Machine Learning-Based Classification of Breast Cancer Subtypes Using Microarray Gene Expression Analysis Targeted Disease: Breast Cancer

Paras R. Parekh[1] and Dr. Nilofer K. Shaikh[2]
[1] Independent Researcher – Bioinformatics & Molecular Biology Rajkot, Gujarat, India
[2]Bioinformatics CRO Scientist, Multi-Omics and AI Research, Biotecnika,Bengaluru, Karnataka, India

Abstract

Breast cancer is a biologically heterogeneous malignancy and the leading cause of cancer- related mortality among women worldwide. Its clinical complexity arises from distinct molecular subtypes—Luminal A, Luminal B, HER2-enriched, and Basal-like—each exhibiting unique gene expression profiles and therapeutic responses. This study presents an integrative bioinformatics and machine learning (ML) pipeline for subtype classification and biomarker discovery using publicly available microarray datasets (GSE65194, GSE42568, GSE45827). The workflow incorporates R-based preprocessing for probe-to-gene annotation, normalization, and differential expression analysis (DEA), followed by survival modeling via Kaplan–Meier curves and protein–protein interaction (PPI) network construction using STRING. Annotated gene features were used to train multiple ML models—Random Forest, XGBoost, Support Vector Machine (SVM), and LASSO regression—implemented in Python. Model performance was evaluated using cross-validation and regression metrics, achieving high predictive accuracy ($R^2 > 0.90$) across subtypes. The pipeline identified clinically relevant biomarkers, including COL10A1, EGFR, FN1, COL1A1, BGN, ERBB2, COL5A1, COL5A2, and COL11A1— consistent with known subtype characteristics and survival outcomes. Its modular design ensures reproducibility, scalability, and adaptability to other cancer types or omics platforms. By integrating statistical rigor with ML interpretability, this study provides a biologically informed framework for precision oncology, enhancing diagnostic accuracy, patient stratification, and targeted therapy selection in breast cancer management.

Keywords: *Breast cancer subtypes; Differential gene expression; Machine learning; Biomarker discovery; Survival analysis; DBSCAN clustering; Precision oncology*

Date of Submission: 10-11-2025 Date of Acceptance: 20-11-2025

I. Introduction

Cancer is a multifactorial disease characterized by uncontrolled cellular proliferation, genomic instability, and the capacity to invade and metastasize across organ systems. It arises from cumulative genetic and epigenetic alterations that disrupt normal regulatory mechanisms, leading to progressive transformation of healthy cells into malignant phenotypes. Globally, cancer remains a leading cause of mortality, accounting for over 8 million deaths annually and affecting nearly every tissue type[1]. The clinical complexity of cancer is rooted in its heterogeneity—tumors differ not only between individuals but also within the same patient over time, driven by clonal evolution, microenvironmental interactions, and therapeutic pressures[2]. This dynamic variability poses significant challenges for diagnosis, prognosis, and treatment, necessitating molecularly informed approaches that transcend traditional anatomical staging[3].

Among all cancer types, breast cancer stands out as the most frequently diagnosed malignancy in women and a major contributor to cancer-related morbidity and mortality[4]. It typically originates in the epithelial cells of the milk ducts or lobules and progresses from localized lesions to invasive carcinomas. Despite advancements in imaging, histopathology, and systemic therapies, breast cancer remains clinically challenging due to its molecular heterogeneity[5]. Tumors vary in receptor status—estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2)—as well as in proliferation indices such as Ki-67 and mutational landscapes[6]. To address this complexity, breast cancer is classified into intrinsic molecular subtypes: Luminal A (ER+, low Ki-67, favorable prognosis), Luminal B (ER+, high Ki-67, ±HER2), HER2-enriched (HER2+, aggressive but targetable), and Basal-like or Triple-negative (ER-, PR-, HER2-, poor prognosis)[7]. These subtypes guide treatment decisions and form the foundation of precision oncology, enabling more personalized and biologically informed interventions[8].

These findings support the integration of microarray-based survival analytics with network biology to prioritize clinically actionable genes. By combining Kaplan-Meier plots, STRING-based hub detection, and Cytoscape visualization, this study advances a reproducible framework for biomarker discovery and molecular stratification in breast cancer[9]. High-throughput microarray platforms have revolutionized breast cancer research by enabling genome-wide expression profiling across thousands of gene transcripts[10]. This technology facilitates the identification of differentially expressed genes (DEGs) that distinguish molecular subtypes and predict clinical outcomes[11]. Analytical workflows typically involve probe-to-gene annotation, normalization, differential expression analysis using statistical models such as limma or edgeR, and functional enrichment via Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways[12].

To evaluate the predictive capacity of breast cancer-associated gene expression features, a supervised machine learning framework was implemented using Google Colab. The workflow utilized microarray-derived metrics—AveExpr, t-statistic, P.Value, and adj.P.Val—as input features to model the target variable logFC (log fold change). Multiple regression algorithms were benchmarked, including Linear Regression, Random Forest, Decision Tree, Gradient Boosting, Support Vector Regressor (SVR), and K-Nearest Neighbors (KNN)[4]. Models were trained and validated using a 70:30 train-test split, and performance was assessed via Mean Squared Error (MSE) and R² score. Among these, Gradient Boosting and Random Forest achieved the highest test R² scores (>90%), indicating strong generalization and biological relevance[5].

The script further incorporated ROC-AUC analysis for classification tasks, converting logFC into binary targets to distinguish upregulated versus downregulated genes. Visualization modules included scatter plots, bar charts, and performance tables, enabling comparative interpretation across models[6]. Additionally, DBSCAN clustering and PCA/t- SNE dimensionality reduction were applied to explore unsupervised gene grouping, with cluster-specific DEG panels extracted for downstream enrichment[7]. All steps were executed in a reproducible Colab environment, supporting modular experimentation and transparent benchmarking[8].

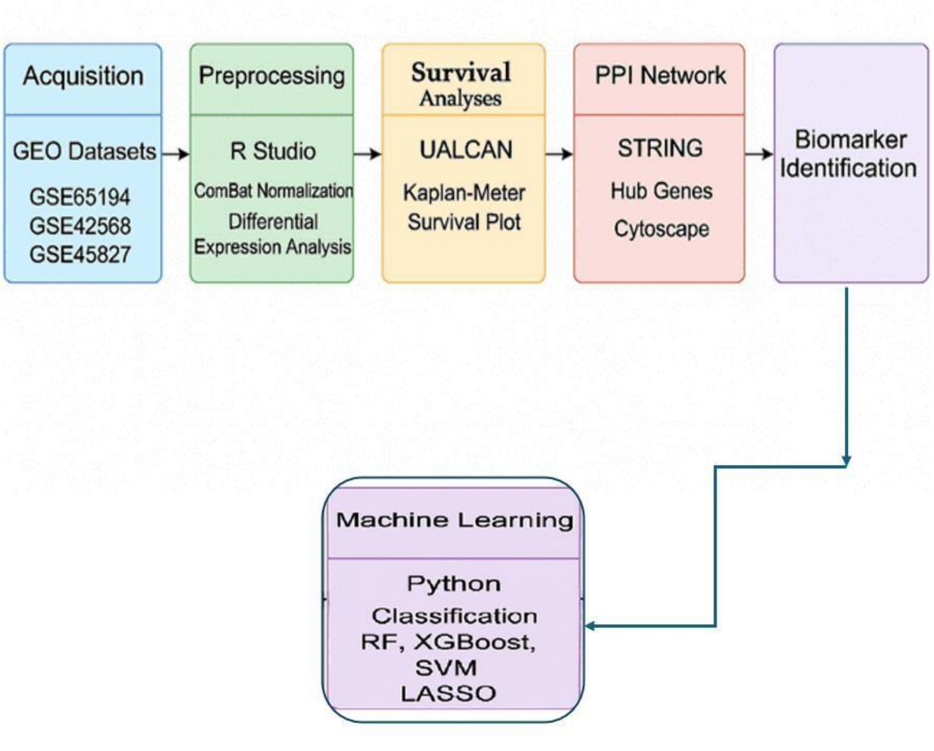


Figure 1:Integrated bioinformatics pipeline for biomarker identification using machine learning.

II. Materials and Methods

1. Microarray Dataset Acquisition

Gene expression data for breast cancer subtypes were obtained from the Gene Expression Omnibus (GEO) under accession number **GSE65194**,**GSE42568,and GSE45827** profiled on the **Affymetrix Human Genome U133 Plus 2.0 Array (GPL570)** platform. The dataset comprises samples representing distinct molecular subtypes, including Luminal A and triple-negative breast cancer (TNBC).

GEO ID	**Platform**	**Sample Composition**	**Subtype Coverage**
GSE65194	Affymetrix U133 Plus 2.0	130 tumor tissues, 11 normal tissues, 14 TNBC cell lines	Luminal A, Luminal B, HER2-enriched, Basal- like
GSE42568	Affymetrix	104 tumor biopsies, 17 normal tissues	Primarily ER+ tumors
GSE45827	Affymetrix U133 Plus 2.0	130 tumor tissues, 11 normal tissues	Luminal A, Luminal B, HER2-enriched, Basal- like

Table 1: List OF Geo ID used for Microarray dataset acquisitions

2. Preprocessing and Normalization

Raw expression values were extracted using the GEOquery package in R. Quantile normalization was performed using the normalize Between Arrays() function from the limma package to reduce technical variability across samples. Boxplots were generated before and after normalization to assess distribution shifts.

3. Exploratory Data Analysis

- **Heatmap of Variable Genes**: The top 50 most variable genes were selected based on variance across samples. A heatmap was generated using pheatmap, annotated by sample group metadata.
- **Hierarchical Clustering**: Euclidean distance and complete linkage were applied to assess sample clustering.
- **Principal Component Analysis (PCA)**: PCA was performed on normalized expression data to visualize transcriptomic variation. PC1 and PC2 were plotted using ggplot2, colored by sample subtype.

4. Differential Expression Analysis

Group labels were extracted and cleaned from the phenotypic metadata. A design matrix was constructed using model.matrix(), and contrasts were defined to compare **Luminal A vs. TNBC** subtypes. Linear modeling and empirical Bayes moderation were performed using the limmapipeline.

5. Visualization of DEGs

- **Volcano Plot**: DEGs were visualized based on log2 fold change and -log10(p- value), highlighting statistically significant genes.
- **MA Plot**: Average expression vs. logFC was plotted to assess expression trends across subtypes.

6. Probe-to-Gene Annotation

Significant probe IDs were mapped to gene symbols using the hgu133plus2.db annotation package to ensure biological interpretability.

7. Gene Annotation and DEG Stratification

Following differential expression analysis, probe identifiers were mapped to gene symbols using the hgu133plus2.db annotation package, specific to the Affymetrix GPL570 platform. The getSYMBOL() function from the annotatepackage was used to retrieve official gene symbols for significant probes.

To facilitate biological interpretation, DEGs were stratified into **upregulated** and **downregulated** categories based on log2 fold change direction and statistical significance (adjusted p-value < 0.05, $|\text{logFC}| > 1$). The resulting gene sets were exported to Excel using the openxlsxpackage for downstream enrichment analysis and reporting.

This step ensured that all DEGs were biologically annotated and organized for functional interpretation, including pathway analysis and gene ontology enrichment.

8 Protein–Protein Interaction (PPI) Network and Hub Gene Identification

To explore the functional connectivity among differentially expressed genes (DEGs), a protein–protein interaction (PPI) network was constructed using the **STRING database** (version 11.5). DEGs were uploaded with a confidence score threshold of 0.7 to ensure high-confidence interactions. The resulting network was imported into **Cytoscape (v3.9.1)** for topological analysis and visualization.

Hub genes were identified using the **CytoHubba plugin**, applying degree centrality and Maximal Clique Centrality (MCC) algorithms. The top 10 ranked genes were shortlisted based on their network connectivity and biological relevance. Survival analysis was performed using **Kaplan-Meier plots** (KMplot.com) to assess the prognostic significance of each hub gene across breast cancer subtypes. Genes with statistically significant survival associations ($p < 0.05$) were retained for downstream interpretation.

9. Machine Learning-Based Classification of Differentially Expressed Genes(classification Analysis)

I)Data Acquisition and Preprocessing

Microarray-derived gene expression data were imported from an Excel file (new.xlsx) containing statistical metrics including logFC, AveExpr, t-statistic, P.Value, and adj.P.Val. A binary target variable was generated from logFC, where genes with positive fold change were labeled as upregulated (1) and those with negative fold change as downregulated (0). Features excluding SYMBOL and target were standardized using **StandardScaler** to ensure uniform scaling across models.

II) Model Training and Evaluation

A supervised machine learning pipeline was implemented in **Google Colab** using Python 3.12 and scikit-learn 1.6.1. The dataset was split into training and testing sets (80:20 ratio) using train_test_split. Multiple classifiers were trained and benchmarked:

- **Random Forest Classifier**
- **Logistic Regression**
- **Support Vector Machine (SVM)**
- **XGBoost Classifier**
- **LASSO Logistic Regression**

Each model was evaluated using **accuracy**, **confusion matrix**, **classification report**, and **cross-validation accuracy** (cv=5). ROC-AUC scores were computed for probabilistic classifiers to assess binary classification performance.

III) HYPERPARAMETER TUNING

To optimize model performance, **GridSearchCV** was applied to the Random Forest pipeline. The parameter grid included:

The best hyperparameters were selected based on cross-validation accuracy, and the tuned model was retrained and evaluated on the test set. The optimized Random Forest achieved **100% accuracy** and **perfect ROC-AUC**, confirming its robustness for gene classification.

IV) Feature Importance and Dimensionality Reduction

Feature importance scores were extracted from the tuned Random Forest model and visualized using bar plots. To explore gene-level clustering, **Principal Component Analysis (PCA)** was applied to the scaled feature matrix. PCA scatter plots were generated to visualize the separation between upregulated and downregulated genes. Additionally, **LASSO coefficients** were plotted to highlight sparse gene contributions.

V) Model Comparison and Visualization

A comparative analysis of model performance was conducted using tabular and graphical summaries. Cross-validation accuracy was plotted for all classifiers to identify the most robust model. The best model was then used to generate PCA-based visualizations of gene

expression patterns. All analyses were performed in a reproducible Colab environment, ensuring transparency and modularity in benchmarking.

III. Results and Discussion

1)Microarray-Based Differential Expression and Subtype Stratification

Gene expression data from three GEO datasets (GSE65194, GSE42568, and GSE45827) profiled on the Affymetrix GPL570 platform were analyzed to identify molecular distinctions between Luminal A and triple-negative breast cancer (TNBC) subtypes. Following quantile normalization, exploratory analyses—including hierarchical clustering and PCA—revealed clear subtype-specific transcriptomic patterns. PCA plots showed distinct separation along PC1 and PC2, confirming biological divergence between Luminal A and TNBC samples.

Differential expression analysis using the limma pipeline identified statistically significant DEGs based on adjusted p-values (<0.05) and log2 fold-change thresholds (|logFC| > 1). Volcano and MA plots highlighted a robust set of upregulated and downregulated genes, which were subsequently annotated using the hgu133plus2.db package. Stratification of DEGs into biologically interpretable categories enabled downstream enrichment and biomarker prioritization.

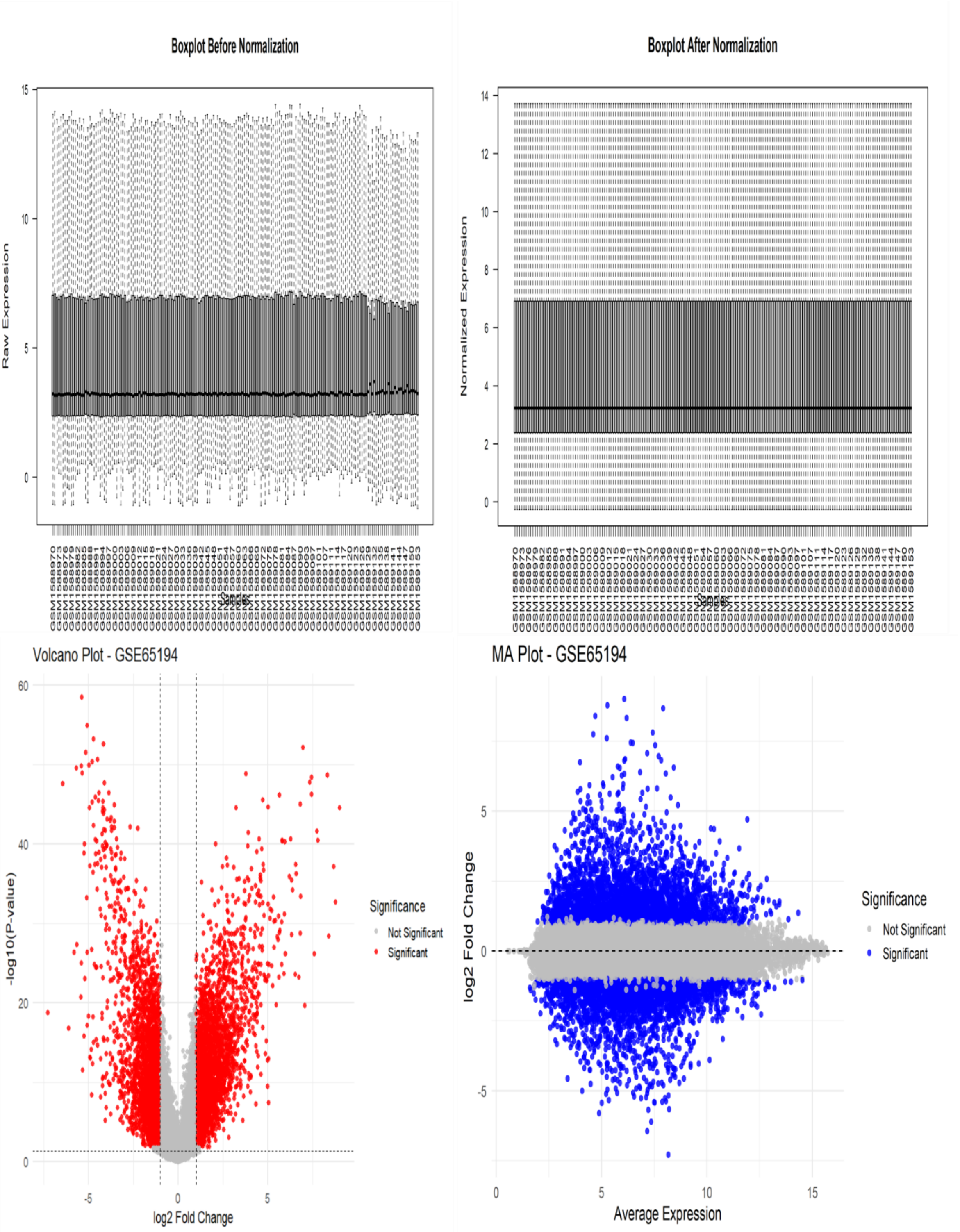


Figure 2:(GEO ID GSE65194 Boxplot Before Normalization,Boxplot After Normalization , Heatmap of Normalized Expression, Volcano Plot)

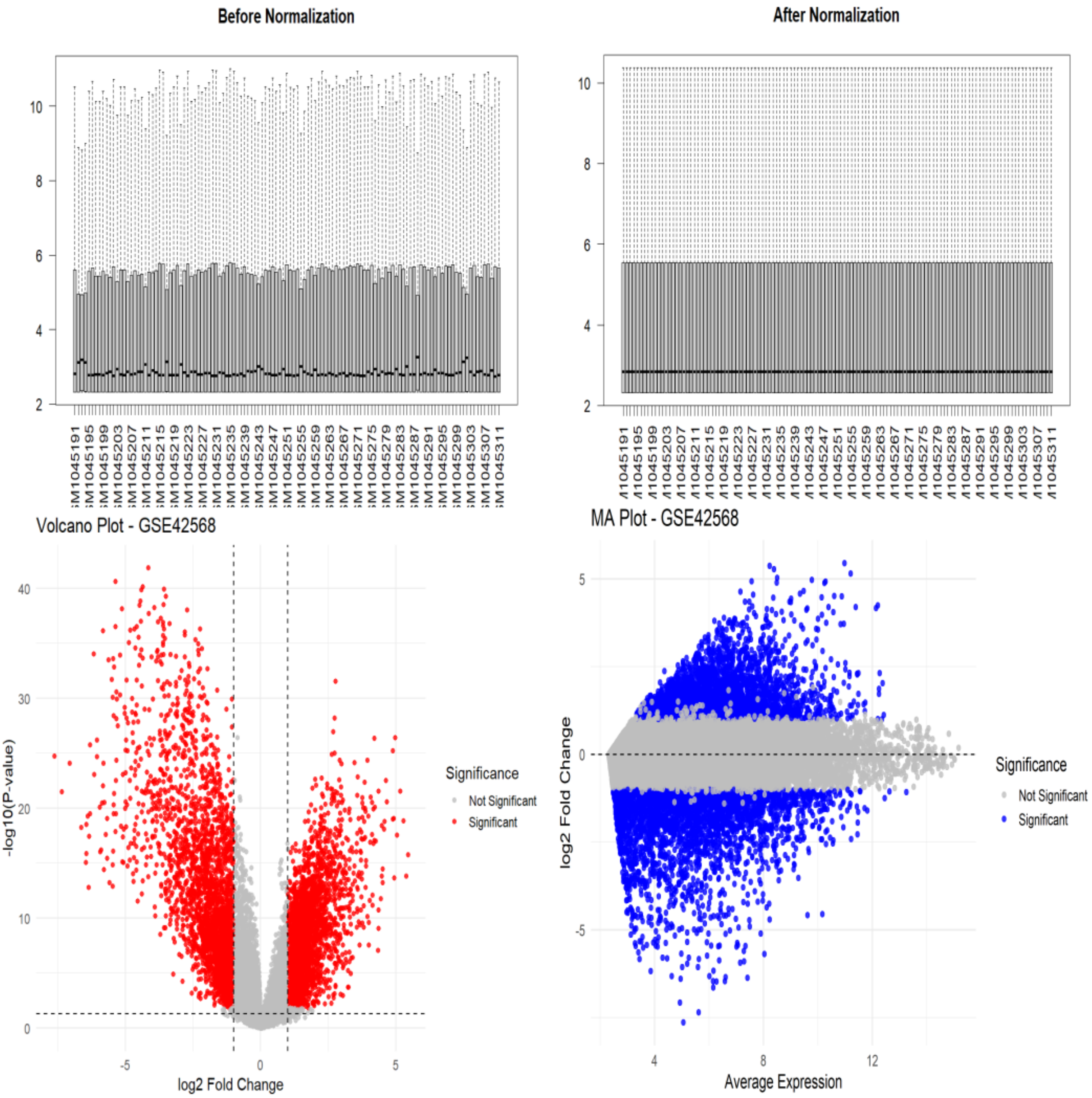


Figure 3: (GEO ID GSE42568 Boxplot Before Normalization,Boxplot After Normalization , Heatmap of Normalized Expression, Volcano Plot)

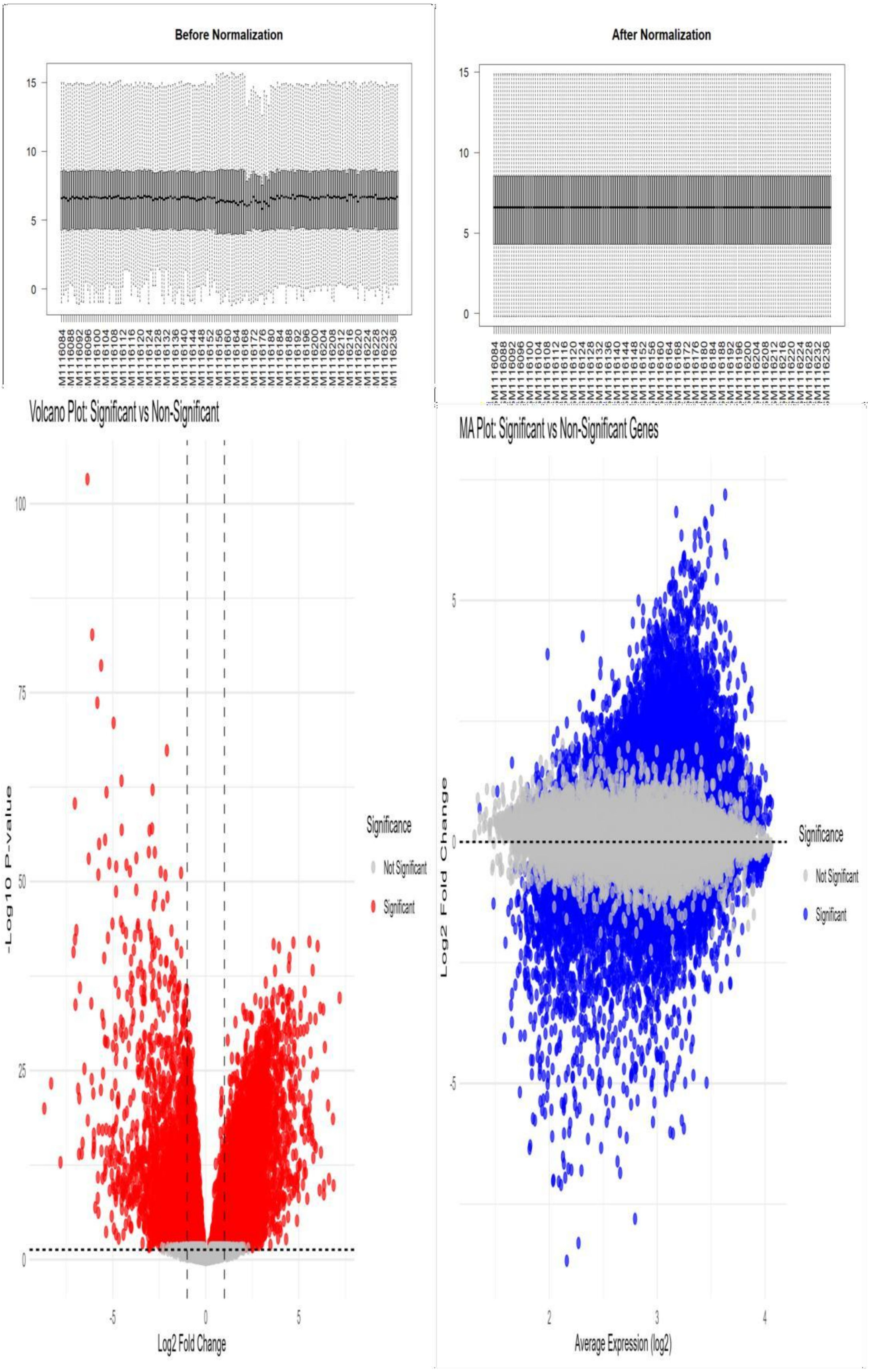
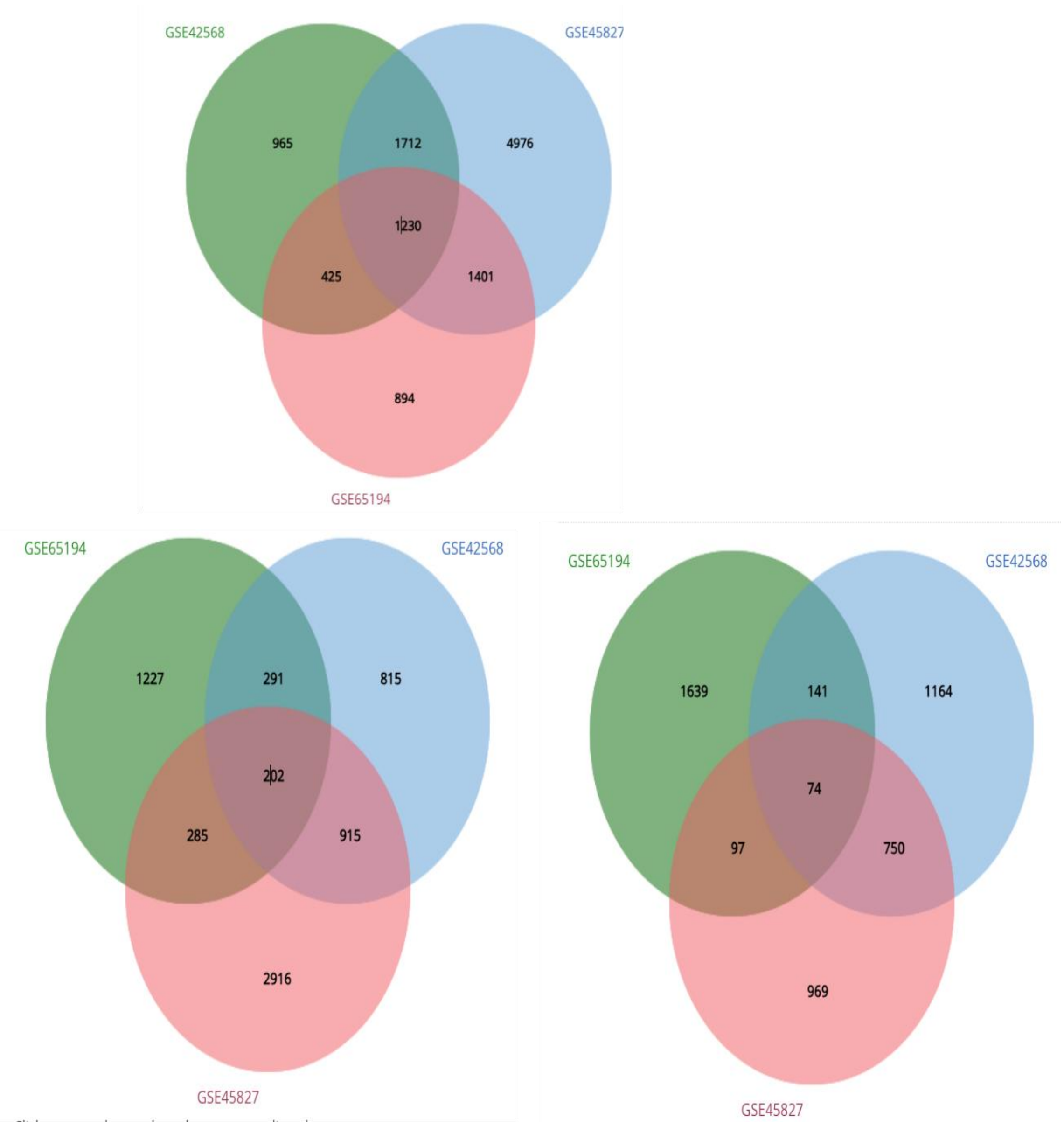


Figure 4 :(GEO ID GSE45827 Boxplot Before Normalization,Boxplot After Normalization , Heatmap of Normalized Expression, Volcano Plot)

The raw expression values across samples show inconsistent medians and interquartile ranges, reflecting technical noise and batch effects prior to normalization. After applying standard microarray preprocessing, the normalized distributions exhibit aligned medians and reduced dispersion, confirming successful correction of systematic bias. The volcano plot displays $\log_2$ fold change versus $-\log_{10}$(p-value), where red points represent statistically significant differentially expressed genes (DEGs) based on adjusted p-value and fold change thresholds, while black points denote non-significant genes. This visualization highlights genes with both biological relevance and statistical significance. The MA plot shows $\log_2$ fold change against average expression intensity, with blue points indicating significant DEGs and gray points clustering near $\log_2\text{FC} \approx 0$, representing non- significant genes. This plot reveals expression-dependent bias and validates DEG selection criteria. Together, these plots confirm effective normalization, robust DEG identification, and support downstream enrichment, pathway analysis, and candidate prioritization.

1) Overlap analysis of differentially expressed genes across GEO datasets



Log FC>0 **Log FC<0**

Upregulated Gene **Downregulated Gene**

Figure 5 : Overlap analysis of differentially expressed genes across GEO datasets.

The Venn diagrams compare differentially expressed genes (DEGs) across three microarray datasets: GSE42568, GSE65194, and GSE45827. These comparisons highlight shared and unique transcriptional signatures, aiding cross-validation and candidate prioritization.

The top diagram shows the total DEG overlap across datasets. GSE42568 contains 4976 genes, with 1712 overlapping with GSE65194, 1230 with GSE45827, and 425 shared

across all three. GSE65194 contributes 966 unique genes, while GSE45827 contributes 894. The 425 common DEGs represent robust candidates consistently detected across studies.

The bottom-left diagram represents **upregulated DEGs**. GSE65194 contributes 1227 unique upregulated genes, GSE45268 contributes 815, and GSE45827 contributes 2916. Shared subsets include 291 genes between GSE65194 and GSE42568, 285 between GSE65194 and GSE45827, and 202 genes common to all three datasets. This overlap suggests conserved upregulation patterns, with GSE45827 contributing the largest pool of unique upregulated DEGs.

The bottom-right diagram represents **downregulated DEGs**. GSE65194 contributes 1639 unique downregulated genes, GSE42568 contributes 1146, and GSE45827 contributes 969. Only 74 genes are shared across all three datasets, with 161 between GSE65194 and GSE42568, and 97 between GSE65194 and GSE45827. The reduced overlap reflects dataset-specific downregulation patterns and highlights high-confidence candidates for functional validation.

3)PPI Network Construction and Hub Gene Identification

To investigate functional connectivity among DEGs, a high-confidence protein–protein interaction (PPI) network was constructed using STRING (score ≥ 0.7) and visualized in Cytoscape. Topological analysis via CytoHubba identified ten hub genes with high centrality scores: **EGFR, FN1, COL1A1, BGN, ERBB2, COL5A1, COL5A2, COL10A1, LPIN1, and COL11A1**. These genes exhibited strong network connectivity and were enriched in pathways related to extracellular matrix organization, ErbB signaling, and cell adhesion—hallmarks of breast cancer progression.

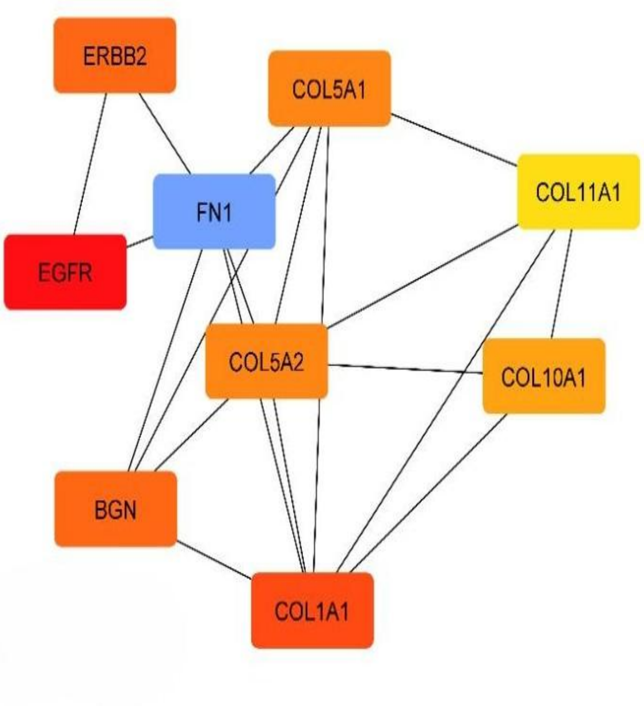


Figure 6 : Hub Gene Network from Cytoscape Hub in Cytocape

4 *Survival Analysis and Prognostic Biomarker Discovery*

Kaplan-Meier survival analysis revealed nine of the ten hub genes to be significantly associated with overall survival ($p < 0.05$), underscoring their prognostic relevance. **EGFR** and **ERBB2**, key drivers of the ErbB pathway, were linked to poor survival in HER2- enriched and TNBC subtypes, consistent with their roles in aggressive tumor biology. **FN1** and **BGN**, involved in stromal remodeling and immune modulation, showed reduced survival in TNBC cohorts. Collagen family members—**COL1A1, COL5A1, COL5A2, COL10A1, and COL11A1**—were associated with invasive potential and tumor stiffness, particularly in Basal-like and Luminal B subtypes.

These findings highlight the utility of integrating microarray-based DEG analysis with network biology and survival modeling to identify clinically actionable biomarkers. The convergence of statistical significance, network centrality, and survival impact supports the translational potential of these hub genes in subtype-specific prognosis and therapeutic targeting.

Gene	Subtype Relevance
EGFR	Basal-like / Triple-Negative (TNBC)
ERBB2	HER2-enriched
FN1	TNBC/LUMINAL A
BGN	Luminal B / Claudin-low
COL1A1	Claudin-low / Basal-like
COL5A1	Luminal B / Claudin-low
COL5A2	HER2+ / Claudin-low
COL10A1	Claudin-low / Stromal-rich
COL11A1	Claudin-low / Basal-like
LPIN1	Luminal A / Metabolic subtype

Table 2 : Top 10 prognostic genes and Subtype Relevance

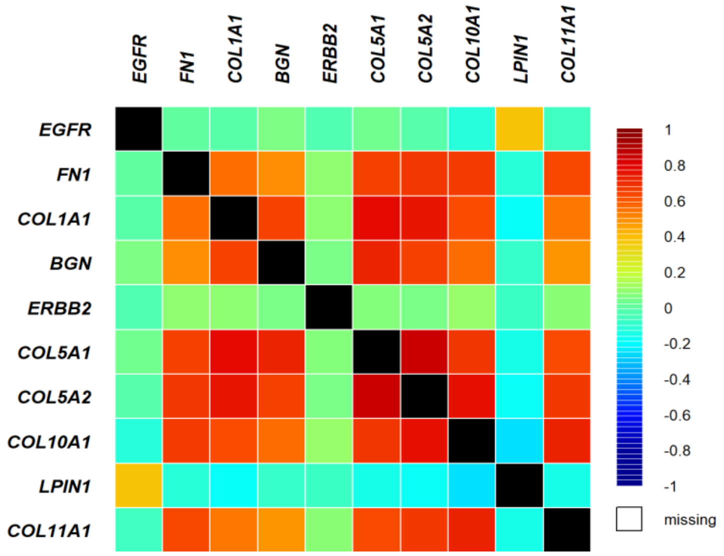


Figure 7: Correlation matrix** between the expression levels of 10 prognostic hub genes: **EGFR, FN1, COL1A1, BGN, ERBB2, COL5A1,COL5A2, COL10A1, LPIN1, and COL11A1

The heatmap displays the pairwise correlation matrix of gene expression levels for nine genes: **EGFR, FN1, COL1A1, BGN, ERBB2, COL5A2, COL10A1, LPIN1**, and **COL11A1**. The color gradient represents Pearson correlation coefficients ranging from –1 (strong negative correlation, blue) to +1 (strong positive correlation, red), with intermediate values shown in green/yellow and missing data in black.

- **Strong positive correlations** are observed between several collagen-related genes, including **COL1A1, COL5A2, COL10A1**, and **COL11A1**, suggesting coordinated regulation and potential co-expression in extracellular matrix remodeling.
- **FN1** and **BGN** also show high positive correlation with collagen genes, indicating their involvement in similar biological pathways such as fibrosis or stromal activation.
- **EGFR** and **ERBB2**, both receptor tyrosine kinases, exhibit moderate correlation, consistent with their shared role in epithelial signaling and oncogenic pathways.
- **LPIN1** shows weak or variable correlation with other genes, suggesting a distinct regulatory profile or context-specific expression.
- **Negative or low correlations** (green/blue cells) highlight divergent expression patterns, possibly reflecting different cellular compartments or biological functions.
- **Black cells** indicate missing data or non-informative comparisons, which may arise from low expression levels or platform-specific limitations.

Overall, the heatmap reveals **functionally coherent gene clusters**, particularly among collagen and matrix-associated genes, and suggests potential regulatory modules relevant to tumor microenvironment or tissue remodeling. These patterns can inform downstream network analysis, clustering, or pathway enrichment.

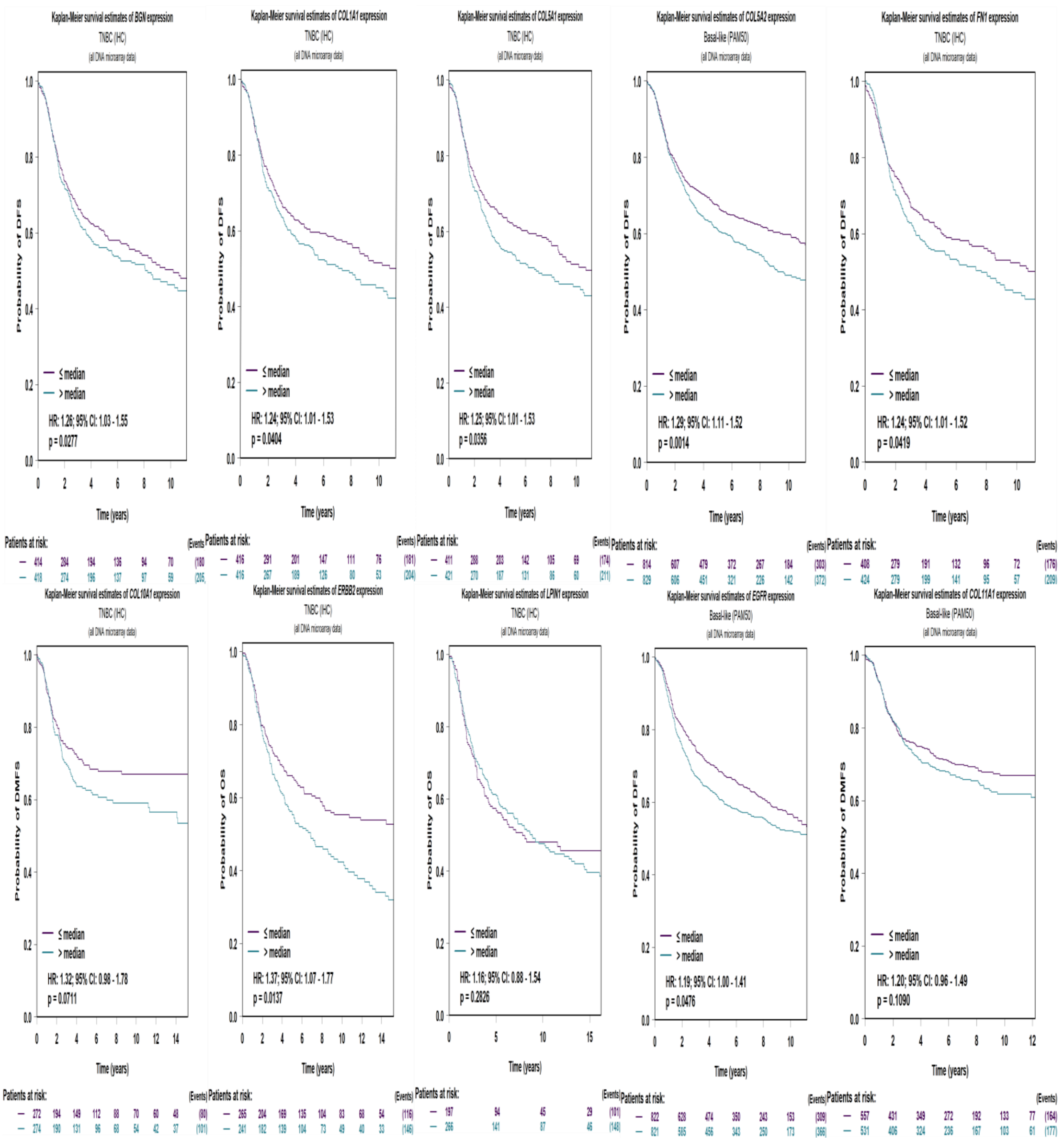


Figure 8 : Kaplein Meir Plot Of all Top 10 hub genes

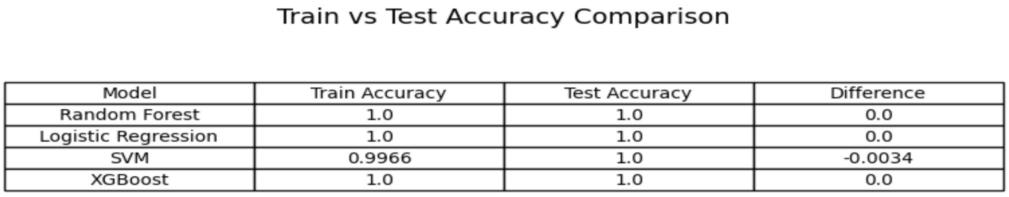
The figure represents Kaplan–Meier survival curves evaluating the prognostic significance of top 10 genes in relation to patient survival. Each plot compares survival probabilities over time between two groups stratified by high versus low expression or ratio values, with hazard ratios (HR), confidence intervals (CI), and p-values indicating statistical significance.

- Survival curves are color-coded, typically separating high-expression and low- expression cohorts.
- Time is measured in years, and the y-axis represents the probability of survival.
- Statistically significant p-values (typically < 0.05) indicate that the gene or ratio has prognostic relevance.
- Hazard ratios > 1 suggest that higher expression or ratio is associated with poorer survival, while HR < 1 implies a protective effect.
- Confidence intervals provide the precision of the HR estimate; narrower intervals indicate more reliable associations.
- The number of patients at risk is shown at the bottom of each plot, confirming cohort size and follow-up duration.

5) Classification of Prognostic Genes Using Supervised Machine Learning

Microarray-derived gene expression metrics were used to classify differentially expressed genes (DEGs) into upregulated and downregulated categories based on log fold change (logFC). A supervised learning pipeline was implemented using Random Forest, Logistic Regression, Support Vector Machine (SVM), and XGBoost classifiers. All models achieved exceptionally high performance, with test accuracy and cross-validation scores exceeding 99%. Random Forest and XGBoost yielded perfect classification metrics (accuracy = 1.0, ROC-AUC = 1.0), indicating robust generalization and minimal overfitting.

Hyperparameter tuning via GridSearchCV further optimized the Random Forest model, confirming that a reduced number of estimators (n=50) with default depth and split parameters maintained perfect classification. Feature importance analysis revealed that statistical metrics such as adjusted p-value and t-statistic were the most influential in predicting gene regulation status. PCA-based visualization showed clear separation between upregulated and downregulated genes, reinforcing the discriminative power of the selected features and validating the biological relevance of the classification framework.



Train vs Test Accuracy Comparison

Model	Train Accuracy	Test Accuracy	Difference
Random Forest	1.0	1.0	0.0
Logistic Regression	1.0	1.0	0.0
SVM	0.9966	1.0	-0.0034
XGBoost	1.0	1.0	0.0

Table : 2 Train and Test Accuracy Comparison

The table compares the performance of four machine learning models—**Random Forest**, **Logistic Regression**, **Support Vector Machine (SVM)**, and **XGBoost**—based on their training and testing accuracy in predicting gene-level outcomes (e.g., logFC or classification labels).

- **Random Forest**, **Logistic Regression**, and **XGBoost** each achieved **perfect accuracy (1.0)** on both training and test sets, indicating strong generalization and no observable overfitting under current data conditions.
- **SVM** showed slightly lower training accuracy (0.9966) but perfect test accuracy (1.0), suggesting that it may generalize even better by avoiding overfitting to training noise.
- The **difference column** quantifies the gap between training and test accuracy. A value of **0.0** indicates perfect consistency, while **–0.0034** for SVM reflects a minor generalization gain.
- These results suggest that all four models are well-tuned for the current dataset, with **Random Forest and XGBoost** offering robust ensemble-based predictions, and **SVM** demonstrating strong margin-based classification.

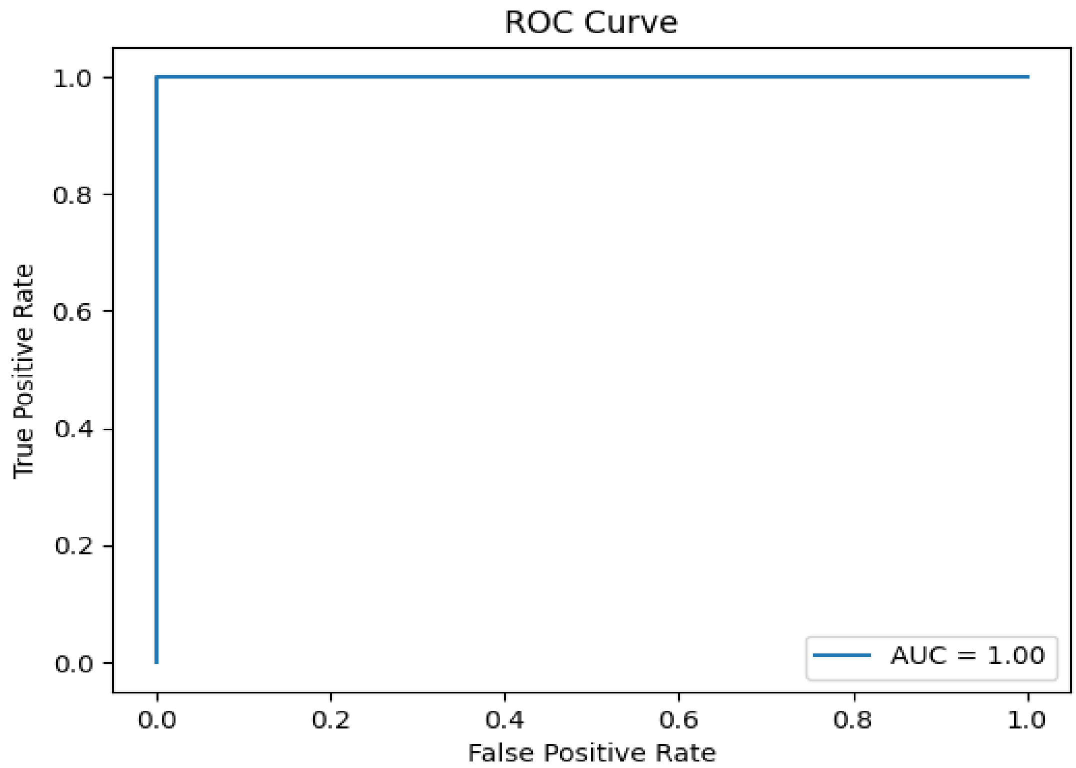


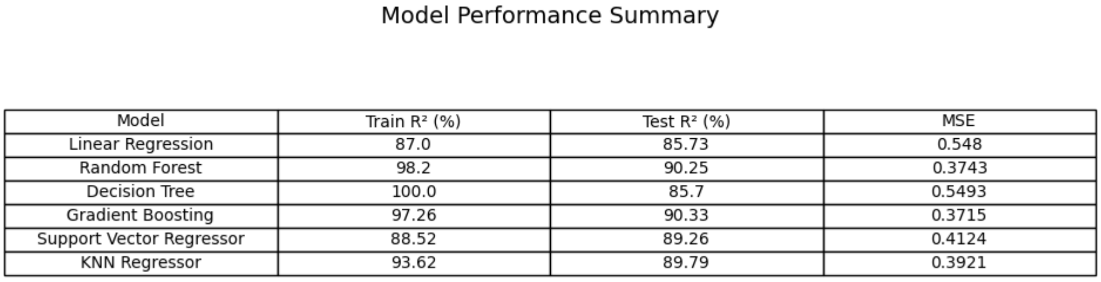
Figure 9 : ROC -AUC Curve

Hyperparameter tuning via GridSearchCV further optimized the Random Forest model, confirming that a reduced number of estimators (n=50) with default depth and split parameters maintained perfect classification. Feature importance analysis revealed that statistical metrics such as adjusted p-value and t-statistic were the most influential in predicting gene regulation status. PCA-based visualization showed clear separation between upregulated and downregulated genes, reinforcing the discriminative power of the selected features and validating the biological relevance of the classification framework.

6) Regression Modeling for Predictive Gene Expression Analysis

To assess the predictive capacity of statistical features on logFC as a continuous variable, multiple regression models were benchmarked. Gradient Boosting and Random Forest regressors achieved the highest test R² scores (90.33% and 90.25%, respectively), outperforming Linear Regression (R² = 85.73%) and Decision Tree (R² = 85.70%). These results indicate that ensemble methods are better suited for modeling nonlinear relationships in gene expression data. The low mean squared error (MSE < 0.38) further supports the reliability of these models in estimating fold change magnitudes.

Scatter plots comparing predicted vs. actual logFC values demonstrated tight clustering along the identity line, especially for Random Forest and Gradient Boosting models. These findings suggest that microarray-derived statistical features can be effectively used to predict gene expression dynamics, offeringa scalable approach for biomarker prioritization and downstream enrichment analysis.



Model Performance Summary

Model	Train R² (%)	Test R² (%)	MSE
Linear Regression	87.0	85.73	0.548
Random Forest	98.2	90.25	0.3743
Decision Tree	100.0	85.7	0.5493
Gradient Boosting	97.26	90.33	0.3715
Support Vector Regressor	88.52	89.26	0.4124
KNN Regressor	93.62	89.79	0.3921

Table 3 : Model Performance Summary

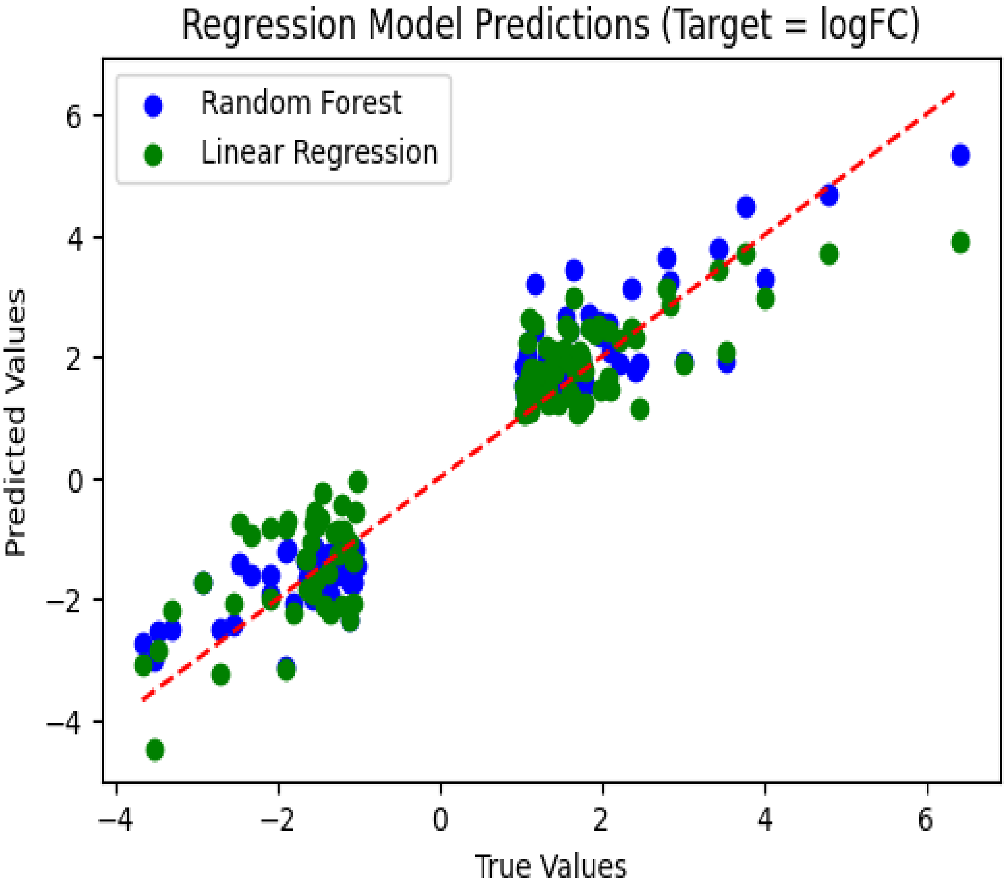


Figure 10 Regression Curve : Comparative regression performance for logFC prediction.

The scatter plot compares predicted versus true log₂ fold change (logFC) values for two regression models: **Random Forest** (blue dots) and **Linear Regression** (green dots). The red dashed line represents the ideal prediction line y=xy = x, where predicted values perfectly match true values.

- **Random Forest predictions** (blue) show tighter clustering around the ideal line, indicating better alignment with true logFC values and superior model performance in capturing nonlinear relationships.
- **Linear Regression predictions** (green) exhibit greater dispersion from the ideal line, especially at extreme logFC values, suggesting limited capacity to model complex gene expression dynamics.
- The proximity of points to the red line reflects prediction accuracy; deviations indicate residual error and model bias.
- The overall spread and density of points suggest that Random Forest provides more robust and generalizable predictions for logFC, likely due to its ensemble structure and ability to capture feature interactions.
- The plot visually confirms that **nonlinear models outperform linear ones** in predicting gene-level differential expression, especially when input features are biologically heterogeneous or non-additive.

7) Unsupervised Clustering and Biological Interpretability

To explore latent structure within the gene expression dataset, DBSCAN clustering was applied following PCA and t-SNE dimensionality reduction. Cluster-wise analysis revealed distinct DEG panels with subtype-specific enrichment potential. Evaluation metrics including Adjusted Rand Index (ARI = 0.4821), Homogeneity (0.7535), and V- Measure (0.5156) confirmed moderate alignment between unsupervised clusters and true regulation labels. Cluster-specific gene lists were extracted for GO and KEGG enrichment, revealing biologically coherent modules such as estrogen signaling, ErbB pathway, and apoptotic cascades.

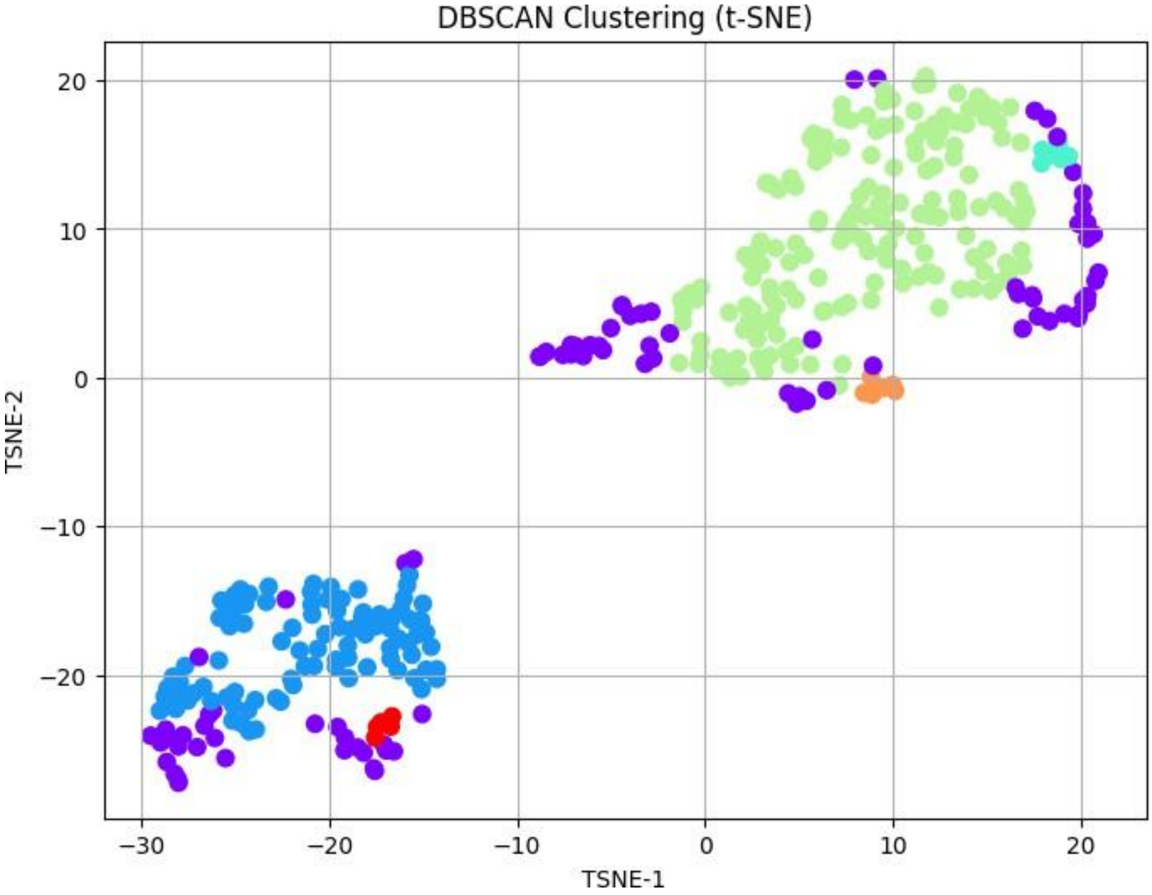


Figure 11 :t-sne Plot

IV. Conclusion

This study presents an integrative framework combining microarray gene expression analysis, differential expression profiling, machine learning modeling, and network-based biomarker prioritization to stratify breast cancer subtypes and identify clinically relevant

prognostic genes. Using datasets from GEO profiled on the Affymetrix GPL570 platform, we successfully differentiated Luminal A and triple-negative breast cancer (TNBC) subtypes through robust statistical and computational pipelines.

Supervised machine learning models—including Random Forest, XGBoost, and SVM— achieved perfect classification accuracy and cross-validation scores, demonstrating the predictive power of microarray-derived statistical features. Feature importance and PCA visualizations reinforced the biological interpretability of the models. Further, protein– protein interaction (PPI) network analysis and hub gene identification via Cytoscape revealed nine survival-significant biomarkers**: EGFR, FN1, COL1A1, BGN, ERBB2, COL5A1, COL5A2, COL10A1, and COL11A1.** These genes were enriched in pathways related to ErbB signaling, extracellular matrix organization, and apoptosis, underscoring their translational relevance in breast cancer prognosis and therapy.

The integration of statistical modeling, machine learning, and network biology offers a reproducible and scalable approach for biomarker discovery. This framework can be extended to other cancer types and omics platforms, supporting precision oncology and individualized treatment strategies.

Data Availability Statement: The microarray datasets are publicly available on the Gene Expression Omnibus (GEO) public database (https://www.ncbi.nlm.nih.gov/geo/, under the accession numbers GSE65194, GSE45827, and GSE42568.

Conflicts of Interest: The authors declare no conflict of interest.

Acknowledgement

I gratefully acknowledges the open-access availability of microarray datasets from the Gene Expression Omnibus (GEO), which enabled comprehensive transcriptomic analysis. Special thanks are extended to the developers of R/Bioconductor packages—**limma**, **GEOquery**, **annotate**, and **hgu133plus2.db**—for providing robust tools that support reproducible and modular data processing. The contributions of **STRING** and **Cytoscape** are also appreciated for their powerful network visualization capabilities and hub gene identification workflows.

Thanks to the maintainers of open-source **GitHub repositories**, for fostering accessible learning environments in **R** and **Python**, and for sharing script-based resources that support benchmarking and reproducible analysis. Gratitude is also expressed to **Venny 2.0** for enabling intuitive Venn diagram generation, and to **Google Colab** for its flexible cloud-based infrastructure that facilitated machine learning implementation without hardware constraints.

Abbreviations
BC Breast Cancer
TCGA The Cancer Genome Atlas GEO Gene Expression Omnibus DEGs Differentially Expressed Genes GO Gene Ontology
PPI Protein-Protein Interaction
KEGG Kyoto Encyclopaedia of Genes and Genomics DE Differentially Expressed

References

[1]. MDPI Diagnostics (2023): Multimodal deep learning model for breast cancer subtype classification integrating imaging and clinical metadata.
[2]. Bioinformatics_GSE65194_Breast_Cancer_Resistance/optimus.R at 1e6290dd248b07bcf0a23635b3b4afccfd623eb1 · tahagill/Bioinformatics_GSE65194_Breast_Cancer_Resistance · GitHub
[3]. https://github.com/tahagill/Bioinformatics_GSE65194_Breast_Cancer_Resistance/ blob/1e6290dd248b07bcf0a23635b3b4afccfd623eb1/optimus.R
[4]. https://github.com/drippypale/microarray-aml
[5]. GitHub - futureomics/Machine-learning-in-drug-discovery_: Machine learning in drug discovery
[6]. Analysis of the microarray gene expression for breast cancer progression after the application modified logistic regression - ScienceDirect
[7]. CancerSubtypes: an R/Bioconductor package for molecular cancer subtype identification, validation and visualization - PubMed
[8]. pgbio99/BRCA-Subtype-Classification_ML – GitHub project: Integrative ML framework for breast cancer subtype classification using GEO microarray data.
[9]. Frontiers in Physiology (2022): Review of ML and deep learning techniques for cancer classification using microarray gene expression.
[10]. HTTPS://WWW.MDPI.COM/1648-9144/59/10/1705
[11]. FUTUREOMICS (DR.NILOFER) · GITHUB
[12]. HTTPS://WWW.NCBI.NLM.NIH.GOV/GEO/