REU Final Report

Jackson Ginn[a], Dezmon Patten[a], Jacob Rottenberg[b]
[a]Department of Mathematics, University of South Carolina, Columbia, South Carolina
[b]Department of Mathematics, University of Massachusetts, Amherst

1 Overview

Background Time series data is a type of data in which observations are collected over time. Common examples of such data are amount of rainfall each month, or quarterly sales reports. Being able to model such data and predict outcomes allows us to plan accordingly. Recurrent Neural Networks (RNN's) are the typical choice to model this fashion of data, but are mostly used when time intervals are uniform. Deeper RNN's be adjusted to model time-series data where the time intervals are not uniform, but this can quickly become memory expensive due to needing to store intermediate quantities. Neural Ordinary Differential Equation networks (NODEs) help tackle this problem.

NODE [1] Consider the RNN $\mathrm{h}_{t+1} = \mathrm{h}_t + f(\mathrm{h}_t, \theta_t)$. By taking smaller steps and adding more layers to the limit an ordinary differential equation can be generated as $\frac{d\mathrm{h}(t)}{dt} = f(\mathrm{h}(t), t, \theta)$. Now consider optimizing the loss function as follows:

$$L(\mathbf{z}(t_1)) = L\left(\mathbf{z}(t_0) + \int_{t_0}^{t_1} f(\mathbf{z}(t), t, \theta)dt\right) = L(\mathrm{ODESolve}(\mathbf{z}(t_0), f, t_0, t_1, \theta))$$

Solving this optimization requires knowing the relationship between L and $\mathrm{z}(t)$. As such, the adjoint is defined:

$$\mathbf{a}(t) = \frac{\partial L}{\partial \mathbf{z}(t)} \rightarrow \frac{d\mathbf{a}(t)}{dt} = -\mathbf{a}(t)^\top \frac{\partial f(\mathbf{t}, t, \theta)}{\partial \mathbf{z}}$$

This concept is the central idea of an NODE.

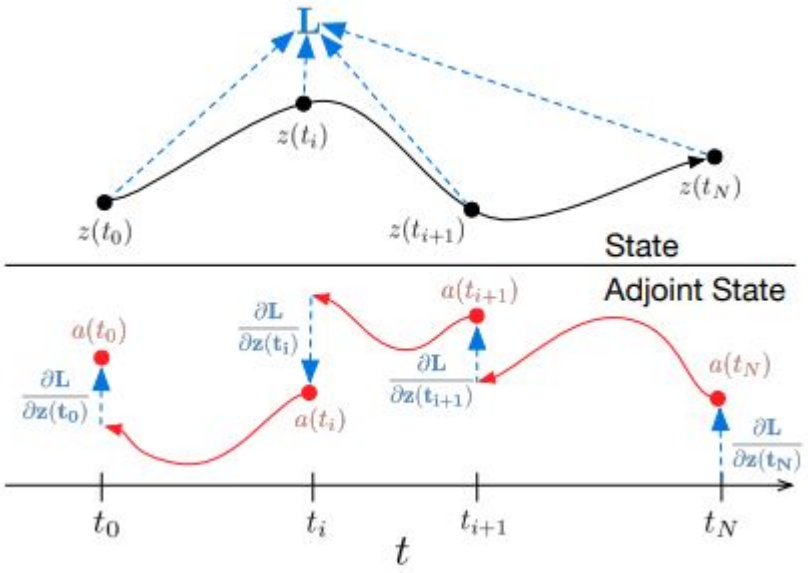

Figure 1: Reverse-mode differentiation of an ODE solution

[1] Chen, Rubanova, Bettencourt, and Duvenaud (2018)

2 Problem Statement

In this project, we attempt to predict various biomarkers of cancer patients using deep learning. These biomarkers follow a dynamical system. Specifically, we use neural ordinary differential equations (NODEs) to analyze patients' metabolic levels and approximate those levels in the short-term. These NODE models are personalized for each patient. Through these personalized models, we hope to provide an accurate prediction of a patient's metabolic levels. With this prediction, doctors could make more informed decisions, allowing for better quality of life and outcomes for the patients.

3 Method

Data Acquisition and Preprocessing We received data containing metabolic panel results for cancer patients taken between 2019 and 2022. Nineteen metabolic indices were measured during the metabolic procedure for all the cancer patients: Glucose, BUN, Creatinine, BUN/Creatinine Ratio, Anion Gap, CO2, Chloride, Sodium, Potassium, Total Protein, Albumin, Globulin, A/G Ratio, Aspartate Aminotransferase (AST), Alanine Aminotransferase (ALT), AST/ALT Ratio, Alkaline Phosphatase, and Total Bilirubin.

We selected 18 biomarkers with the minimal missing data rate as the time series dataset to establish the time-dependent discrete dynamical system, where time is represented by the number of minutes since the first data entry was taken. First we calculated the standard deviation and mean of each biomarker. Then used this to replace outliers, which we define as data points that were not in the range $(\mu \pm 3 * \sigma)$, with the value for the boundary $(\mu \pm 3 * \sigma)$. Next we standardized the data using the formula, (value - mean)/standard deviation, such that when all the models are incorporated in the model the loss is not dominated by one feature that has higher averages than the others. Next, we calculated the correlation coefficient of each biomarker to obtain the optimal number of biomarkers from our data that will be used in building our model, by removing correlated biomarkers. The biomarkers included in our model were Glucose, BUN, Creatinine, Anion Gap, CO2, Chloride, Sodium, Potassium, Total Protein, Albumin, Globulin, A/G Ratio, Aspartate Aminotransferase (AST), Alanine Aminotransferase (ALT), AST/ALT Ratio, Alkaline Phosphatase, and Total Bilirubin. At this point in our data preprocessing process the graphs for each biomarker were irregular as they were composed of short waves which can best be described as non-smooth areas of the graph , long waves which are smooth areas of the graph, and noise. By calculating the sliding window average we were able to remove noise and short waves, leaving us with a smoother curve for our model. Finally we split our data into test/train/validation sets for our model to help prevent overfitting.

4 Analysis

One problem we faced when using the NODE models was creating a long term prediction of the data. By long term prediction, we mean providing an initial point to the NODE model at or near the beginning of the dataset and having the model predict at each of the time points in the data set afterwards. Our NODE models were not very effective in predicting farther than a few points away from the initial value. The long term predictions often simply predicted that the value at a given time would be approximately the mean rather than predicting any fluctuations. After failing to find a method to accurately predict in the long term, we decided to shift our initial point to be near the end of the data set to provide much more accurate predictions, albeit only in the short term. In the model's actual use case, this is not much of an issue as doctors will likely only be concerned with short term predictions of biomarkers. We were able to create short term predictions with less than 10% relative error, which is acceptable in a clinical setting.

The datasets we were using to train our NODE models were not entirely ideal. The data from the spinal cord patient was collected across about two and a half years, and the data from the pineoblastoma patient was collected across about a year. During both of these time periods, data was sampled extremely irregularly. Some measurements were taken within days or even hours of each other, while other pairs of measurements had over a week separating them. Initially, we corrected for this by using linear interpolation to create a consistent distance between all points. We only interpolated our spinal cord dataset, which increased our number of points from 26 to 1521. We found that the NODE model would accurately

learn the training data by essentially treating it as a piecewise linear function, which translated poorly to predicting the validation dataset. We decided against interpolating the entire dataset and later used a sliding window average and standardization to smooth the data so that the NODE model could function more accurately.

5 Results

After fine-tuning our NODE models, we were able to successfully predict some biomarkers using short term predictions. This final NODE model's most accurately predicted biomarker was Sodium, which had a relative error of approximately 7.9%. However, this model was not accurate with all biomarkers, as it struggled to accurately predict bilirubin for example. The graphs shown in the figure below contains a line connecting the points of the original data marker values , which are plotted along side our predicted values obtained from our NODE model. As mentioned before some biomarkers worked better with our model than others but for the most part the graphs show that this model if tuned more accurately could be used to predict patient biomarkers values over a set time period. In terms of cancer patients, this models ability to make short term predictions can be used by doctors to assist with making a more regimented treatment plans with the ultimate goal being a more targeted approach to eliminating the cancer at a successive rate.

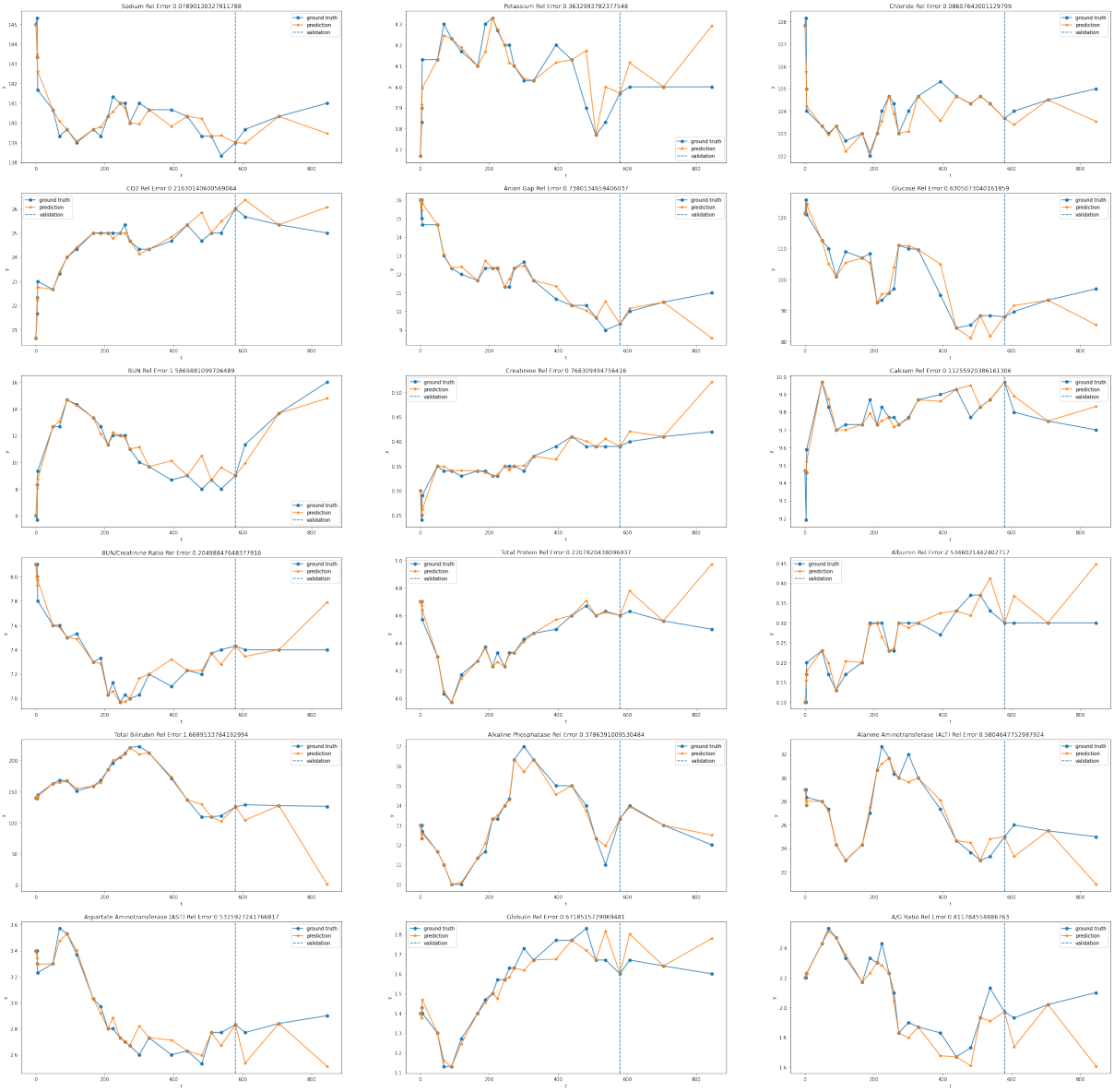

Figure 2: Results for each biomarker

6 Conclusion

With all the information presented, the focus of our research project is more concise as we were able to shift our focus from using deep learning techniques to predict long term metabolic rates of cancer patients to predicting more accurate short term metabolic rates. With our model being able to make short term predictions for two patients, this would then allow it to be expanded to work with a larger sample size. With an unpredictable disease such as cancer, it can be challenging to make long term predictions. By focusing more on obtaining more reliable short term predictions with a low relative error, this was able to create more sustainable outcomes for medical professionals. The model's ability to learn over time helps narrow the biomarkers' values to create a more useful treatment plan for the patient. Although the focus of the research was to create a long term prediction method, our data demonstrates that learning and understanding just how important the short term predictions are provided a large level of benefit. By shifting the focus of our research, we were able to find more conclusive results that are better applied to time series data.

References

Chen, T. Q., Rubanova, Y., Bettencourt, J., & Duvenaud, D. (2018). Neural ordinary differential equations. *CoRR*, *abs/1806.07366*. Retrieved from http://arxiv.org/abs/1806.07366