Genetic Feature Detection on Breast Cancer Subtypes

By: Melika Nassizadeh and Diego Mallen

Video Presentation: Panopto
Slides: Google Slides

Code Base: Google Collab

Introduction

Breast cancer is one of the most common cancers among women, accounting for about one in four cancer cases worldwide. According to the World Health Organization, in 2020, more than 2.3 million women were diagnosed [1]. Age is a factor in susceptibility, and patients are most often over 50 years old. However, breast cancer can affect young individuals, especially women with a genetic predisposition or family history. Other risk factors include genetic mutations such as BRCA1 and BRCA2, hormonal influences, lifestyle choices such as alcohol consumption and obesity [2].

The complexity of breast cancer makes the illness difficult to prevent or pinpoint within a population. Caused by gene mutations, it presents itself through distinct, classified subtypes. Basal-like, HER2-positive, luminal A, and luminal B are labeled according to molecular characteristics, including gene expression profiles [3]. Treatment varies based on the specific diagnosis and stage of breast cancer, scaled from zero to four. It can include surgery, radiation therapy, chemotherapy, hormone therapy, and targeted biological therapies such as trastuzumab for HER2-positive tumors [4].

The diagnosis procedure traditionally starts with a physical examination for lumps. Women older than the age of 40 are recommended to take an image called a mammogram every two years. If concerns are present, the procedure can follow ultrasound, MRI, or tissue biopsy for histopathological evaluation [5]. More recently, physicians can use molecular diagnostic tests like Oncotype DX, MammaPrint, and PAM50, which analyze expression patterns in tumor tissue. Using these tests clarifies the risk of recurrence and the benefit of chemotherapy. Physicians can have better precision in diagnosis and personalize treatment plans for patients.[6]. Preventive strategies emphasize identifying individuals with a high genetic risk. Genetic counseling and testing for hereditary mutations allow for earlier awareness and preventive measures, such as prophylactic mastectomy or chemoprevention [2]. On the other hand, emerging research in gene therapy directly corrects or compensates for genetic defects at the molecular level [7]. With a large quantity of patient data and precise AI models, medical professionals can detect breast cancer significantly earlier, enabling patients with higher-quality treatment options [8].

Previous Works

In 20211, researchers from the University of North Carolina at Chapel Hill, led by Prat and Perou [3], analyzed the molecular subtypes of breast cancer by collecting gene expression profiles from 1,877 breast cancer tumor samples. The combined public datasets included The Cancer Genome Atlas (TCGA), a comprehensive project compiling genetic and clinical data from thousands of cancer patients. They confirmed the existence of intrinsic subtypes, including basal-like, HER2-enriched, luminal A, and luminal B tumors, using hierarchical clustering and supervised classifications. The best prognosis is Luminal A cancer, which is typically hormone receptor-positive and slow to grow. In contrast, luminal B tumors can express higher levels of the protein HER2, which promotes cancer cell growth. The limitations of the research are hinted at in the reliance on batch-corrected public datasets, which can introduce variability from direct clinical samples. Batch correction, while necessary to standardize data across studies, can inadvertently remove true biological differences or create artificial patterns. Often, machine learning models trained on such data perform poorly when applied to real clinical samples.

In 2006, researchers at the University of Alberta, Cruz and Wishart [9] published a comparison review of machine learning models utilized for cancer prediction and prognosis. Summarizing over 100 studies, the paper emphasized that models such as decision trees, support vector machines (SVMs), and ensemble methods consistently outperformed traditional clinical factors when analyzing large-scale gene expression datasets, often based on hundreds to thousands of patients. While the review encompassed multitudes of cancers, they highlighted a key challenge common in early studies: the risk of overfitting, where models trained on small sample sizes relative to the high variability of gene features may fail to generalize to new patient data.

More recently, researchers from Sun Yat-sen University in China, led by Jiang et al. [10], applied deep learning methods to breast cancer data, combining both gene expression profiles and histopathological images. Using over 1,000 samples from TCGA, they demonstrated that convolutional neural networks (CNNs) could outperform traditional SVMs in predicting breast cancer outcomes and subtypes. However, one limitation of their approach was the reduced interpretability of deep learning models, often called "black-box" behavior. Deep neural networks, while powerful, make it difficult to trace exactly how specific genes or image features influence predictions. In clinical contexts, where explainability is critical for physician trust and regulatory approval, this lack of transparency poses a major barrier to adoption.

In summary, these studies show that previous works on machine learning applied to gene expression and clinical imaging data. They highlight the assistance of models in classification, despite challenges with overfitting, dataset variability, and model interpretability.

Research Question and Hypothesis

Using a supervised machine learning model, we aim to mimic gene expression feature detection for six different types of breast cancer: basal-like, HER2-positive, luminal A, luminal B, normal, and cell line. The gene activity levels are distributed among 152 columns. The hypothesis is that we can discreetly associate the genes most likely to upregulate or downregulate towards each breast cancer subtype. Not only will the neural network model provide insights into key genetic markers within this dataset, but it will also highlight the significance of using machine learning techniques in medicine for preventive measures.

Methodology

We used the CUMIDA Breast Cancer Gene Expression dataset [11], which contains gene expression profiles for breast cancer samples labeled by subtype. The project aimed to train a supervised machine learning model to classify samples into one of six subtypes: basal-like, HER2-enriched, luminal A, luminal B, normal tissue, and cell line.

First, we loaded the dataset using Pandas and filtered the feature set to include only numeric gene expression columns. The sample identifiers were dropped, and the type column, which contained the cancer subtype labels, was separated as the target variable. We normalized the gene expression scale to a range between 0 and 1 using a Min-Max method. The preprocessing step ensures that data with a large standard deviation does not lose certain features during the neural network calculations.

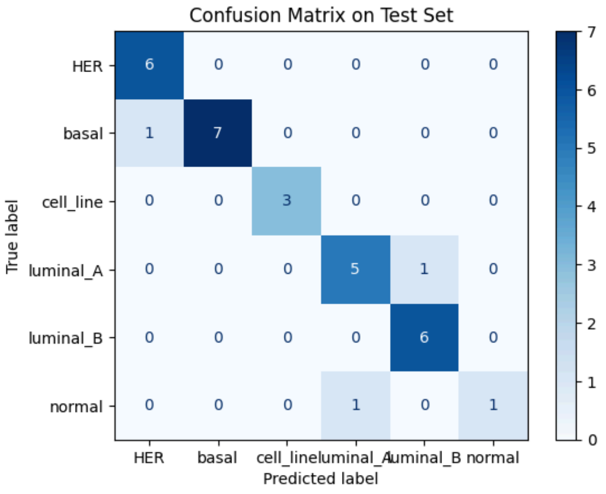
The dataset was split into training and testing sets using an 80/20 ratio to evaluate the model’s generalization performance. We implemented a simple fully connected neural network using TensorFlow/Keras with three layers. The input layer corresponds to the number of gene features, and is followed by one hidden layer with 128 ReLU-activated neurons. The output layer has six neurons activated by a softmax function, corresponding to the six cancer subtypes.

We compiled the model using the AdamW optimizer and categorical cross-entropy loss, appropriate for multi-class classification tasks. Training was conducted over 50 epochs with a batch size of 32. Model performance was evaluated on the held-out test set using accuracy as the primary metric. Throughout the process, categorical labels were one-hot encoded to match the output format of the neural network.

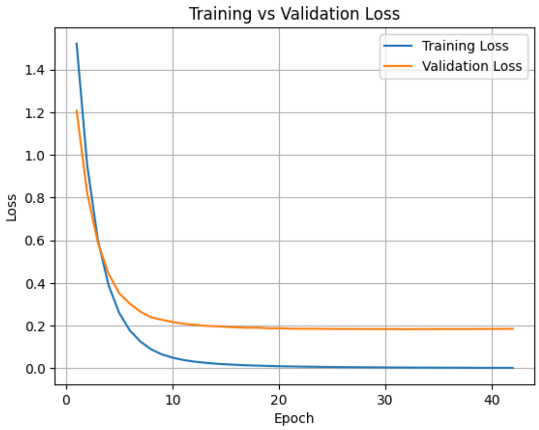
This methodology allowed us to efficiently preprocess high-dimensional gene expression data and apply a deep learning approach to classify breast cancer subtypes based on distinct molecular signatures.

Results and Evaluation

The neural network model resulted in a breast cancer classification with 90.32% accuracy, suggesting that machine learning can distinguish molecular signatures associated with each breast cancer subtype.

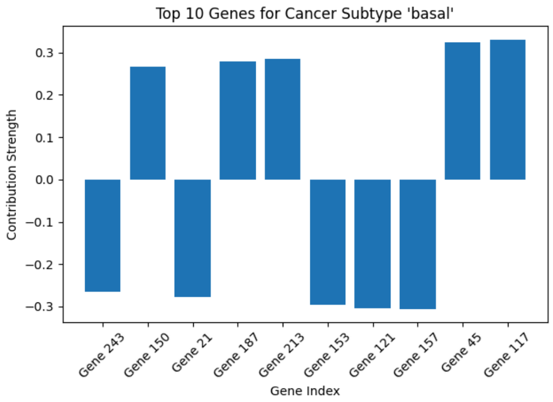


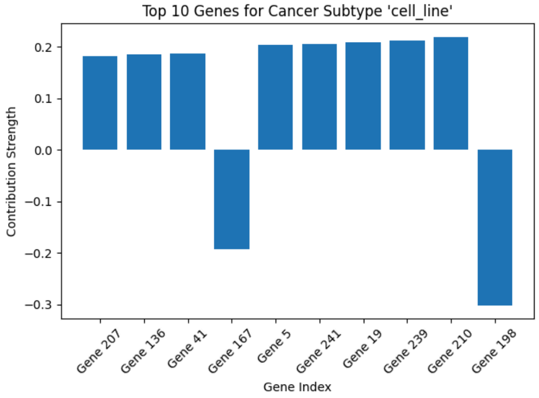
A highly accurate confusion matrix with all six subtypes, HER2, basal-like, cell line, luminal A, luminal B, and normal tissue, is shown to the right. HER2 and Basal-like samples showed only one incorrectly overlapping classification. Cell line samples were correctly identified in all cases, although only three were labeled. Luminal A displayed slight overlap with luminal B, reflecting biological similarity between these subtypes. Normal samples exhibited minor confusion with luminal subtypes, likely due to similar baseline gene expression patterns. In conclusion, misclassification rates are low and occur between biologically related subtypes.

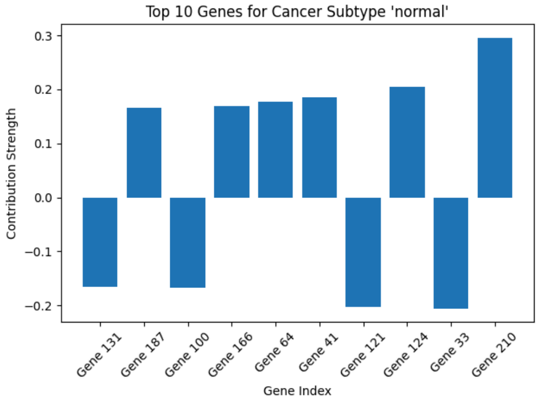


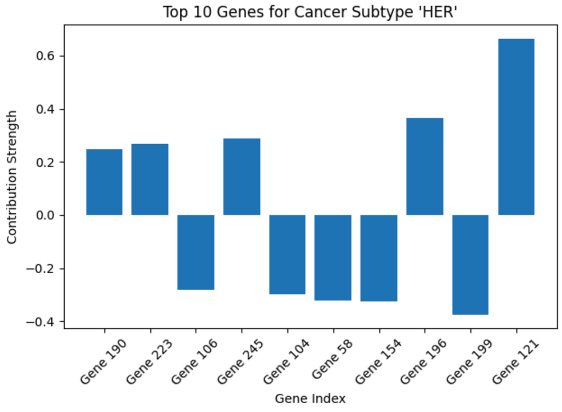
The training and validation loss curves figure above shows the model rapidly decreasing within the first 10 epochs, then stabilizing with less than 0.3 loss. Since the testing data plateaued close to the training loss, the model was not overfitted and can be applied to unseen data. Verifying the size of the dataset, the learning pattern also suggests that the network layers are appropriate for the complexity of the data.

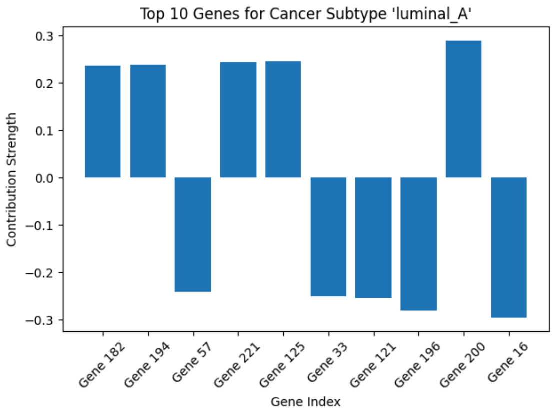
Using feature detection, we showcased the most influential genes for each cancer subtype. Each gene was correlated with a positive or negative numeric contribution strength; both are shown in the bar graphs. Intuitively, positive values highlight the promotion of classification, whereas negative contributions show which genes are tumor suppressors for the subtype.

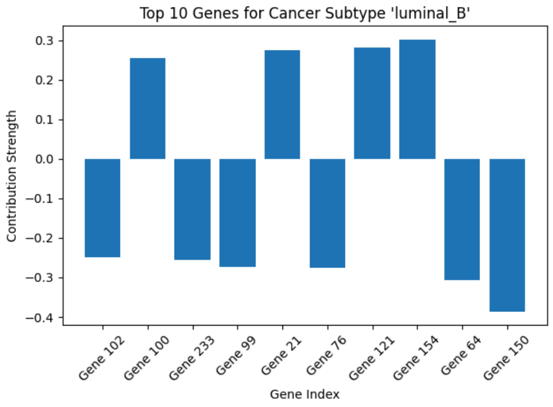












For HER2 subtype predictions, Gene 121 showed the strongest positive contribution, while Gene 199 had a strong negative contribution, suggesting that elevated expression of Gene 121 serves as a key marker for HER2-positive cancer, while suppression of Gene 199 reinforces accurate subtype discrimination. In basal-like tumors, genes such as Gene 45 and Gene 117 positively influenced predictions, whereas Gene 243 exerted a negative effect, indicating that the model balances activators and suppressors to differentiate basal subtypes. For cell line samples, positive contributions from Gene 207 and Gene 136 and negative influence from Gene 198 likely reflect molecular adaptations unique to cultured cell lines compared to primary tumors. Luminal A and luminal B subtypes displayed distinct yet overlapping influential genes, with Gene 121 again emerging as an important feature across both classes, highlighting known biological similarities between these luminal subtypes. Finally, in normal tissue classification, Gene 210 contributed positively, while Gene 131 and Gene 100 negatively influenced predictions, suggesting that the regulation of these genes plays an important role in maintaining normal cellular function and distinguishing healthy tissue from malignancy. Together, these patterns validate the model's ability to not only achieve high classification accuracy but also reveal biologically meaningful gene regulation patterns across different breast cancer subtypes.

Discussion and Future Work

This study demonstrated that gene expression profiles can accurately predict breast cancer subtypes through a deep learning model. Achieving a test accuracy of 90.32%, the model successfully distinguished between HER2, basal-like, luminal A, luminal B, cell line, and normal tissue samples. Gene contribution analysis revealed biologically meaningful patterns, identifying genes whose regulation either promoted or suppressed subtype classification. This interpretability strengthens the potential clinical relevance of AI models by linking predictions to underlying biological mechanisms.

Several limitations suggest directions for future improvement. The model was trained on a single dataset without external validation, limiting conclusions about generalizability. Evaluating the model on independent cohorts would provide stronger evidence of clinical utility. Additionally, the inclusion of all available genes as input features may have introduced noise; applying feature selection techniques such as LASSO regression could enhance both accuracy and interpretability. While the basic neural network architecture captured key gene patterns, future models incorporating graph neural networks or attention mechanisms may better model complex gene-gene relationships. Furthermore, integrating more advanced interpretability methods such as SHAP values or Integrated Gradients could offer finer-grained explanations for model predictions. Finally, expanding the analysis to incorporate multi-omics data, which combines different biological layers such as gene expression, protein levels, and epigenetic modifications, would provide a more comprehensive understanding of breast cancer subtypes. In particular, proteomics, the study of protein abundance and activity, could reveal functional changes that gene expression data alone may not capture.

In conclusion, this work highlights the promise of AI-driven gene expression analysis for breast cancer subtype classification. Future research focusing on dataset expansion, feature refinement, model complexity, and integration of multi-omics data will be crucial for advancing the clinical impact of this approach.

References

[1] World Health Organization, “Breast cancer,” World Health Organization, 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/breast-cancer.

[2] Centers for Disease Control and Prevention, “Breast Cancer Risk Factors,” CDC, 2021. [Online]. Available: https://www.cdc.gov/cancer/breast/basic_info/risk_factors.htm.

[3] A. Prat and C. M. Perou, “Deconstructing the molecular portraits of breast cancer,” *Molecular Oncology*, vol. 5, no. 1, pp. 5–23, 2011. doi: https://doi.org/10.1016/j.molonc.2010.11.003.

[4] D. Slamon, W. Godolphin, L. A. Jones, et al., “Studies of the HER-2/neu proto-oncogene in human breast and ovarian cancer,” *Science*, vol. 244, no. 4905, pp. 707–712, 1989. doi: https://doi.org/10.1126/science.2470152.

[5] S. S. D’Orsi, E. A. Sickles, E. B. Mendelson, et al., “ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System,” American College of Radiology, 2013.

[6] S. Paik, S. Shak, G. Tang, et al., “A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer,” *New England Journal of Medicine*, vol. 351, no. 27, pp. 2817–2826, 2004. doi: https://doi.org/10.1056/NEJMoa041588.

[7] Y. Li and H. L. Zhang, “Gene therapy targeting breast cancer: progress and challenges,” *Frontiers in Oncology*, vol. 11, 2021. doi: https://doi.org/10.3389/fonc.2021.768796.

[8] T. D. Nguyen, M. W. Nguyen, and S. Ravi, “Artificial intelligence in breast cancer prognosis: A review,” *Cancers*, vol. 13, no. 15, p. 3896, 2021. doi: https://doi.org/10.3390/cancers13153896.

[9] J. A. Cruz and D. S. Wishart, “Applications of machine learning in cancer prediction and prognosis,” *Cancer Informatics*, vol. 2, pp. 59–77, 2006. doi: https://doi.org/10.1177/117693510600200030 .

[10] Y. Jiang, C. Chen, S. Li, X. Xie, Y. Wang, and Y. Hu, “Breast cancer histopathological image classification using deep neural networks,” *International Journal of Computational Intelligence Systems*, vol. 10, no. 1, pp. 812–820, 2017. doi: https://doi.org/10.2991/ijcis.2017.10.1.54.

[11] B. Grisci, “Breast Cancer Gene Expression (CUMIDA),” Kaggle, 2021. [Online]. Available: https://www.kaggle.com/datasets/brunogrisci/breast-cancer-gene-expression-cumida.