Improving Breast Cancer Detection in Fine-Needle Aspiration Biopsies Through Machine Learning

Ivaturi, Dhanvee (School: Silver Creek High School)

Kabranov, Philip (School: Silver Creek High School)

This project evaluates multiple machine learning (ML) algorithms to determine optimal accuracy in detecting breast cancer. It also analyzes the effect of applying Principal Component Analysis (PCA) for feature reduction. Python is the implementation language, SciKit-Learn and Tensorflow are used for core ML algorithms, and Matplotlib is used for data visualization. The dataset consists of 569 entries of 30-dimensional feature vectors (describing the characteristics of cells in the tissue sample), obtained from the Wisconsin Breast Cancer Dataset. It is randomly split into test and training subsets, used to train and test the ML algorithm. A PCA-reduced dataset is also used to train the algorithms. Training time is also considered as an efficiency measurement. Visualizations show that the dataset is linearly separable, meaning that a logistic regressor (LR) and support vector machine (SVM) can be used. The LR, SVM, and neural networks were trained using the same dataset. For both data reduced to 5 features by PCA and for non-reduced data (30 features), LR produced optimal mean accuracy over approx. 20 training cycles. The accuracy for PCA-reduced and non-reduced data is comparable within a small margin. These accuracies resemble or exceed human ability. Our results show that the most accurate algorithm for this particular dataset is the logistic regressor, producing the highest accuracy with the shortest training time. Neural networks are on par with SVMs but take about 100 times longer to train. This is likely due to the computational complexity of backpropagation, unlike the simpler training for logistic regression. In addition, the relatively low performance of the SVM is likely due to the imperfect separation of the data into clusters, allowing inaccuracies.