**Machine Learning Analysis of Breast Cancer Expression Profiles**

**Introduction**

### Breast cancer can develop at any age after puberty, with higher incidence rates in later life. In 2022, around 2.3 million women were diagnosed with breast cancer, resulting in 670,000 deaths globally (WHO, 2024). The integration of big data and AI in biomarker discovery has enhanced the potential for personalised medicine (Restrepo *et al.*, 2023). In this analysis, machine learning techniques were used to identify key genetic biomarkers that differentiate tumour from normal tissues in breast cancer patients.

**Methodology**

**Preprocessing of Dataset for Machine Learning Analysis**

The dataset consisted of twenty gene expression profiles from normal and breast cancer samples, accompanied by metadata on tissue type (tumour or normal). Gene expression data was aligned with metadata by matching sample IDs. The data was transposed so that genes became features, and the top 1000 most variable genes were selected based on standard deviation.

Preprocessing steps included removing near-zero variance features to reduce noise, centering the data to standardise values, and removing highly correlated features (correlation > 0.5) to mitigate multicollinearity.

**Machine Learning Analysis**

A Random Forest model was implemented using the ranger package. The dataset was split into a training set (80%) and a testing set (20%) for model development and evaluation. A 10-fold cross-validation was used to ensure model reliability and prevent overfitting. Key parameters, such as the number of predictors, minimum node size, and Gini impurity, were optimised through grid search. Feature importance was assessed using permutation-based methods to identify significant biomarkers.

### 

**Result and Model Performance**

The Random Forest model showed a low prediction error during training, indicating good performance in classifying tissue types. This plot (figure 1), highlights the top genes that contribute the most to the Random Forest model's classification. Genes like TGFBR3 and COL10A1 are highly influential in distinguishing between tumor and normal tissues. The top 10 most important genes were visualised to effectively communicate the findings guiding further biological research or treatment strategies.

**Conclusion and Future Directions**

This machine learning analysis identified the genes that could potentially be biomarkers for breast cancer using gene expression data. These identified genes can be used as critical biomarkers for early detection and targeted therapies in breast cancer. Integration of clinical data with other machine learning models could be further applied in the future to improve the predictive accuracy and develop a deeper insight into the biology of breast cancer.

**References**

HACKBIO. Machine Learning in Genomics Course. 2024. <https://thehackbio.com/courses/60>

Restrepo, J. C., Dueñas, D., Corredor, Z., & Liscano, Y. (2023). Advances in Genomic Data and Biomarkers: Revolutionizing NSCLC Diagnosis and Treatment. *Cancers*, *15*(13), 3474. <https://doi.org/10.3390/cancers15133474>

World Health Organization (WHO). 2024. Breast Cancer. Retrieved 25th September 2024. <https://www.who.int/news-room/fact-sheets/detail/breast-cancer>