Paper ID:	BMEiCON2024-009
Paper Title:	A Simplified Machine Learning Model for Breast Cancer Classification using Reverse Engineering
Authors:	Wadcharin Pana; Suejit Pechprasarrn (College of Biomedical Engineering, Rangsit University)
Email:	suejit.p@rsu.ac.th

Abstract

According to the World Health Organization (WHO), approximately 2.3 million women were diagnosed with breast cancer in 2023, resulting in 670,000 deaths. Early detection and effective treatment are critical, as they improve the chances of successful outcomes. This study aims to detect breast cancer at an early stage using machine learning methods. This research utilizes the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, which contains data from 569 patients—212 with malignant tumors and 357 with benign tumors—to develop classification models. The dataset was divided into training (90%) and validation (10%) subsets. The highest accuracy of 97.6% was achieved using a Quadratic Support Vector Machine (SVM) model on the training data. Similarly, an accuracy of 100% was achieved on the test dataset using Quadratic SVM, Linear SVM, and Medium Gaussian SVM models. Here, Principal Component Analysis (PCA) was employed to reduce feature dimensionality and eliminate correlated information. Feature selection techniques were also used to rank the importance of clinical attributes in the WDBC dataset. The results identified four crucial features: 1) Area, 2) Area Standard Error, 3) Worst Parameter, and 4) Worst Area. Several feature selection algorithms, including Chi-squared (Chi2), ReliefF, ANOVA, and Kruskal-Wallis, also highlighted four key features: 1) Worst Perimeter, 2) Worst Radius, 3) Worst Area, and 4) Mean Concave Points. A comparative analysis was performed between PCA and feature selection algorithms, using the four selected predictors to train the Quadratic SVM model. Feature selection yielded a minor underfitting performance o only 0.3% between the training and test cases, while PCA showed a larger discrepancy of 1.7%. Additionally, reverse engineering was conducted using 20 images from the WDBC dataset (10 benign and 10 malignant), along with 20 additional images (10 benign and 10 malignant) from other sources. This process involved adjusting image sizes by rescaling the magnification of the objective lens to restore them to their appropriate pixel dimensions, as per the WDBC dataset. Feature extraction functions were then applied to capture the four key features. The simplified model, incorporating these reverse-engineered feature extraction functions, demonstrated effective classification of breast cancer images from fine needle aspiration samples of breast tumors.