



Classification of pCR and non-pCR in HR+/HER2+ Breast Cancer Subtype Using ADC Maps and Machine Learning Models

M.H.S. Kavindi¹, W.M.K. De Silva^{2}, M.L. Jayathilake³, A.D.S. S. Sanjeeewa⁴,
B.S. Weerakoon³, S. M. Vijithananda⁵, P.D.S. H. Upul⁶*

¹ Department of Physics, University of Ruhuna

² Department of Physics, Faculty of Science, University of Ruhuna

³ Department of Radiography/Radiotherapy, Faculty of Allied Health Sciences,
University of Peradeniya

⁴ Faculty of Medicine, University of Ruhuna

⁵ Department of Radiology, Faculty of Medicine, University of Peradeniya

⁶ Department of Physics, National Cancer Institute, Maharagama

**Corresponding author: sandunikakavindi07@gmail.com*

Pathological complete response (pCR) following neoadjuvant chemotherapy is a key prognostic indicator in HR+/HER2+ (triple-positive) breast cancer. Early non-invasive prediction of treatment response could help optimize therapeutic strategies and minimize exposure to toxic treatments. This study aimed to predict pCR using radiomics features extracted from apparent diffusion coefficient (ADC) maps derived from diffusion-weighted MRI (DWI) trace images acquired at the early treatment stage (T0) from The Cancer Imaging Archive (TCIA). The dataset included 215 ADC maps from 53 HR+/HER2+ patients, comprising 78 pCR and 137 non-pCR cases. ADC maps were generated by combining DWI trace images with b-values of 0 and 800 s/mm². The dataset was split into 70% training and 30% testing sets. Tumor regions of interest were manually delineated, and first-order statistical, textural, and morphological radiomics features were extracted. Correlation-based feature selection was applied to the training set, removing highly correlated features with correlation >0.9. Class imbalance was addressed using the SMOTE technique to upsample the minority class while minimizing overfitting. Several supervised machine learning classifiers were evaluated. Random Forest achieved the highest test accuracy (67.69%), followed by Decision Tree (61.54%), LightGBM (61.54%), Logistic Regression (56.92%), Support Vector Machine (56.92%), XGBoost (58.46%), and Gradient Boosting (47.69%). Five-fold cross-validation of Random Forest yielded 64.65% accuracy, with an area under the ROC curve of 0.54. The optimal Random Forest parameters were: n_estimators = 800, criterion = 'entropy', max_depth = 5, min_samples_split = 10, min_samples_leaf = 8, max_features = 0.05, bootstrap = True, random_state = 42. Limitations of this study include its retrospective multicenter design and small sample size. These findings should be validated in larger cohorts. Nevertheless, this study provides useful insights for predicting pCR in clinical practice, offering potential to guide personalized treatment decisions in HR+/HER2+ breast cancer.

Keywords: Magnetic resonance imaging, Apparent diffusion coefficient, Diffusion-weighted imaging, pCR and non-pCR, HR+/HER2+ (triple-positive breast cancer subtype), Random Forest