Biotecnika Info Labs

Department of Bioinformatics, AI & Machine Learning

Machine Learning and Bioinformatics Framework for Breast Cancer Subtype Classification and Biomarker Discovery

By

Prashansha Goel

A report submitted in partial fulfilment of the summer internship program in Bioinformatics, AI and Machine Learning organized by Biotecnika

Supervised by

Dr. Nilofer Shaikh

ABSTRACT

The molecular intricacy of breast cancer makes prompt diagnosis and personalized treatment extremely difficult. In order to find important indicators for breast cancer subtyping, we used both bioinformatics and machine learning techniques in this study to methodically examine gene expression data from two GEO microarray datasets. Following the discovery of a group of common and subtype-specific genes by differential expression analysis, the building of a protein-protein interaction network highlighted key hub genes implicated in pathways relevant to cancer. The clinical significance of several hub genes was highlighted by enrichment and survival analysis. 35 potential biomarkers were prioritized through machine learning methods, particularly Random Forest, which performed well in subtype classification. These discoveries offer intriguing clues for upcoming advancements in diagnosis and treatment. The limited sample size of the examined datasets may restrict the generalizability of the findings, even though this integrative approach provides novel insights into breast cancer molecular subtyping and biomarker identification. Confirming the robustness and therapeutic value of the identified biomarkers will require additional validation in larger, independent cohorts and with different types of data.

Keywords: Breast cancer , Biomarker discovery, Machine learning, Subtype classification, Functional enrichment

1. INTRODUCTION

Breast cancer is the most frequent cancer among women globally; in 2022, there were an estimated 2.3 million new cases and 670,000 deaths . Breast cancer prevalence continues to rise globally, especially in developing nations where access to prompt diagnosis and care remains limited. In contrast, improvements in early detection and treatment have increased survival rate in high-income countries. There are still notable regional differences: whereas one in eight American women may receive a breast cancer diagnosis at some point in their lives, the disease is spreading most quickly in regions of Asia, Africa, and South America. Although confined breast cancer has a 99% five-year survival rate, late-stage diagnosis still

results in a high death rate, therefore early detection is essential [1-4]. There are several molecular subtypes of breast cancer, each with unique morphological and clinical traits, such as Basal-like, Her2-enriched, Luminal A, Luminal B, and Normal-like. The present understanding of BRCA heterogeneity was shaped by Sørlie et al.'s initial classification of these subgroups based on their distinct gene expression profiles [5]. This increasing global burden emphasizes how urgently trustworthy early-stage biomarker research is needed to enhance diagnosis and outcomes for women worldwide.

The identification of breast cancer biomarkers and the classification of cancer subtypes have been substantially improved by recent developments in bioinformatics and artificial intelligence. Researchers used one method to identify 733 differentially expressed genes (DEGs) through examining gene expression data from two GEO breast cancer datasets. They then used enrichment analysis and the creation of protein-protein interaction networks to identify 10 important hub genes, including CDCA8, MELK, and BIRC5, that are closely associated with the onset and progression of breast cancer. Through survival analysis, these hub genes were further confirmed, highlighting their potential as early diagnostic indicators [6]. Another recent study found weighted DEGs for each subtype of breast cancer by combining differential expression and gene regulatory network analysis. The researchers discovered biological activities and connections specific to subtypes by building gene co-expression networks and creating a novel GO enrichment approach (GOEGCN). They used these weighted DEGs to construct binary classifiers, leveraging artificial intelligence, and were able to differentiate BRCA subtypes with high accuracy [7]. Together, these approaches demonstrate the effectiveness of AI-driven machine learning models and advanced bioinformatics analyses for identifying breast cancer subtypes and discovering reliable biomarkers.

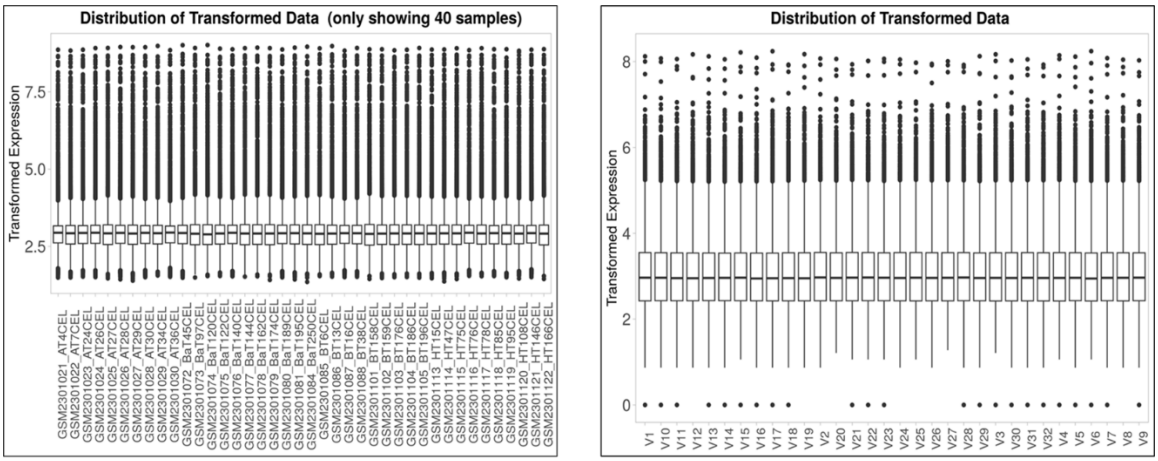
By combining several GEO microarray datasets, this effort seeks to find strong differentially expressed genes linked to breast cancer subtypes, building on previous bioinformatics and AI-based research. It uses protein-protein interaction network analysis to find biologically important hub genes then functional enrichment and survival analyses to confirm their clinical significance. To find distinct molecular signatures, subtype-specific differential expression analysis will also be carried out. Furthermore, in order to precisely identify breast cancer

subtypes and rank potential biomarkers, machine learning classifiers will be created and assessed. The goal of this integrative approach is to help patients with breast cancer receive specific treatment plans and assist early diagnosis.

2. METHDOLOGY

2.1 Comprehensive Data Retrieval and Preprocessing

In order to ensure thorough coverage, we used appropriate keywords to conduct a systematic search of the GEO database for microarray datasets including information on breast cancer subtypes during the data collection phase. GSE86374, which contains 10 normal and 50 tumor samples, and GSE57297, which contains 7 normal and 25 tumor samples, were the two datasets chosen for study. R software was used to process the raw data from the datasets. To assure comparability and data quality across samples, quantile normalization was carried out on GSE57297 and RMA normalization was applied to the GSE86374 dataset (Figure 1). This consistent preprocessing created a strong basis for further analysis.



GSE86374 GSE57297

In Figure 1: Normalization box plots for the GEO ids, (here v denotes sample id).

2.2 Differential Gene Expression Profiling

We employed R's Limma package for differential expression analysis (DEA), which is well-known for its strong statistical foundation and applicability for evaluating microarray data. To find important genes that differentiated between normal and tumor samples for the

bioinformatics-based biomarker discovery approach, DEA was carried out independently on each dataset. To identify subtype-specific genes for further machine learning modeling, we also performed subtype-associated DEA for each BRCA subtype in both datasets using a one-vs-rest method. The foundation of subsequent biomarker prioritization and classification tasks was laid by the dual analysis, which secured thorough detection of both general and subtype-specific DEGs (differentially expressed genes).

Common significant general DEGs were considered for further network analysis, whereas subtype-specific significant DEGs were combined for pre-processing of data to perform ML modeling [for significant genes, adj.P.Val<0.05].

2.3 Protein-Protein Interaction Network Construction

Using the collection of common significant DEGs found in both datasets, we built a protein-protein interaction (PPI) network. The PPI network was assembled and visualized using the STRING database, which integrates both predicted and experimental protein interactions. We identified hub genes using network topology metrics like node degree and centrality after exporting the resultant network to the Cytoscape software for additional analysis. These hub genes were then chosen for survival and downstream functional enrichment analysis to assess their potential as important breast cancer biomarkers.

2.4 Integrated Functional and Prognostic Analysis of Hub Genes

Enrichr was used to functionally enrich hub genes, and TNM plotter was used to examine gene expression patterns. To evaluate prognostic value, survival studies were performed using KM Plotter and GEPIA2. Furthermore, gene modification frequencies were assessed using cBioPortal, which shed light on their clinical significance in breast cancer. A comprehensive assessment of the putative biomarkers functional prognostic, and genetic importance was made possible by this multi-database approach.

2.5 Development and Evaluation of Machine Learning Models

We used two supervised machine learning algorithms, K-Nearest Neighbors (KNN) and Random Forest (RF), to classify breast cancer subtypes.

In KNN, a sample is assigned to the majority class among its k nearest neighbors.

Euclidean distance is used to determine the feature space's k nearest neighbors :

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

The Random Forest Classifier is an ensemble of decision trees; each trained on a bootstrap sample and split using a random subset of features. The final prediction is determined by majority vote across all trees.

Using cross-validation and stratified train-test splits, the model's performance was assessed using measures such as F1-score, accuracy, sensitivity, and specificity.

The python codes used in this research is available at, https://colab.research.google.com/github/pgbio99/Breast-cancer-/blob/main/Random%20forest%20_%20BRCA%20subtypes%20(2).ipynb.

2.6 Feature Selection and Prioritization of Candidate Biomarkers

Promoting possible biomarkers and interpreting model predictions:

For the KNN model, permutation importance was computed by quantifying the drop in model accuracy when the values of each feature were randomly shuffled.

For the Random Forest, feature importance was determined by the mean decrease in Gini impurity for each gene is:

$$\text{Gini Importance}(j) = \sum_{t \in T} p(t)\Delta i(t, j)$$

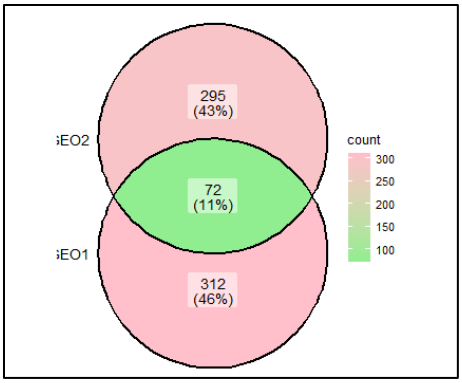
Where $p(t)$, is the proportion of samples reaching node t, and $\Delta i(t,j)$ is the decrease in impurity from splitting on feature j.

As potential subtype-specific indicators, the top-ranked genes from the two models were contrasted and examined further. By taking these actions, a strong and comprehensive machine learning framework for categorizing breast cancer subtypes and finding biomarkers was achieved.

3. RESULTS

3.1 Discovery of DEGs

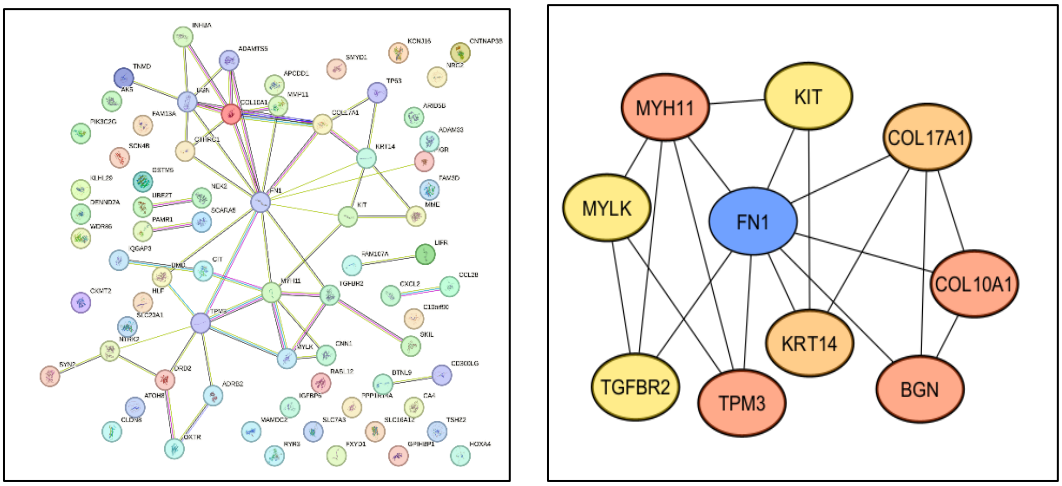
The GSE86374 dataset contained 306 significant differentially expressed genes, while the GSE57297 dataset contained 426 DEGs. These gene sets intersected to produce a shortlist of 72 common DEGs for additional examination (Figure 2).



In Figure 2: Venn diagram visualized using R software, where GEO1 represents GSE86374 and GEO2 represents GSE57297.

3.2 Network-Based Identification of Central Hub Genes

Using the STRING database, a PPI network was created using the 72 frequent DEGs, and Cytoscape was used to visualize the top 10 cytohub genes network (Figure 3). Using degree centrality, we identified hub genes, including KIT, COL17A1, COL10A1, TPM3, MYLK, MYH11, FN1, BGN, KRT14,and TGFBR2.

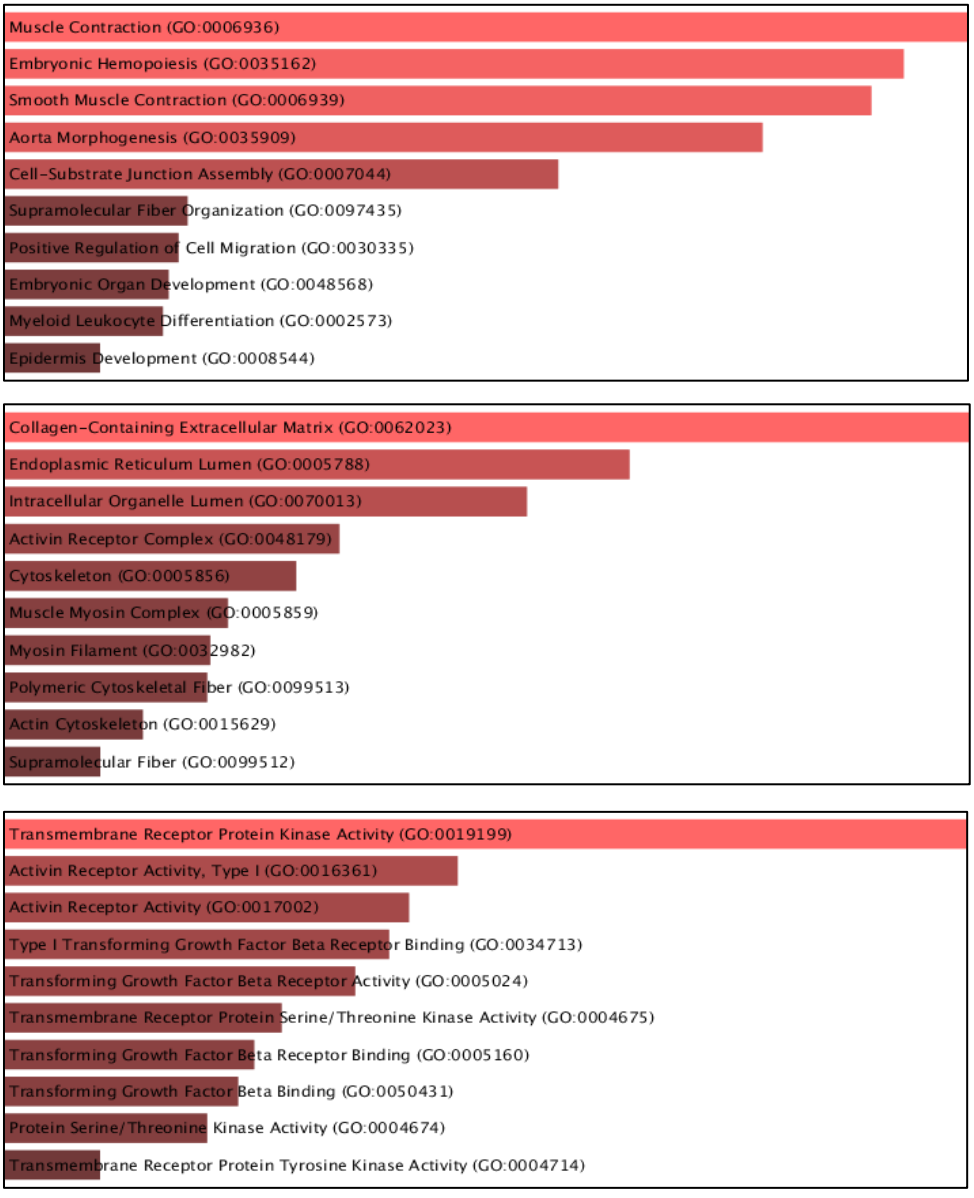


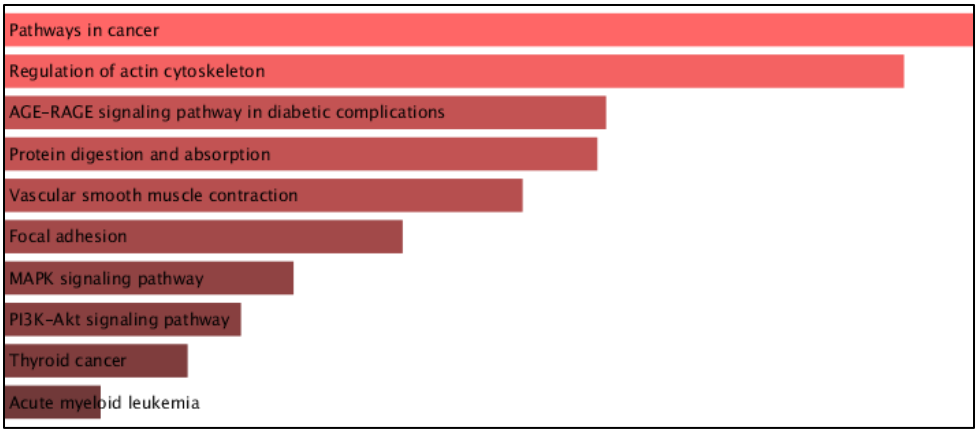
In Figure 3: Protein-Protein Interaction of 72 genes (left network) and 10 hub genes (right network).

3.3 Functional Annotation and Pathway Mapping of Hub Genes

Enrichr database was used to analyze functions and pathways associated with these 10 hub genes. Gene Ontology and KEGG analyses revealed that these 10 hub genes are involved in top-ranking pathways such as Muscle Contraction, Pathways in cancer,

Regulation of actin cytoskeleton, Vascular smooth muscle contraction, Transmembrane receptor protein Kinase activity, Type I transforming growth factor Beta receptor binding, Activin receptor activity, Aorta morphogenesis, Embryonic hemopoiesis, Collagen-containing extracellular matrix, and Cytoskeleton organization (Figure 4).

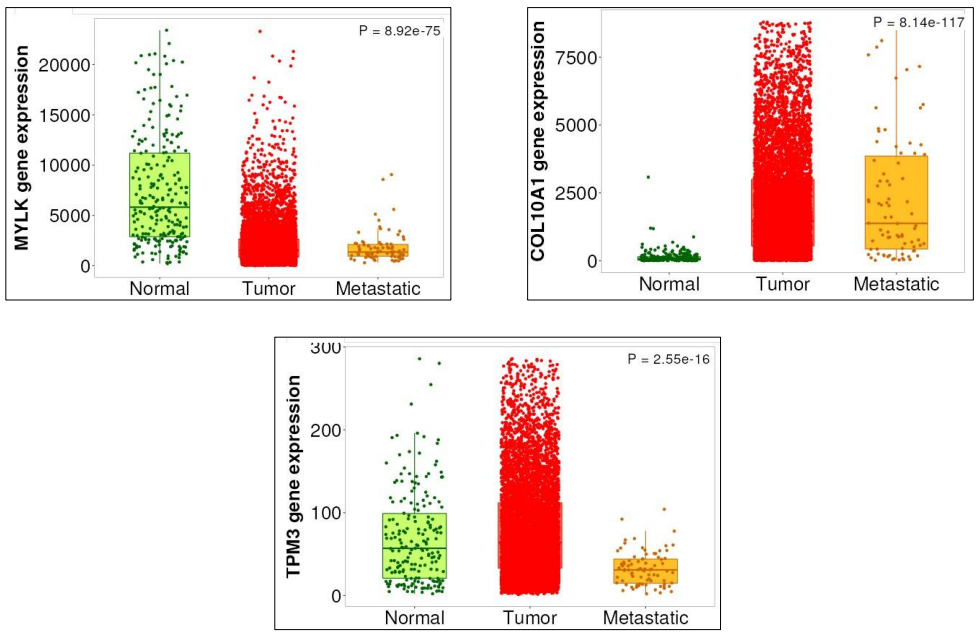




In Figure 4: These are the chart representations of GO-Biological Processes, GO-Cellular Components , GO-Molecular Functions and KEGG pathways.

3.4 Expression Patterns and Genomic Alteration Analysis

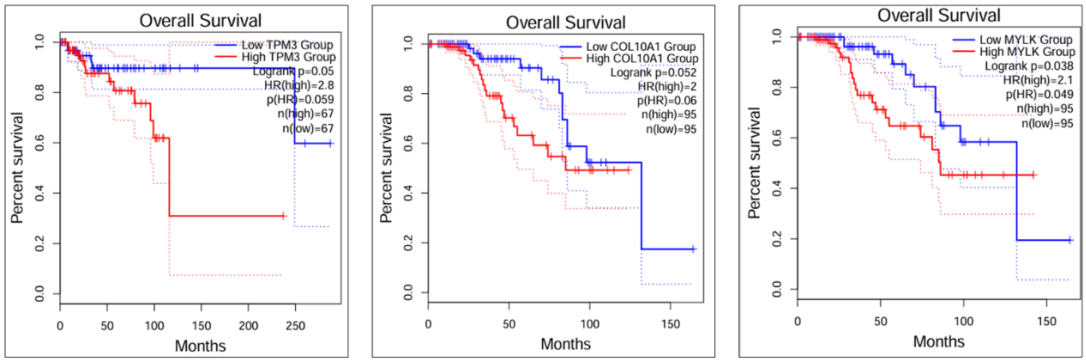
TNMplot expression analysis revealed that TPM3 and COL10A1 are increased in both tumor and metastatic tissues, whereas MYLK is marginally downregulated in tumor and significantly downregulated in metastatic samples (Figure 5). Genes show variable levels of amplification and deletion across breast cancer patients, according to alteration frequency analysis using cBioPortal. TPM3 had the highest alteration frequency.



In Figure 5: Gene expression in normal, tumor and metastatic tissues by TNMplot.

3.5 Assessment of Prognostic Value via Survival Analysis

Survival analyses using KM Plotter and GEPIA2, highlighted several hub genes with significant prognostic value. Notably, MYLK, TPM3 and COL10A1 were associated with poorer survival outcomes. In Luminal-B subtype, gene COL10A1 and MYLK show significant p value for survival analyses (Figure 6).



TPM3
P-val: 0.05, HR : 2.8

COL10A1
P-val: 0.052, HR : 2
(for LUM-B subtype)

MYLK
P-val: 0.038, HR : 2.1
(for LUM-B subtype)

In Figure 6: Survival analyses curve using GEPIA2.

3.6 Subtype Classification Performance of Machine Learning Models

As we used two supervised models Random Forest and KNN for classification purpose; Random Forest consistently outperformed KNN in terms of classification metrics:

Accuracy: ~91%

F1 Score: ~89%

Sensitivity: ~90%

Specificity: ~92%

KNN yielded moderate performance:

Accuracy: ~86%

F1 Score: ~84%

Sensitivity: ~85%

Specificity: ~87%

These results demonstrate the effectiveness of tree-based ensemble methods like Random forest in handling high-dimensional microarray data with complex subtype distributions.

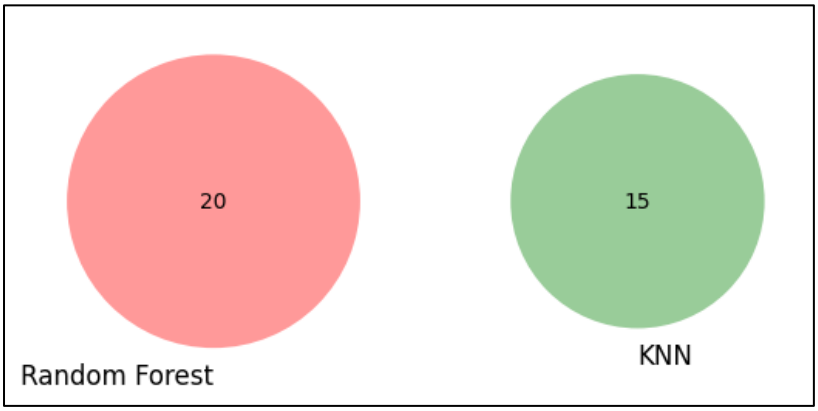
3.7 Feature Importance Analysis and Biomarker Candidate Selection

We prioritized genes based on their contributions to model prediction:

Random Forest, gene importance was ranked using the mean decrease in Gini impurity.

KNN, applied permutation importance, where feature values were randomly shuffled to assess impact on accuracy.

Interestingly, there was no overlap among the top-ranked features identified by the two models that is 20 from Random Forest and 15 from KNN (Figure 7).



In Figure 7: Venn diagram showing no intersection between top features of RF and KNN.

Due to lack of overlapping top features, we considered all union genes from both models as a pool of candidates of biomarker discovery. The following genes emerged as high-confidence candidates from at least one mode: TMEM45B, POF1B, STIL, SNORD116-3, SLC19A2, IFRD1, RARA, CMBL, PRR15, HSPA5, IQCB1, DNA2, CISD1, CALML4, E2F7, PARD6B, TMPRSS6, BTF3L4, REEP6, CDH22, TPX2, NRIP1, H2BP1, RAB17, ASB18, SERPINA3, CHD7, ANKRD30A, COQ3, CYP2B6, ZC3HAV1L, SAA2-SAA4, SLITRK6, RPL5, ZNF814.

This approach ensured no potentially informative gene was missed due to algorithm-specific biases. These 35 subtype-specific DEGs represent strong biomarker candidates for downstream validation and functional analysis.

4. DISCUSSION

With the aim to improve classification of breast cancer subtypes and find reliable biomarkers, this work effectively integrated machine learning modeling with bioinformatic

analysis. We created a protein-protein interaction network and discovered 72 common differentially expressed genes (DEGs) by combining two separate GEO microarray datasets. This network highlighted 10 hub genes that are crucial to the biology of breast cancer. Notably, the literature has extensively documented the roles of genes such as KIT, COL10A1, TPM3, COL1A1, MYLK, and MYH11 in cytoskeletal remodeling, cell migration, and extracellular matrix organization-all of which are essential for tumor growth and metastasis.

By demonstrating their involvement in important pathways like muscle contraction, actin cytoskeleton control, and transforming growth factor-beta signaling, functional enrichment analysis provided additional evidence of these hub genes biological significance. It is well recognized that these pathways affect the invasiveness, motility, and dynamic interactions of cancer cells with their surroundings, Their possible roles in disease progression and metastasis are suggested by the downregulation of MYLK in metastatic samples and the reported overexpression of TPM3 and COL17A1 in tumor and metastatic tissues. Furthermore, TPM3'high change frequency among breast cancer patients suggests that it may be important as a genomic driver or an indicator of genomic instability.

The clinical significance of these findings was highlighted by survival studies. Poorer patient outcomes were substantially linked to genes including MYLK, TPM3 and COL17A1, especially in aggressive subtypes like Luminal B. This implies that these genes may function as prognostic indications in addition to diagnostic markers, supporting risk assessment and personalized therapy planning.

The usefulness of gene expression profiles for subtype classification was further confirmed using machine learning algorithms. The Random Forest model outperformed KNN and showed the power of ensemble approaches in managing intricate, high-dimensional biological data with its excellent accuracy, sensitivity, and specificity. Crucially, a varied group of 35 potential biomarkers was identified by feature importance analysis; several of them have established or developing functions in cancer biology (e.g., TMEM45B, DNA2, E2F7, HSPA5, PRR15). The robustness of the union set of candidate genes and the complementary nature of various algorithms are highlighted by the lack of overlap between the top features from Random Forest and KNN.

This study is peculiar since it uses machine learning and bioinformatics to prioritize both new and established biomarkers for breast cancer subtyping in a systematic manner. Compared to conventional single-method research, our technique captures a wider range of physiologically and clinically relevant genes by combining network-based hub gene identification with precise machine learning feature selection. The small sample size of the GEO datasets examined, however, is a significant drawback that could compromise the conclusions statistical power and generalizability. The need for more research to confirm and expand these findings is further highlighted by the lack of external validation using different platforms or independent datasets (such RNA-seq). Translating these potential biomarkers into clinical practice will require extending this integrative approach to bigger and more varied patient groups.

Finding these subtypes-specific DEGs provides important information on the molecular diversity of breast cancer. Many of these genes are implicated in pathways that are often dysregulated in cancer, including those that regulate the cell cycle, DNA repair, apoptosis, and cellular stress responses. Their potential as biomarkers may help with improved prognostic evaluation, more precise subtyping and earlier discovery, which would ultimately lead to more specific and successful treatment plans.

5. CONCLUSION

To sum up, this work shows how effective it is to combine machine learning and bioinformatics techniques in order to find reliable biomarkers for breast cancer. We discovered important hub genes and a group of 35 subtype-specific DEGs with significant biological and clinical significance by examining two separate microarray datasets. The prognostic significance of these genes signatures is further supported by Random Forest model's excellent performance categorizing breast cancer subtypes. Crucially, a number of discovered genes are connected to important pathways that affect tumor growth, metastasis, and patient outcome. These results demonstrate the potential of these biomarkers to enhance early diagnosis. Prognostic evaluation, and personalized therapy plans in the management of breast cancer and provide a strong basis for further experimental validation. Interestingly, there was no overlap between the top biomarkers determined by bioinformatics analysis and those ranked by machine learning models, demonstrating the

complimentary advantages of both methodologies. Our analysis reduces the possibility of method-specific biases omitting clinically significant signs by taking into account the union of these candidate genes, which captures a larger and possibly more comprehensive range of biomarkers.

REFERENCES

1. Huang, J., et al., *Global incidence and mortality of breast cancer: a trend analysis*. Aging (Albany NY), 2021. **13**(4): p. 5748-5803.
2. Sedeta, E.T., B. Jobre, and B. Avezbakiyev, *Breast cancer: Global patterns of incidence, mortality, and trends*. Journal of Clinical Oncology, 2023. **41**(16_suppl): p. 10528-10528.
3. Sung, H., et al., *Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries*. CA Cancer J Clin, 2021. **71**(3): p. 209-249.
4. Siegel, R.L., et al., *Cancer statistics, 2023*. CA Cancer J Clin, 2023. **73**(1): p. 17-48.
5. Sørlie, T., et al., *Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications*. Proc Natl Acad Sci U S A, 2001. **98**(19): p. 10869-74.
6. Yan, S. and S. Yue, *Identification of early diagnostic biomarkers for breast cancer through bioinformatics analysis*. Medicine (Baltimore), 2023. **102**(37): p. e35273.
7. Yu, Z., et al., *RNA-Seq-Based Breast Cancer Subtypes Classification Using Machine Learning Approaches*. Comput Intell Neurosci, 2020. **2020**: p. 4737969.