CLINICAL DATA PREPARATION AND FEATURE ENGINEERING FOR MACHINE LEARNING CLASSIFICATION OF BREAST CANCER CASES AND COLORECTAL CANCER CONTROLS IN SRI LANKA

K.M.S.S.Muhandiramge[1*], J.H.D.S.P.Tissera[1], H.S.Karunaratne[1], S.A.T. Chaturanga[2], B.M. Rathnavibushana[2]

[1] *Department of Statistics, University of Colombo, Colombo, Sri Lanka*

[2] *National Cancer Institute, Maharagama, Sri Lanka*

*Corresponding author: sandunikasmuhandiramge@gmail.com

Abstract

Real-world clinical records contain missing values, mixed variable types and correlated predictors, requiring careful preparation before machine learning. This study examines preparation challenges and feature-engineering decisions in a newly assembled Sri Lankan hospital-based dataset. A retrospective case-control study used 500 female Bed Head Ticket records from Apeksha Hospital, Maharagama (2022–2025): 321 breast cancer cases and 179 colorectal cancer controls. Colorectal cancer controls were used because healthy controls were not readily available; this may introduce selection and spectrum bias, limiting interpretation to distinguishing these two cancer groups. Variables included demographic, reproductive, tumour-stage and biological characteristics. Missing values were handled using Multiple Imputation by Chained Equations (MICE). Mixed-type associations were assessed using Pearson's correlation, Cramér's V and point-biserial correlation. Because age and menopausal status were strongly associated, an Age_Meno_Group composite was created to reduce redundancy before modelling. Seven classifiers were compared using an 80:20 train-test split, with tuning and cross-validation on training data. XGBoost achieved the highest test accuracy (0.81) and F1-score (0.86). SVM showed more consistent performance, with training and testing accuracy of 0.79 and 0.78, respectively, and F1-scores of 0.85 on both sets. SHAP identified T2 and T3 tumour stages, Ki-67 index and N2 stage as important contributors to classification. The timing of tumour-stage and Ki-67 measurements relative to diagnosis requires verification; therefore, the model should not be interpreted as a pre-diagnostic detection tool. Further evaluation using additional performance metrics and uncertainty estimates, and validation in larger multi-centre datasets including healthy controls, are needed. The study highlights the importance of systematic clinical data preparation and clinically informed feature engineering before modelling.

Keywords: Clinical Data Preparation, Feature Engineering, Breast Cancer, Machine Learning, Sri Lanka