Accelerating Breast Cancer Diagnoses: Leveraging Machine Learning to Accurately Predict Breast Cancer Presence and Growth With Hormonal Data

Kim, Joshua (School: McCallie School)

Mammography, currently the primary method implemented for breast cancer screening/detection, often produces false negatives (1 in 8). Additionally, there exist prolonged delays during diagnosis (median 61 days from first symptoms to diagnosis). These two setbacks allow for cancers to spread, enabling a Stage 1 tumor with a 99% 5-year survival rate to quickly metastasize into Stage 4 cancer with a 30% 5-year survival rate. The underlying objective of this research was to determine the feasibility of developing a more accurate, alternative detection method: a machine-learning model able to predict tumor status (tumor presence binary prediction) and stage (per AJCC guidelines) of breast cancer using hormonal data. Implementing patient case datasets from the NCI GDC Data Portal, the researcher developed models from various machine-learning algorithms, of which the Random Forest Classifier (RFC) and the Histogram-based Gradient Boosting Classifier (HGBC) performed best, each displaying 96.00% tumor status prediction accuracy. Compared to 1 out of 8 cancer-containing mammograms that produce false negatives, the RFC model displayed only 1 out of 67 false negatives. The models could not as effectively predict tumor stage, however, with the best model (RFC) displaying a mere 66.03% accuracy. The results not only provide novel insight into the correlation of female-specific hormones with the metastasis of breast cancer but also offer significant evidence for a more accurate, alternative method for breast cancer detection. The researcher anticipates in vitro validations to test the machine-learning models' applicability to real cancer tissue, marking a significant step toward clinical implementation.