IST 718 Final Project Proposal

Group 1

Mike DeMaria, Lu Guo, Haotian Shen, Casey Walsh

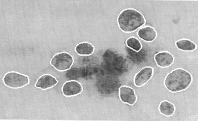
**Scenario**

Cancer has traditionally been diagnosed through an invasive biopsy, where a large sample is removed from the body through a surgical cut. This procedure can be painful, intrusive and leave the patient at risk of scarring and infection. Patients may also be less likely to undergo a biopsy for these very same reasons.

Medical researchers were interested in finding a minimally invasive technique. A new procedure was developed called fine needle aspirations (FNA). A small needle is inserted into the tumor and withdraws a sample. Using this process results in far greater patient comfort, but produces a much smaller sample. Only a small number of cells are withdrawn in any given sample.

It is believed that a machine learning algorithm can be developed to assist in determining if a tumor is malignant or benign. The cells in each sample are digitized, scanned, measured, and averaged. Our goal is to see if we can predict malignancy based on these values. This model

was initially published in a paper written by Street et al. (1993). The link to access the paper is https://minds.wisconsin.edu/bitstream/handle/1793/59692/TR1131.pdf;jsessionid=F231854E80A 00FB8803122FDF9847940?sequence=1.



*Image of cell nuclei in a sample*

The initial model was constrained by the computational power of the early 90s, we intend to use modern hardware and techniques to recreate and improve this model.

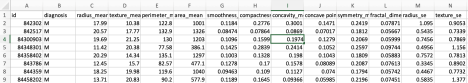
**Dataset**

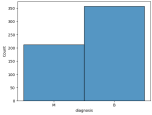
We will use the same UCI dataset WDBC as Street et al. (1993). The dataset is a benchmark dataset for breast cancer diagnosis tasks, which is still used in recent research projects (Vijayakumar et al., 2021). We have obtained the Breast Cancer Wisconsin (Diagnostic) data set from Kaggle: https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data/data

This data set contains statistical measurements of cell nuclei as obtained from a fine needle aspiration. 569 rows are included, and each observation represents a patient. The data set has a total of 32 columns: 30 input variables, 1 output variable (diagnosis column), and patient ID

(which will not be used in our task). All of the input variables are numeric. The dataset is clean with no missing values, outliers, or data type mismatches.

The prediction variable “diagnosis” has two values: B (benign) or M (malignant). Each observation contains measurements across 10 categories: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. The data set includes 3 values for each feature: the mean, the standard error, and the mean of the three largest cells. The below table is the data samples:

**Data Exploration** 

The dataset contains 357 (63%) benign and 212 (37%) malignant observations. 

Looking at the histogram, most of the data appears to be either normally distributed or right-skewed. The "Mean" column appears to be the most standardized, while the standard error and "Worst" values have the greatest deviation.

**Problem Statement and Solution Proposed**

Our primary question is whether we can determine if a tumor is malignant from the input variables. We will pursue a few other questions as well. Can we generate clusters? Does the accuracy of the model increase if we ignore the ‘worst’ cells, or are they good indicators? Is the standard error higher or lower when faced with a malignant diagnosis?

**Expected Results and Evaluation**

1. Model Performance Metrics:

Accuracy

Precision

Recall (Sensitivity)

F1-score

ROC-AUC: Area under the Receiver Operating Characteristic (ROC) curve, which measures the model's ability to discriminate between classes.

2. Feature Importance:

Insights into which input variables (features) are most important in predicting whether a tumor is malignant or benign. This information can help in understanding the underlying factors contributing to tumor classification.

The expected results would be a well-performing model with high accuracy, precision, recall, and F1-score, supported by strong feature importance analysis.

The data set is not tremendously large, but we believe we can still obtain a respectable split for training and testing data. We plan to try out several supervised machine learning models. The original paper was able to obtain 97.3% accuracy. We hope to obtain a similar accuracy.

**References:**

Street, W. N., Wolberg, W. H., & Mangasarian, O. L. (1993, July). Nuclear feature extraction for breast tumor diagnosis. In Biomedical image processing and biomedical visualization (Vol. 1905, pp. 861-870). SPIE.

Vijayakumar, K., Kadam, V. J., & Sharma, S. K. (2021). Breast cancer diagnosis using multiple activation deep neural network. Concurrent Engineering, 29(3), 275-284.