Identification of Abnormal Breast Growths as Potential Cancers Using Machine Learning

Team PCDT (Pre-Clinical Diagnosis and Testing)
Rishi Basdeo, Navodit Chandra, Pranav Katragadda
TA Adviser: Shivani Saboo

1. MOTIVATION

Breast cancer is one of the most common types of cancer, especially among women. Classified into malignant (able to spread throughout the body) or benign (non-spreading), the former is highly dangerous and requires early detection for the best chance of survival. To diagnose these growths, a minimally invasive procedure called fine needle aspiration is conducted, whereby an ultrathin needle is inserted into the growth to extract a cellular sample, which is then examined under a microscope. This procedure must be performed by a pathologist, potentially increasing the required amount of time for conclusive results. **[1]** A machine learning approach is a proposed improvement to current practices, as it may reduce the turnaround time for these diagnoses.

2. DATA SOURCE

The proposed machine learning models will be based on data generated by Dr. William H. Wolberg Et. Al at the University of Wisconsin. It consists of ten dimensions extracted from microscopic images of fine needle aspirate cell nuclei of 569 patients. Dimensions include radius, texture (based on variations in grey-scale values), and shape variations. **[2]**

3. MACHINE LEARNING

The proposed models for evaluation are classification-based techniques, such as logistic regression, k-nearest neighbors (k-NNs), decision trees, and Support vector machines (SVMs). These models are especially effective for smaller datasets, such as the one used in this proposal. To further prevent overfitting (a common challenge with small sets), feature engineering practices will be employed, such as experimenting with different combinations of features in the training data. Additionally, the number of samples in each category will be balanced, so as to eliminate any inherent biases in the model, while also experimenting with the effects of Synthetic Minority Oversampling Technique (SMOTE) to augment the minority class.

4. EVALUATION

The following performance measures will then be used to evaluate the results:[1]

1. **Accuracy**: the proportion of correctly identified data
2. **Precision**: the ratio of correctly labelled positive observations to the overall predicted positive observations.
3. **Recall**: the ratio of correctly predicted positive observations to the total positive observations
4. **Specificity**: the proportion of true negatives which were labelled correctly
5. **F1 Score**: the overall measure of precision and recall

These values will also be expressed graphically so as to better visualise their comparison and, further expressed with a confusion matrix which compares predicted and true classification values. Additionally, each model will be evaluated using 5-fold cross-validation, and their mean statistics will be reported.

6. REFERENCES

[1] Virtual Medical Centre (2017). Fine Needle Aspiration Biopsy (FNA).
[2] Wolbert, Et Al. (1995) Breast Cancer Wisconsin (Diagnostic) Data Set.
[3] Microsoft Azure Machine Learning Documentation. (2019) SMOTE

[1] TP: True Positive; TN: True Negative: FP: False Positive; FN; False Negative