Conference Paper Volume 2883, Issue 1, AIP Conference Proceedings 2024

Measurement of data complexity using machine learning for breast cancer classification

According to statistics, breast cancer is one of the most frequent malignancies among women globally. Early detection can dramatically improve patients' prognoses and chances of survival. Detection using machine learning is highly proved as successful method in early detection. The purpose of this project is to analyze the complexity of data in machine learning using different set sample sizes of breast cancer samples. There are four phases in this project. Phase one is to collect data and pre-processes the data. For pre-processing, PCA is used as a data dimension reduction method. In phase two, data samples are fed into classifiers (SVM, NB, and PNN) for training. In phase three, the performance of the each breast cancer sample set is validated by measuring statistically method(Anova test) and the accuracy of each classifier (using the K-fold across-validation method). The highest accuracy is achieved in detecting breast cancer when using 1000 samples (98%) while the lowest accuracy is achieved in detecting breast cancer when using 200 samples (50%). Among SVM, NB, and PNN, NB outperforms others by achieving 98%. These results show the higher the number of data samples the better the detection of breast cancer by using the machine learning.