CALIFORNIA STATE SCIENCE FAIR
2014 PROJECT SUMMARY

Name(s)

Dibya Jyoti Ghosh

Project Number

34399

Project Title

Implementing Novel Kernel Discriminant Analysis to Optimize Breast Cancer Diagnosis

Abstract

Objectives/Goals

The objective is to determine if pre-processing mammographic data with KDA (Kernel Discriminant Analysis) increases the accuracy of a machine-learning breast cancer diagnosis algorithm by increasing detections and removing false positives. This will result in the simultaneous advancement of the medical and machine learning fields.

Methods/Materials

Mammographic images of 50 normal and 50 cancerous patients were obtained from the DDSM. These files were processed, filtered, segmented, and distilled into a set of 11 points each. These points were then processed with Quadratic (QDA) and Linear Discriminant Analysis (LDA), and stored separately. Three different machine learning algorithms of varying efficiencies were set up and trained with subsections of the data, to analyze the effectivity of the Kernel Discriminant Analysis pre-processing technique over a range of data sizes and base training algorithms.

Results

Among the two methods tested, quadratic and linear, Quadratic Discriminant Analysis provides the greatest accuracy boost. Trends in results were uncannily similar between the three base machine learning algorithms, and followed the same general pattern. With small-scale training data sets (10-60 subjects) , QDA provides a 60% accuracy boost over the control group, and LDA a 30% increase. Under training data sets with size over 60, QDA provides comparable accuracy as the control, but LDA deteriorates and a 15% decrease in accuracy.

Conclusions/Discussion

The data collected shows that quadratic discriminant analysis is a viable pre-processing step for current laboratory breast cancer diagnostics. The algorithm has a marked increase in accuracy with smaller data sets, and adds low overhead, enough to justify implementation into diagnostic programs.

Summary Statement

I created a preprocessing algorithm for breast cancer detection programs that increases accuracy by mutating the mammographic data to better fit the algorithmic model.

Help Received

UCSF for providing mammographic data and suitable analysis methods. My dad for helping me debug my program at times