Theresa Sheets
Machine Learning Midterm Report
May 4, 2019

Cystic Fibrosis (CF) is a progressive genetic disease. The disease is caused by an incorrectly functioning protein that causes the mucus in the organs of individuals with CF to become thick and sticky. This mucus builds up in the lungs and traps germs leading to infections and respiratory distress. Many individuals with CF also develop diabetes. While 9.4% of adults in the general population have diabetes, 40-50% of adults with CF develop diabetes.

In CF related diabetes, sticky mucus causes scarring of the pancreas. This prevents the pancreas from producing adequate amounts insulin. Many people with CF related diabetes do not know they have diabetes until they are tested, so part of the CF care guidelines recommend annual diabetes testing of CF patients. CF related diabetes can be well managed with insulin and blood sugar monitoring. An effective way of predicting diabetes diagnosis could help to identify individuals at highest risk and get them treatment more quickly.

As CF is a progressive disease, it is important to be able to predict survivorship based on physiological measures. The first goal of this project will be to identify the features most likely to predict 5 year survivorship of patients with CF. The second goal of this project will be to identify the features most likely to predict a diabetes diagnosis within the next five years.

Previous research has identified forced expiratory volume (FEV) as the most significant predictor of 5 year survivorship of CF patients. While FEV is an incredibly powerful tool to determine those who are most dangerously sick, it is a poor predictor amongst patients with high FEV. We hope using decision trees to partition the data will identify the most powerful predictors amongst groups of individuals with high FEV and low FEV. The structures of those trees might will then be compared to elucidate the nature of the disease.

A model has been developed using logistic regression to include previously identified features. This project will use the perceptron algorithm to identify a classifier to predict five year survivorship. Then this model will be compared to the previous model to see if it is a better method to predict survivorship. I was originally planning on using data about diabetes in these patients to examine the relevance of diabetes diagnosis, but the participants with diabetes were a very small percentage of the total dataset.

The data which will be used to conduct this research are a set of ~48000 patients divided into 4 five year cohorts each with ~60 time points in which they were assessed by a physician. The first year of

physiological measures for each patient will be used as the input parameters, and the patient's survivorship at the end of the five years will be used as the labels for the dataset.

In order to complete this project, I have completed the necessary trainings to receive IRB permission to conduct research using human subjects and this particular dataset. I have conducted a literature review exploring common features of cystic fibrosis and current models of the disease progression. I have edited my code to be able to handle the attributes from the data. I have met with the researchers who have previously used this data to identify their most pressing questions regarding the data.

To complete this project I used two different machine learning algorithms to evaluate the dataset. The ID3 algorithm which makes a decision tree to better understand the data, and perceptron which is an algorithm for developing a linear classifier. Before using those algorithms on the dataset, I first needed to make it usable. The original dataset was made of two large datasets with a combined total of 81 parameters, and multiple rows for each individual. I took the first recorded instance of the individual and concatenated it with their five year survival rate. I then chose 22 parameters without identifying information to use in the analysis. Finally, as I was unclear how to effectively handle missing data for this set, I removed all participants with values of NA for any of the 22 parameters I chose. In the end, I was left with a dataset containing observations of 30,792 individuals diagnosed with CF. For each of those individuals, I had information on: 'mssa', 'mrsa', 'h_flu', 'pseudo', 'burkho_complex', 'alcalig', 'steno', 'enterobacter', 'serratia_marcescens', 'aspergillus', 'candida', 'scedosporium', 'mabscessus', 'mai', 'sex', 'suff', 'diabet', 'impglu', 'fev1pct_best', 'zscore_best', 'trunc03', 'dflag5'. Most of these parameters are binary classifications regarding the individual's colonization with a variety of pathogens while 'fev1pct_best' measures forced expiratory volume (related to lung capacity) and 'zscore_best' measures relative weight for age. Both of these parameters were turned into binary classifications for the ID3 algorithm based on the median of the parameter. Dflag5 is the binary survivorship label. To train the data, I randomly selected 10% of the data (3061 observations) and used that dataset to train the algorithms.

Unfortunately, my analysis using the ID3 algorithm proved relatively unsuccessful. Of the training set, only 4% of the observations were individuals who passed in 5 years. This means that regardless of the depth of the decision tree, the prediction was always the same, as the majority of individuals survived 5 years, and thus the labels of most of the leaves of the tree were also the surviving label. I know my algorithm was working correctly, as I successfully tested it on the bank dataset used in our class. In the future, I hope to explore other methods to artificially inflate the proportion of individuals who do not survive 5 years, so that the decision tree will be better able to identify those individuals most at risk. I did not choose to explore those options for this project, because I was concerned about unfairly

biasing the data towards the traits of a few individuals. While I did not learn much from the consistent accuracy of the ID3 algorithm 94.9402442193% regardless of the depth, the structure of the decision tree was compelling as it agreed with the researchers prior understanding that 'fev1pct_best' and 'zscore_best' are the best predictors of five year survivorship. The decision tree split first on fev1pct_best and then on zscore_best at every single node. On the third layer, things got a bit more interesting, and they split on 'suff', 'pseudo', and 'burkho_complex'. When the parameter 'trunc03' was included it was the first split for the decision tree. I chose to not include it because the researchers had some questions about it's validity, it is a measure of number of acute pulmonary exacerbations and there were questions about it's consistency across the years.

The results of the perceptron algorithm are much more clear. The algorithm appears to converge on a classifier successfully with small returns on testing accuracy as the number of epochs increase. The table below shows the testing and training accuracy as the number of epochs increase. The weight vector found at T=10 is [mssa=-0.2, mrsa=-0.2,h_flu= -0.3, pseudo=12.400000000000041, burkho_complex=4.099999999999993, alcalig=0.15000000000000002, steno=0.8500000000000002, entarobacter=-0.15000000000000002, serratia_marcescens=-0.1, aspergillus=0.8000000000000002, candida=-0.35, scedosporium=0.0, mabscessus=0.44999999999999996, mai=-0.1, sex=10.100000000000009, suff=14.500000000000071, diabet=2.799999999999998, impglu=0.9000000000000002, fev1pct_best=-3.696916568113793, zscore_best=-17.904626055695747, bias=14.550000000000072]. These results are interestingly different from those found by the ID3. Focusing on magnitude alone, it seems apparent that 'pseudo', 'burkho_complex', 'sex', 'suff', 'diabet', 'fev1pct_best', and 'zscore_best' are particularly important. This makes sense, given the disease and matches nicely with the paper by Liou et al which found: gender, fev, weight for age z_scores, suff, diabet, staph, burkho_complex, and the number of acute exacerbations to be the most important parameters for predicting survivorship.

Epochs	Training Accuracy	Testing Accuracy
1	0.95459	0.95008
10	0.95557	0.95223
50	0.95949	0.95297
100	0.95949	0.95323

Ultimately I feel satisfied that my results agree with those found by previous researchers. I hope to expand upon these results to find more meaningful and different information than previously explored. In the future, I hope to find a better way to train this data on a decision tree. There were options in the handling of continuous data, and I chose to simply turn them into binary values. There may be important subgroups within those continuous values, which are more effectively divided into thirds or quartiles. I hope to include more information about diabetes diagnosis and prediction as there is less research on those than CF alone. I also hope to compare across cohorts. This analysis combined all cohorts into one group, but it would be easy to compare across groups and see if that helps the predictive capability.

A github repository containing my code can be found at:
https://github.com/holysheets/MachineLearning/tree/master/FinalProject

References:

https://academic.oup.com/aje/article/153/4/345/129039
https://www.ncbi.nlm.nih.gov/pubmed/27248696
https://www.cff.org/What-is-CF/About-Cystic-Fibrosis/
https://cysticfibrosisnewstoday.com/cystic-fibrosis-related-diabetes-cfrd/
https://www.cff.org/Life-With-CF/Daily-Life/Cystic-Fibrosis-Related-Diabetes/