Using Machine Learning Models and Principal Component Analysis to Identify Critical Diagnostic Variables for Breast Cancer Cells

Suejit Pechprasarn[1*], Ohmthong Wattanapermpool[2], Maninya Warunlawan[2], Pornchaya Homsud[2], and Thumpussorn Akarajarasroj[2]

[1]College of Biomedical Engineering, Rangsit University, Pathum Thani, Thailand
[2]Satriwithaya School, Wat Bowon Niwet, Phra Nakhon, Bangkok, Thailand
The authors have contributed equally.
*Corresponding author; E-mail: suejit.p@rsu.ac.th

Abstract

Breast cancer (BC) is now recognized as a disease that substantially affects mortality and morbidity that is on the rise and endemic throughout the world. This study utilizes a publicly available clinical dataset from the University of Wisconsin containing 699 patients and 9 parameters: (1) clump thickness, (2) uniformity of cell size, (3) uniformity of cell shape, (4) marginal adhesion, (5) single epithelial cell size, (6) bare nuclei, (7) bland chromatin, (8) normal nucleoli, and (9) mitoses. Here, we utilize this data to assure its objectivity and precision. We use machine learning algorithms and the analysis of principal components to determine the variables involved in identifying a tumor as malignant or benign. This study examines and analyses the classification accuracy of several machine learning models, such as tree models, discriminant models, logistic regression, Naive Bayes, support vector machine models, K-nearest neighbor models, ensemble models, and neural network models kernel models. The most accurate models are the coarse Gaussian SVM, the cosine KNN, and the medium Gaussian support vector machine, with a classification performance of 96.5%. The principal component analysis approach is then used to identify vital components and construct a precise model with fewer parameters. The medium Gaussian SVM has the best classification accuracy based on cross-validation at 96.98% and requires just 3 predictor variables: (1) normal nucleoli, (2) bare nuclei, and (3) cell size uniformity.

Keywords: *Breast Cancer Classification, Principal Component Analysis, Complexity-Reduced Model, Machine Learning*