Multi-Omic Integration for Breast Cancer Subtype Classification Using Advanced AI Techniques

Sajib Acharjee Dip, Liqing Zhang
Department of Computer Science
Virginia Tech, USA

Abstract

Breast cancer is a highly heterogeneous disease with diverse molecular subtypes that exhibit distinct clinical behaviors and treatment responses. Current classification methods often rely on single-omic data, such as gene expression, failing to capture the full biological complexity. In this paper, we propose a multi-omic integration framework for breast cancer subtype classification using gene expression, DNA methylation, and miRNA expression data. We employ a novel cross-attention neural network combined with contrastive loss (SimCLR) and autoencoders to align and fuse the omic features. Our method achieves state-of-the-art classification performance and demonstrates the potential for discovering novel subtypes of breast cancer. This work contributes to precision oncology by offering a more comprehensive and interpretable classification model for breast cancer subtypes.

Breast cancer, with its diverse molecular and clinical profiles, is one of the most studied cancers in the world. Yet, despite significant advances, the current methods for breast cancer subtype classification, typically based on single-omic data, fail to fully address the heterogeneity of the disease. Different subtypes, such as Luminal A, Luminal B, and HER2-enriched, exhibit varying responses to treatment, necessitating the need for more accurate, comprehensive, and interpretable models.

Integrating multi-omic data for breast cancer subtyping presents several challenges due to the inherent heterogeneity of the data, where gene expression, DNA methylation, and miRNA expression come from distinct biological processes and are measured at varying scales, making it difficult to combine them effectively while preserving their biological significance. Additionally, large-scale cancer datasets like TCGA often suffer from missing or noisy data, which complicates model training and can lead to inaccurate or unreliable predictions if not properly addressed. Another significant challenge is the interpretability of machine learning models, especially deep learning models, which often operate as "black boxes," making it difficult to understand how different omic layers contribute to the final classification. This lack of transparency is a barrier to scientific validation and clinical application. Finally, capturing the cross-relationships between omic layers—such as how DNA methylation might influence gene expression or how miRNA regulates both—requires advanced techniques for feature alignment and fusion, making the integration of multi-omic data a complex and computationally intensive task.

Existing classification methods often rely on gene expression data alone or, at best, integrate a limited number of omic layers, missing the cross-relationships between these biological levels. The integration of multi-omic data—combining gene expression, DNA methylation, and miRNA expression—offers a more holistic view of the disease but introduces significant computational and methodological challenges. This paper positions the use of advanced machine learning models, particularly multi-omic attention neural networks, as a promising approach to overcome these barriers and improve the predictive accuracy of breast cancer subtype classification.

To address these challenges, this work introduces a multi-omic classification framework for breast cancer using a cross-attention neural network architecture. The key innovations of this model are:

Data Alignment Using SimCLR Contrastive Loss: The model employs SimCLR, a contrastive loss function, to align the multi-omic features (gene expression, DNA methylation, miRNA) into a common representation space. This ensures that the features from different omics layers are meaningfully aligned, allowing the model to learn cross-relationships between them.

Cross-Attention for Data Fusion: The model utilizes cross-attention mechanisms, an extension of the self-attention mechanism, to fuse multi-omic data. This technique learns relationships across different modalities at multiple scales, enabling the model to capture both global and local dependencies across the omics layers.

Enhanced Prediction Accuracy: The multi-omic fusion framework significantly improves classification performance, achieving high accuracy (87%), precision (86%), recall (87%), and F1-score (86%), while maintaining balanced accuracy (80%) on the test dataset.