EE / CprE / SE 491 – sdmay24-47
Accurate Cancer Prediction Using AI
Bi-weekly Report 2

Jan 28 – Feb 10
Client: Ashraf Gaffar
Faculty Advisor: Ashraf Gaffar

Team Members:

Jack Sebahar — *Project Manager / Client Liason*
Nicholas Otto — *Model Design Lead*
Mason Wichman — *UI Design Lead*
Lal Siama - QA Engineer, Tester
Helen Lau - UI Design
Isaiah Mundy - UI Design Test

Weekly Summary

So far this semester, our main goal is to find a meaningful way to preprocess the data into a model. Our data is formatted as a CSV file for each patient representing a cell culture image. Each patient has a corresponding file with cancer recurrence information. As of right now, it is unknown if there are correlations between the outcome and data itself. The past 2 weeks have consisted of creating simple models as a baseline to gague the potential accuracy of using our data. We have experimented by using data preprocessing libraries, like Pandas, and extracting useful things from the data. Via this method, we have developed a base model with very low accuracy. Simultaneously, we have intialized our development environment on AWS. A part of our project is to potentially compare results on GCP and AWS. We have started our process on GCP and now need to begin developing on AWS. Since none of our team is very familiar with AWS, this has been not as trivial. In the upcoming weeks, we anticipate beginning development on AWS and increasing the accuracy of our model.

Past Week Accomplishments

Data preprocessing methods

- Our data consists of CVS files representing a cell culture image along with a corresponding patient for each file and their time of recurrence. A large goal of ours is determining (if any) a meaningful way to parse this data as an input for a model. Some options we have investigated include the Python Pandas library, which is used for data manipulation and analysis. We have also been researching base Keras models, and seeing if any of them are designed for a use case similar to ours

Initial Data Preprocessing and Model Training

- Created preprocessing functions for filtering data to retrieve peak data points and for normalizing x and y values. These functions will be usable for training future models and have the potential to improve accuracy.
- Attempted to visualize the raw data using MATLAB data cleaner. The survival column was heavily analyzed for logistic function. The data only showed the linearly decreasing trend. I also ran a logistic growth equation to calculate the growth rate of the cell culture, r, and generate the line of best bit for the survival data. There were a lot of errors and missing data. Results showed nothing much.

- Wrote a Python script to separate data into usable samples that we have survival data for, and incomplete samples without survival data. Also did the same thing for the survival data to separate the data that we have samples for from the data we have corrupt/no samples for.
- Began training a simple model on Google Colab, starting with a very small sample size (25) to try to predict time of survival/recurrence based on patient sample data. Initially, the model had very poor accuracy but was able to see some improvements by retraining the model on all of our usable data (340 samples). Also experimented with adding more layers to the neural network, which improved the accuracy of the model by a small amount. Currently, the average MSE the model outputs is ~3400, which translates to roughly 58 months in our context, so it is still far from ideal.

```
Epoch 1/9
9/9 [==============================] - 1s 3ms/step - loss: 23331.3848
Epoch 2/9
9/9 [==============================] - 0s 3ms/step - loss: 4240.2876
Epoch 3/9
9/9 [==============================] - 0s 2ms/step - loss: 3448.0798
Epoch 4/9
9/9 [==============================] - 0s 3ms/step - loss: 3020.9810
Epoch 5/9
9/9 [==============================] - 0s 2ms/step - loss: 3048.1758
Epoch 6/9
9/9 [==============================] - 0s 2ms/step - loss: 2973.0918
Epoch 7/9
9/9 [==============================] - 0s 2ms/step - loss: 2881.8486
Epoch 8/9
9/9 [==============================] - 0s 3ms/step - loss: 2863.4856
Epoch 9/9
9/9 [==============================] - 0s 2ms/step - loss: 2873.8875
3/3 [==============================] - 0s 5ms/step - loss: 3443.1614
Mean Squared Error on Test Data: 3443.161376953125
```

Began investigating AWS approach

- An important portion of our project is to generate models on different platforms. The purpose of this is to determine if there are differences between the accuracies of models on different platforms. Our current development environment consists of GCP and Colab. We began initializing our AWS environment. This has not been as trivial as GCP since nobody on our team has prior experience with it.

Pending Issues

- Determine collaborative environment for AWS
    - GCP uses colab for collaborative development, we need to set up our AWS environment in a similar fashion.
    - AWS uses Sagemaker notebook snapshots, but this means changes made won't affect other notebooks
    - We need to figure out how to create a similar development environment as in GCP
- Determine billing options for AWS
    - Some features of AWS may require a billing account. We have been approved to use funds for our project, but this will have to be consulted with our advisor

Individual Contributions

Team Member	Contribution	Weekly Hours	Total Hours
Jack Sebahar	● Data distribution ● Preprocessing methods research ○ Pandas ○ Keras	6	18
Nicholas Otto	● Data preprocessing ● initial model training ○ First python test	8	23
Mason Wichman	● Data preprocessing ○ Data peak extraction script ● AWS initialization	8	20
Lal Siama	● Data preprocessing ○ Methods research	6	18
Helen Lau	● Data preprocessing ○ Methods research	6	18
Isaiah Mundy	● Data Preprocessin ● Model Design	6	18

Plans for Coming Week

- Investigate more effective options for data preprocessing - Helen, Siama
  - Determine if there are more suitable options for preprocessing our data
  - Determine if any Keras models are designed to work with data like ours
- Increase Model Accuracy - Nick, Jack, Isaiah
  - Try normalizing the input data
  - Experiment with different types of models and layers
  - Potentially try different ways of formatting the model input data
- Fully initialize the AWS environment - Mason
  - Current development is in GCP
  - Part of goal is to compare results on different platforms
  - Using Sagmaker, similar to Colab
  - Determine if there are any billing requirements