MLKD 2024
The First International Conference on
Machine Learning and Knowledge Discovery
Amirkabir University of Technology, December 18-19, 2024

Improving Breast Cancer Prognosis via Multi-Gene Machine Learning Model

Fateme Azizi[*] Bahram Sadeghi Bigham[†] Mahboobe Zarrabi[‡]

Abstract

Breast cancer remains one of the most prevalent and lethal diseases among women, with challenges in treatment stemming from the biological and genetic heterogeneity of tumors. While prior studies have developed grading models using machine learning to improve prognostic accuracy, they reached a peak accuracy of 91%. This paper advances this work by employing a more extensive dataset and refined data selection methods, achieving an accuracy improvement to 92%. Gene expression datasets were collected from the Gene Expression Omnibus (GEO) repository, undergoing preprocessing, integration, and normalization, before being analyzed by the XGBoost algorithm to develop a predictive tumor grading model. Evaluation results show that our expanded dataset and modified biomarker panel of 70 markers contribute to enhanced grading accuracy, particularly in classifying grade 2 and indeterminate tumors, which are often challenging to diagnose and treat. This model underscores the effectiveness of combining expanded transcriptomic data with advanced machine learning techniques. Furthermore, it highlights key genes associated with prognosis, offering insights into potential biomarkers for future research and clinical applications.

Keywords: Gene Expression, Machine Learning, Cancer Prognosis

[*] Department of Computer Science, Faculty of Mathematical Sciences, Alzahra University, Tehran, Iran, azizif799@gmail.com

[†] Department of Computer Science, Faculty of Mathematical Sciences, b_sadeghi_b@iasbs.ac.ir

[‡] Department of Biotechnology, Faculty of Biological Sciences, Alzahra University, Tehran, Iran, mzarrabi@alzahra.ac.ir