts.org/en/research-and-analysis/blogs/stateline/2020/08/03/lack-of-public-data-hampers-covid-19-fight

Wieczorek, M., Silka, J., & Wozniak, M. (2020, November). Neural network powered COVID-19 spread forecasting model. *Science Direct*, *140*(Chaos, Solitons, and Fractals), 5. https://doi.org/10.1016/j.chaos.2020.110203