**Breast Cancer Prediction and Classification using Neural Network**

**Jack Crum Liuqing Yang Sixian Zhu Biying Zhu**

**Team Name: GW Navy**

Our project aims to design a machine learning algorithm for breast cancer risk prediction and diagnosis. Breast Cancer is found to be the second leading cause of death among women and the fifth major cause of cancer death. According to National Breast Cancer Foundation, an estimation of over 252,710 women in the United States will be diagnosed with breast cancer each year, and more than 40,500 will die. Early detection and diagnosis are the most effective ways to prevent death from breast cancer and a machine learning algorithm that can predict malignancy through tumor images would greatly decrease detection time and misdiagnosis.

The dataset used in this project is the Breast Cancer Wisconsin (Diagnostic) Data Set, which is available through UCI Machine Learning Repository. Attributes information in this dataset includes ID number and Diagnosis (M = malignant, B = benign). The features are generated from image analysis of fine needle aspirates (FNA) of breast masses to describe the cell nucleus. Overall, dataset consists of 569 observations, 357 benign and 212 malignant cases, and 31 features of tumor attributes. This data set contains a large enough sample size and number of attributes, though some dimensional reduction will be necessary, to train and evaluate the algorithm to determine model performance.

As some features in our dataset are on different scales from one another, we will preprocess the data by normalizing all features to the same scale between 0 and 1. We will adjust weight and bias by supervised learning rule using class labels with features. The majority of the dataset will be used as the training set to obtain the optimal parameters of the predictive model. Since there are two types of diagnosis and ten attributes, a perceptron neural network with ten-dimensional inputs can be applied in this case. The output of perceptron can be 1 when the diagnosis result is benign, and -1 when the result is malignant. In addition to the training set, some groups will be extracted as the test set to evaluate how well the algorithm performs. By comparing the predictive results with actual results in the test set, we will be able to minimize the optimization index. The metrics we will use for measuring model quality will include classification accuracy and AUC. We will also develop decision tree and random forest models and use the associated metrics of entropy, gini index, and classification error to maximize information gain and select the best features for predictive classification.

There are several past literatures that provide references to using perceptron neural network for breast cancer prediction. For example, Mojarad, Dlay, Woo and Sherbet (2011) designed a multilayer perceptron to evaluate four cellular and molecular breast cancer markers, further confirming the ability of neural network in capturing the complex patterns in breast cancer tumor markers. Dural and Ganesh (2017) proposed a novel iterative linear regressive perceptron classifier to predict the instances of breast cancer, and the model achieved a highest accuracy of 100%. Multilayer Perceptron with Dense Scale Invariant Feature Transform (DSIFT) features is also found to be effective for the breast cancer detection (Jadoon et al., 2017).

**Time Frame:**

10/3-10/16:

First meeting.

Make agreements on dataset and project methods selections.

Finish project proposal.

11/1-11/20:

Process the data in advance.

Visualize the data set.

Second meeting for communicating the understand of the data set.

Decide the specific way of analyzing.

11/24-11/27:

Analyze the data and get the model.

Third meeting for communication of the model adjustment.

11/27-12/4:

Fourth meeting for the conclusion of the project.

Finish final report and prepare for the presentation.