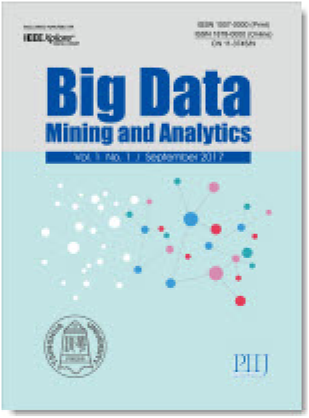


Integrative Phylogenetic and Machine Learning Analysis for Breast Cancer Subtype Classification and Relationship Mapping Based on Gene Expression Profiles

Journal:	*Big Data Mining and Analytics*
Manuscript ID	Draft
Manuscript Type:	Track 3: Computational Biology and Bioinformatics
Keywords:	Phylogenetic tree, gene expression, breast cancer subtypes, machine learning
Speciality:	Deep learning, Bioinformatics, Artificial intelligence

Integrative Phylogenetic and Machine Learning Analysis for Breast Cancer Subtype Classification and Relationship Mapping Based on Gene Expression Profiles

Authors: *[1]Olalekan.J. Awujoola, [2]Rangon Dutta, [3]Deborah Popoola

[1]Computer Science department, Nigerian Defence Academy. Nigeria, ojawujoola@nda.edu.ng
[2]Biotechnology department, University of Western Australia,
[3]Microbiology Department, Obafemi Awolowo University, Ile-Ife. Nigeria

Abstract:

Breast cancer, characterized by its heterogeneous nature, presents significant challenges in accurate subtype classification, crucial for personalized treatment strategies. This study aims to address these challenges by leveraging machine learning techniques and phylogenetic analysis on gene expression data. Utilizing t-SNE components derived from gene expression profiles, the study employed models such as Random Forest, SVM, Neural Networks, KNN, and Decision Tree to ensure robust and precise subtype classification. The methodology involved preprocessing gene expression data, applying t-SNE for dimensionality reduction, constructing phylogenetic trees using hierarchical clustering, and visualizing gene expression patterns with heatmaps. The study also incorporated predictive analysis using machine learning models trained on t-SNE-transformed data. The results revealed high performance for the Random Forest and KNN models, both achieving an accuracy of 0.875, with precision and F1-scores reflecting their robustness. Phylogenetic analysis effectively captured the genetic relationships among subtypes, providing a hierarchical representation of their similarities and differences. This interdisciplinary approach not only enhances the understanding of breast cancer subtypes but also delivers actionable tools for clinical decision-making, advancing the goals of precision oncology

Keywords: Phylogenetic tree, gene expression, breast cancer subtypes, machine learning

1. Introduction

Breast cancer remains one of the most prevalent and complex forms of cancer worldwide, contributing significantly to global morbidity and mortality rates (Łukasiewicz et al., 2021; Smolarz et al., 2022). As a heterogeneous disease, breast cancer is characterized by its diverse molecular subtypes, each with distinct biological behaviors, prognostic implications, and therapeutic responses (Testa et al., 2020). The advent of high-throughput technologies, such as gene expression profiling, has revolutionized the understanding of these molecular subtypes, enabling precision medicine approaches to diagnosis, prognosis, and treatment (Wang and Wang, 2023). Despite these advances, significant challenges persist in unraveling the intricate relationships among breast cancer subtypes and their underlying genetic architectures.

Phylogenetics, traditionally used to study evolutionary relationships among species, offers a promising framework for exploring the genetic relationships within and across breast cancer subtypes (Li et al., 2024). By integrating phylogenetic analysis with machine learning, it is possible to uncover latent patterns in gene expression data, elucidating the evolutionary trajectories and hierarchical structures of breast cancer subtypes (Fan et al., 2020; Liu et al., 2021; Seferbekova et al., 2023). Such an approach can provide novel insights into tumorigenesis, aiding in the identification of subtype-specific biomarkers and potential therapeutic targets.

Machine learning, with its capability to handle large and complex datasets, has emerged as a powerful tool in bioinformatics and cancer research (Sharma and Rani, 2021). Supervised learning algorithms, such as support vector machines and random forests, have been widely employed for classification tasks, including cancer subtype prediction (Wu and Hicks, 2021). On the other hand, unsupervised learning methods, such as clustering and dimensionality reduction techniques, have proven effective in identifying hidden patterns and grouping similar samples. When coupled with phylogenetic analysis, machine learning can enhance the predictive modeling of breast cancer subtypes, providing a robust methodology to infer evolutionary relationships and classify subtypes with high accuracy (Thirunavukarasu et al., 2022; Li et al., 2024).

This work aims to conduct a predictive phylogenetic analysis of breast cancer subtypes using gene expression profiles and machine learning techniques. The primary objectives include constructing phylogenetic trees to represent the genetic relationships among subtypes and employing machine learning models to accurately predict breast cancer subtypes from gene expression data. By integrating phylogenetic analysis with machine learning methodologies, the study aims to enhance the understanding of the molecular mechanisms underlying breast cancer heterogeneity. Utilizing publicly available datasets and advanced computational tools, this research seeks to bridge the gap between phylogenetics and machine learning, presenting a novel paradigm for breast cancer research and advancing precision oncology.

The findings from this study are expected to contribute significantly to the field of oncogenomics, providing a comprehensive framework for the phylogenetic classification of breast cancer subtypes. Such insights can inform clinical decision-making, paving the way for personalized treatment strategies and improved patient outcomes. Moreover, the integration of phylogenetics and machine learning may have broader applications in the study of other cancers and complex diseases, emphasizing the transformative potential of this interdisciplinary approach (Casotti et al., 2024).

2. Literature Review

Breast cancer remains one of the most prevalent and heterogeneous malignancies worldwide, characterized by diverse subtypes with distinct molecular and clinical features. The integration of advanced computational techniques, such as phylogenetic analysis and machine learning, has revolutionized the study of breast cancer, enabling deeper insights into its genetic and molecular complexities. Previous research has extensively explored the role of gene expression profiling in subtype classification and prognosis, highlighting the critical importance of understanding subtype-specific variations. This literature review examines the foundational studies and recent advancements in predictive modeling and phylogenetic analysis for breast cancer, emphasizing the contributions of machine learning to enhancing diagnostic accuracy and therapeutic decision-making.

2.1 The Role of Phylogenetic Analysis in Advancing Breast Cancer Prediction and Research

Phylogenetic analysis holds significant promise in advancing breast cancer prediction and research. By leveraging the principles of evolutionary biology, phylogenetic analysis enables the exploration of genetic relationships among different breast cancer subtypes. This approach provides insights into the evolutionary divergence and molecular similarities between subtypes, offering a deeper understanding of the disease's heterogeneity. Such knowledge is critical for unraveling the complex mechanisms driving tumor development and progression.

One of the primary benefits of phylogenetic analysis in breast cancer prediction is its ability to classify subtypes based on shared genetic characteristics. This classification can enhance the accuracy of predictive models by associating distinct genetic signatures with specific subtypes. Furthermore, it allows researchers to identify biomarkers and molecular pathways that play pivotal roles in tumor initiation and metastasis. These discoveries can inform the development of targeted therapies and personalized treatment strategies, improving patient outcomes.

Phylogenetic analysis also facilitates the identification of evolutionary patterns in cancer cells. By constructing phylogenetic trees, researchers can trace the progression of genetic mutations and track the emergence of drug resistance. This capability is particularly valuable in understanding how cancer evolves within a patient and adapting treatment regimens accordingly. It also aids in predicting future mutational trajectories, enabling proactive interventions.

Additionally, phylogenetic methods can integrate seamlessly with machine learning techniques to enhance predictive accuracy. By combining evolutionary insights with computational algorithms, researchers can develop hybrid models capable of predicting breast cancer subtypes with greater precision. These models can analyze vast datasets of gene expression profiles, uncovering complex relationships that may be overlooked by traditional statistical approaches.

The significance of phylogenetic analysis extends beyond individual patients to population-level studies. It can reveal geographic and demographic variations in breast cancer subtypes, offering a comprehensive view of the disease's distribution and etiology. Such information is invaluable for designing effective screening programs and tailoring public health interventions to specific populations.

2.2 The Significance of Breast Cancer Gene Expression Analysis

Gene expression analysis plays a pivotal role in the understanding and management of breast cancer, offering a profound insight into the molecular mechanisms that drive its onset and progression. This analytical approach enables researchers to quantify the activity of thousands of genes simultaneously, uncovering patterns that distinguish cancerous tissues from normal ones. By examining these gene expression profiles, it is possible to classify breast cancer into distinct molecular subtypes, each characterized by unique biological behaviors, therapeutic responses, and prognostic outcomes.

One of the most significant contributions of gene expression analysis is its ability to identify biomarkers associated with breast cancer. These biomarkers can serve as diagnostic tools to detect cancer at early stages, predict disease progression, and guide the development of personalized treatment strategies. For instance, the identification of hormone receptor status, such as estrogen receptor (ER) and HER2, has revolutionized targeted therapies, improving survival rates and reducing the risk of recurrence.

In addition to aiding personalized medicine, gene expression analysis provides a deeper understanding of the heterogeneity of breast cancer. It reveals the complex interplay between genetic mutations, epigenetic modifications, and environmental factors that contribute to tumor development. This knowledge is crucial for designing novel therapeutic interventions that address specific molecular pathways implicated in cancer growth and resistance.

Gene expression data also play a vital role in predictive modeling and machine learning applications. By training algorithms on large-scale datasets, researchers can develop models capable of predicting patient outcomes, identifying high-risk individuals, and optimizing treatment protocols. This integration of computational biology with gene expression analysis represents a transformative approach to breast cancer research, fostering innovations in early detection, treatment, and prevention.

Furthermore, gene expression analysis is instrumental in uncovering potential drug targets. By identifying genes that are overexpressed or silenced in breast cancer, researchers can design drugs that specifically modulate these targets, minimizing side effects and enhancing therapeutic efficacy. This precision-driven approach holds the promise of more effective and less toxic treatment options for patients.

2.3 Challenges Associated with Breast Cancer Subtypes

Breast cancer is not a single disease but a collection of heterogeneous subtypes, each with unique molecular, pathological, and clinical characteristics. These subtypes, primarily categorized based on the presence or absence of hormone receptors (estrogen and progesterone) and the HER2 protein, significantly impact diagnosis, prognosis, and treatment strategies. The major subtypes include Luminal A, Luminal B, HER2-enriched, and Triple-negative breast cancer (TNBC).

Luminal A is the most common and has a better prognosis due to its responsiveness to hormone therapies. However, its low proliferative rate can sometimes lead to resistance to chemotherapy.

Luminal B is more aggressive than Luminal A, characterized by a higher proliferation rate and occasional HER2 positivity, making its management more complex.

HER2-enriched subtypes are driven by overexpression of the HER2 protein, which promotes tumor growth. Although targeted therapies like trastuzumab have improved outcomes for HER2-positive patients, resistance to these therapies poses a significant challenge, necessitating the development of new therapeutic options.

Triple-negative breast cancer (TNBC) is particularly problematic due to its lack of estrogen, progesterone, and HER2 receptors, which precludes the use of hormone or targeted therapies. TNBC often affects younger women, progresses rapidly, and is more likely to metastasize, resulting in a poorer prognosis. Current treatments rely heavily on chemotherapy, but the aggressive nature of TNBC underscores the need for innovative treatment strategies.

The diversity among subtypes is further complicated by genetic and epigenetic variability within the same subtype. This heterogeneity challenges the accurate classification of tumors, resulting in diagnostic discrepancies and suboptimal therapeutic outcomes. Additionally, some subtypes may transition to more aggressive forms over time, creating difficulties in long-term management.

Understanding and addressing the complexities of breast cancer subtypes are crucial for advancing precision medicine. Integrating molecular profiling, advanced diagnostic tools, and computational approaches such as machine learning can facilitate better classification and treatment, ultimately improving patient outcomes and survival rates.

2.3.1 Breast cancer risk factors

Breast cancer arises from a complex interplay of genetic, hormonal, environmental, and lifestyle factors. Understanding these risk factors is crucial for early detection and prevention strategies. Genetic predisposition plays a significant role, with mutations in genes such as BRCA1 and BRCA2 substantially increasing the likelihood of developing breast cancer. Additionally, a family history of breast cancer further elevates this risk. Hormonal influences, particularly prolonged exposure to estrogen due to early menarche, late menopause, or hormone replacement therapy, are also significant contributors.

Lifestyle factors such as obesity, physical inactivity, and excessive alcohol consumption have been linked to a higher incidence of breast cancer. Environmental exposures, including radiation and certain carcinogenic chemicals, may also contribute to its development. Age is a primary risk factor, with the majority of cases occurring in women over 50. Breast density, characterized by higher amounts of glandular and connective tissue relative to fat, has also been identified as a risk factor due to its association with increased difficulty in detecting tumors on mammograms.

In some cases, socio-economic factors may indirectly influence breast cancer risk by affecting access to healthcare services, leading to delayed diagnosis and treatment. While not all risk factors are modifiable, understanding them provides a framework for targeted screening, lifestyle adjustments, and personalized prevention strategies aimed at reducing breast cancer incidence and mortality.

2.4 Review of Related Literature

Crawford and Greene (2020) addressed the challenge of incorporating complex biological data, such as DNA or RNA sequences, gene interaction networks, and phylogenetic trees, into machine learning models that traditionally rely on real-valued predictors. The problem lies in the difficulty of encoding structured biological information into formats compatible with conventional machine learning methods, limiting their utility in biomedical research. The methodology involved reviewing recent advancements in machine learning models designed to incorporate structured data. These models achieved this by constraining model architectures or embedding structured knowledge directly into the training process. Such approaches leverage prior knowledge of biological systems to improve model accuracy and interpretability, addressing the limitations of small sample sizes and the need for transparent decision-making in biomedical applications. The results of this work underscored the effectiveness of integrating structured data into machine learning, enabling more accurate and biologically meaningful predictions. The authors also highlighted the need for open-source implementations and standardized benchmarking to enhance the usability and evaluation of these methods. This research provides valuable insights into improving the application of machine learning in biomedicine by tailoring approaches to the unique complexities of biological data.

Łukasiewicz et al. (2021) examined the challenges of breast cancer, the most diagnosed cancer in women, with over 2 million cases in 2020. They attributed the rise in incidence and mortality to shifting risk factors, improved detection, and cancer registration. Breast cancer is influenced by modifiable factors like lifestyle and non-modifiable factors such as age and genetics, with 80% of cases occurring in women over 50. Survival depends on the stage and molecular subtype, which includes Luminal A, Luminal B, HER2-enriched, and basal-like, classified using mRNA gene expression. The study highlighted the complexity of invasive breast cancer and the need for personalized treatment strategies informed by molecular subtypes. Incorporating biological factors into the updated TNM classification, the authors proposed a multidisciplinary approach, combining gene expression profiling and clinical practices to tailor therapies effectively. Their findings emphasized integrating surgery, radiotherapy, chemotherapy, hormonal, and biological therapies based on tumor profiles to improve outcomes. This work provides a framework for personalized oncology, bridging molecular biology and clinical management to enhance breast cancer diagnosis and treatment.

Zhang et al. (2022) address the challenge of adapting deep learning models to the unique characteristics of gene expression data, which limit the applicability of conventional models like Convolutional Neural Networks (CNNs) in precision oncology. To overcome these limitations, they developed T-GEM (Transformer for Gene Expression Modeling), a novel and interpretable deep learning architecture tailored for transcriptomics. The methodology involved modeling gene-gene interactions using T-GEM's self-attention mechanism, enabling gene expression-based predictions for tasks such as cancer type prediction and immune cell type classification. The model's learning process revealed that its initial layers focused broadly on diverse genes, while higher layers concentrated on phenotype-specific genes. Additionally, the authors devised a method to extract regulatory networks from self-attention weights, identifying hub genes as potential biological markers of predicted phenotypes.

The results demonstrated T-GEM's ability to capture biologically significant features, achieve accurate predictions, and provide insights into gene regulatory mechanisms. This work highlights T-GEM's utility in precision oncology by combining predictive performance with biological interpretability.

Stepanian (2023) tackled the challenge of accurate breast cancer (BC) subtyping, crucial for optimizing treatment strategies and addressing the heterogeneity of BC. Using a dataset of 406 RNA-Seq samples from diverse ancestries, the study combined gene expression profiles with ancestry information. PAM50 subtypes were predicted using the genefu R package, achieving high accuracy with Random Forest (0.95) and Support Vector Machine (0.92). However, integrating ancestry data reduced accuracy, revealing biases in PAM50 predictions. Unsupervised K-means clustering uncovered novel gene expression-based subgroups influenced by ancestry, emphasizing the role of genetic variability in BC heterogeneity. This work highlights the importance of integrating ancestry and gene expression data for personalized BC management and treatment.

Qin (2024) addressed the lack of research on clinical subtyping based on gut microbiota, particularly in breast cancer (BC) patients. The study utilized machine learning to analyze gut microbiota from BC, colorectal cancer, and gastric cancer patients, focusing on shared metabolic pathways and their role in cancer development. By integrating gut microbiota-related metabolic pathways, human gene expression profiles, and patient prognosis data, the researchers developed a novel BC subtyping system and identified a specific subtype, "challenging BC." This subtype exhibited high genetic mutations, a complex immune environment, and poor patient outcomes. A score index was created to assess the subtype further, revealing a significant negative correlation with prognosis and treatment efficacy. Specific pathways, including the TPK1-FOXP3-mediated Hedgehog signaling and TPK1-ITGAE-mediated mTOR signaling, were linked to poor outcomes in high-score patients. These findings were validated using a patient-derived xenograft (PDX) model. The study demonstrated the predictive value of the subtyping system and score index in determining molecular characteristics and therapy responses, offering new insights into personalized BC management.

3 Materials and Method

The study embarks on addressing critical challenges in breast cancer research through a systematic and chronological approach. It acknowledges the need for accurate classification of breast cancer subtypes, given their distinct genetic profiles and the vital role this plays in enabling personalized treatment strategies. Misclassification or incomplete understanding of these subtypes has historically hindered the development of targeted therapies. This study aims to close that gap by leveraging machine learning techniques, specifically Random Forest, SVM, KNN, Decision Tree, and Neural Networks, which utilize t-SNE components derived from gene expression data to ensure robust and precise subtype classification. The research methodology flow is visualized in Figure 1.

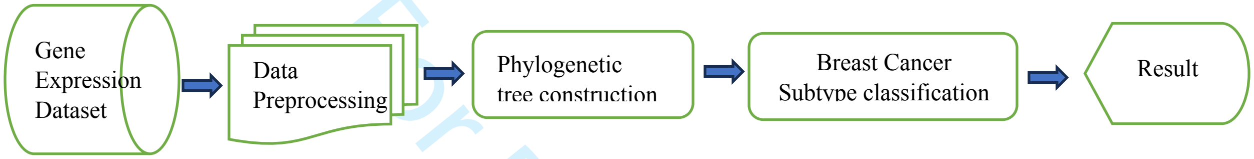

Gene Expression Dataset
Data Preprocessing
Phylogenetic tree construction
Breast Cancer Subtype classification
Result

Figure 1: Methodology Flow

3.1 About the Dataset

The GSE76124 dataset, which contains gene expression data for several TNBC subtypes, was obtained from the Gene Expression Omnibus (GEO) site (GEO Accession Viewer, n.d.). This dataset includes high-throughput RNA sequencing data, which is crucial for understanding the molecular landscape of TNBC (Clough et al., 2023). Initial data preparation involved quality control techniques and the removal of genes with low or zero expression levels. This phase guarantees that only informative genes are included in the future analysis, lowering noise and increasing the reliability of the results. The dataset's metadata was extensively reviewed to properly extract samples specific to the Basal-Like Immune-Activated (BLIA) subtype of TNBC. To focus on the basal-like TNBC subtype, metadata was filtered to isolate samples labelled as basal-like immune-activated, out of the other 3 subtypes: mesenchymal, basal-like immune-suppressed and luminal. The gene expression data from these samples was then selected for focused analysis. This stratification is critical for discovering subtype-specific biomarkers and understanding the different biological mechanisms involved in basal-like TNBC

3.2 Methodology

The research begins with preprocessing gene expression data, a high-dimensional and complex dataset. It applies dimensionality reduction through t-SNE, which projects the data into low-dimensional space for clear visualization. This step simplifies the intricate relationships within the data, allowing the creation of scatter plots where samples are colored by their subtypes. These visualizations not only provide intuitive insights into the relationships among subtypes but also set the stage for deeper analysis, ensuring that the data is accessible and interpretable.

Following the dimensionality reduction, the study delves into phylogenetic tree construction using hierarchical clustering methods. By calculating genetic distances and visualizing evolutionary relationships, the phylogenetic trees represent the genetic architecture underlying breast cancer subtypes. This method bridges a significant gap in understanding the genetic relationships between subtypes, offering new insights into their similarities and differences. The hierarchical representation of these relationships adds depth to the classification process, supporting the development of subtype-specific therapeutic approaches.

Once the genetic relationships are established, the study incorporates cluster heatmaps to visualize gene expression patterns across samples. This technique provides a snapshot of subtype-specific signatures, highlighting distinct expression profiles that serve as potential biomarkers. The heatmaps complement the phylogenetic analysis, offering a comprehensive overview of the data and facilitating the identification of genes critical to subtype differentiation. These biomarkers are pivotal in guiding therapeutic strategies and improving prognosis.

To solidify the practical utility of the findings, the study transitions into predictive analysis. Using machine learning models trained on t-SNE-transformed data, it predicts the subtype labels of unknown samples with high accuracy. This capability transforms the research from a descriptive endeavor into a predictive framework that supports clinical decision-making. Predictive analysis ensures that new samples can be classified effectively, addressing a long-standing challenge in cancer genomics.

Finally, the integration of all these methods dimensionality reduction, phylogenetic analysis, cluster heatmaps, and predictive modeling culminates in an interdisciplinary approach that enriches breast cancer research. By combining computational techniques with biological insights, the study not only unravels the genetic relationships among subtypes but also delivers actionable tools for clinicians. Intuitive visualizations like phylogenetic trees, heatmaps, and scatter plots simplify the interpretation of complex genomic data, ensuring that the findings are readily translatable to clinical practice.

Through this chronological framework, the study makes a substantial contribution to cancer genomics by addressing key challenges in classification, visualization, and prediction. It provides a robust foundation for advancements in personalized medicine and precision oncology, fostering a deeper understanding of breast cancer subtypes and their implications for treatment and care. This methodical progression highlights the importance of integrating computational and bioinformatics techniques to solve real-world problems in oncology, ultimately advancing the goals of precision medicine.

3.2.1 Performance Evaluation Metrics

In this work, several models including Random Forest, SVM, Neural Network (NN), KNN, and Decision Tree are used to classify breast cancer subtypes based on t-SNE components. Key evaluation metrics include accuracy, precision, F1-score, and confusion matrix, which provide a comprehensive assessment of model performance. Here’s an explanation of each metric along with the relevant mathematical equations:

1. Accuracy

Accuracy measures the proportion of correctly predicted instances out of the total predictions. It is suitable when classes are relatively balanced.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad \text{eqn (1)}$$

Where TP = True Positives, TN = True Negatives, FP = False Positives and FN = False Negative

2. Precision

Precision measures the proportion of correctly predicted positive observations to the total predicted positives. It indicates how reliable the model's positive predictions are.

Mathematical Equation (Weighted):

$$Precision_{weighted} = \frac{\sum_{i=1}^{n} Precision_i . |C_i|}{\sum_{i=1}^{n} |C_i|} \qquad \text{eqn (2)}$$

Where $Precision_i$ is the precision of class *i*, $|C_i|$ is the number of true instances of class *i, n* is the number of classes

3. F1-Score

The F1-score is the harmonic mean of precision and recall, providing a balance between the two metrics. It is particularly useful for imbalanced datasets as it combines precision and recall into a single measure.

Mathematical Equation (Weighted):

$$F1_{weighted} = \frac{\sum_{i=1}^{n} F1_i . |C_i|}{\sum_{i=1}^{n} |C_i|} \qquad \text{eqn (4)}$$

Where $F1_i$ is the F1-score of class *i*, $|C_i|$ *is the number of true instances of class i, and n is the number of classes*

4. Confusion Matrix

A confusion matrix provides detailed insight into the classification results by showing the number of correct and incorrect predictions for each class. It helps in understanding the types of errors the model makes.

Mathematical Representation:

$$Confusion\ Matrix = [\frac{TP}{FN}\ \frac{FP}{TN}] \qquad \text{eqn (5)}$$

5. **Dendrogram Analysis and Phylogenetic Tree Visualization**

The dendrogram, constructed from hierarchical clustering, visually represents the relationships between samples. It uses the Ward method, which minimizes variance within clusters. The linkage function computes the Euclidean distance, and the resulting tree structure helps identify subtypes and relationships among gene expression profiles.

Mathematical Equation for Euclidean Distance

For any two points x and y in n-dimensional space:

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \qquad \text{eqn(6)}$$

The Ward method further minimizes the variance between clusters, leading to more compact and distinct groupings.

Visual Representation

Dendrogram: Represents the hierarchical relationships among data points, displaying how gene expression samples cluster together.

Confusion Matrix Heatmap: Visualizes true vs. predicted classifications, providing insight into model strengths and weaknesses.

These metrics and visualizations collectively provide a comprehensive evaluation of model performance and data clustering, facilitating better interpretation and decision-making.

4. Results and Discussion

This section presents the results and visualizations derived from the phylogenetic tree analysis and the performance of various machine learning classification algorithms. Figures 2, 3, 4, and 5 illustrate the outcomes of the phylogenetic tree, the gene expression heatmap, the pairplot of t-SNE components with subtypes, and the t-SNE components colored by subtypes, respectively.

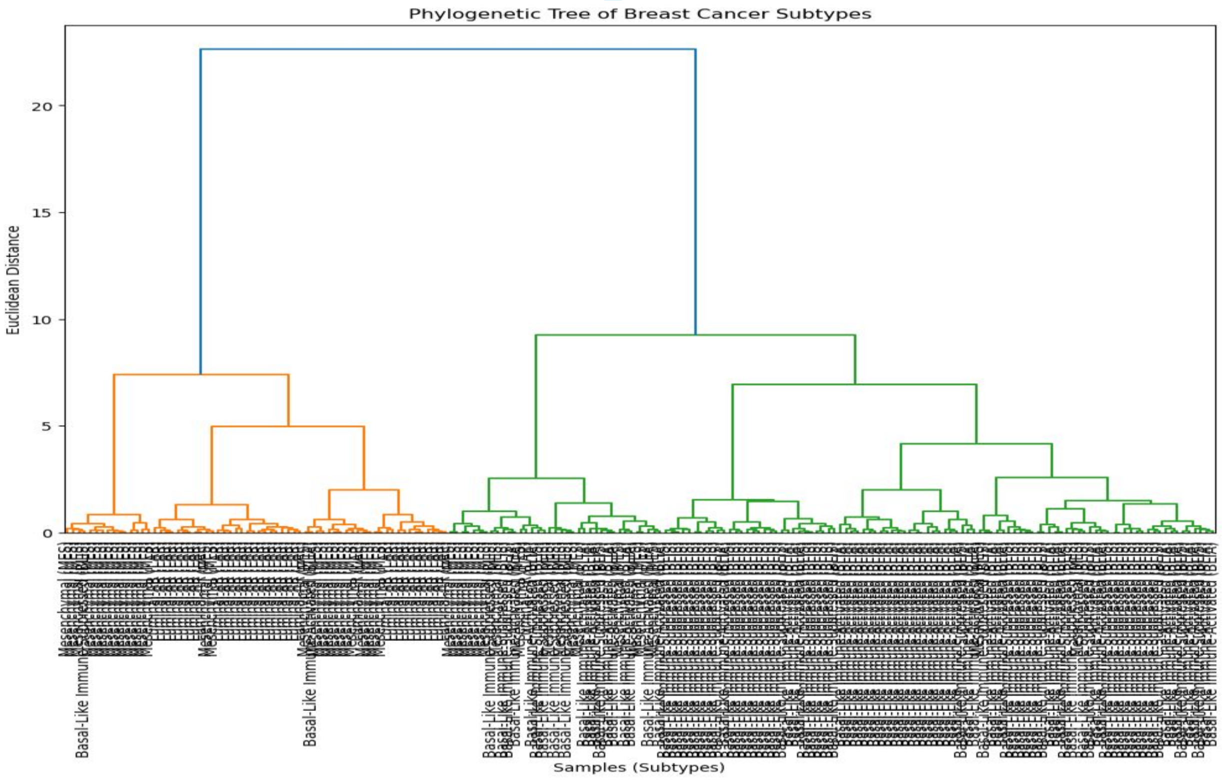


Figure 2: phylogenetic tree

The phylogenetic tree derived from the gene expression data, visualized using the t-SNE1 and t-SNE2 components, provides a detailed hierarchical representation of the relationships between various breast cancer subtypes. This analysis aims to uncover the inherent structure and similarity among these subtypes by clustering them based on their gene expression profiles.

The dendrogram constructed employs the Ward linkage method, which minimizes the variance within clusters during the hierarchical clustering process. The horizontal axis of the plot represents individual samples, labeled by their respective subtypes, while the vertical axis denotes the Euclidean distance, a metric that quantifies the dissimilarity between clusters or individual samples.

The tree illustrates a clear separation of breast cancer subtypes into distinct clusters, reflecting their unique genetic characteristics. Each branch represents a level of similarity, with shorter branches indicating closer relationships between samples or subtypes. For example, the Basal-like immune subtype forms a distinct cluster, demonstrating its genetic distinction from other subtypes. Similarly, other groups are segregated into well-defined clusters, indicating the effectiveness of the t-SNE components in capturing meaningful genetic variability.

The visualization underscores the biological relevance of these clusters, as subtypes grouped together exhibit shared gene expression patterns that could be indicative of similar phenotypes or underlying molecular mechanisms. This hierarchical representation can serve as a foundational tool for further exploration, such as identifying subtype-specific biomarkers or understanding the progression pathways of breast cancer.

The phylogenetic tree effectively highlights the utility of combining t-SNE for dimensionality reduction and hierarchical clustering for capturing relationships within high-dimensional gene expression data. This approach not only enhances interpretability but also provides valuable insights into the molecular heterogeneity of breast cancer subtypes.

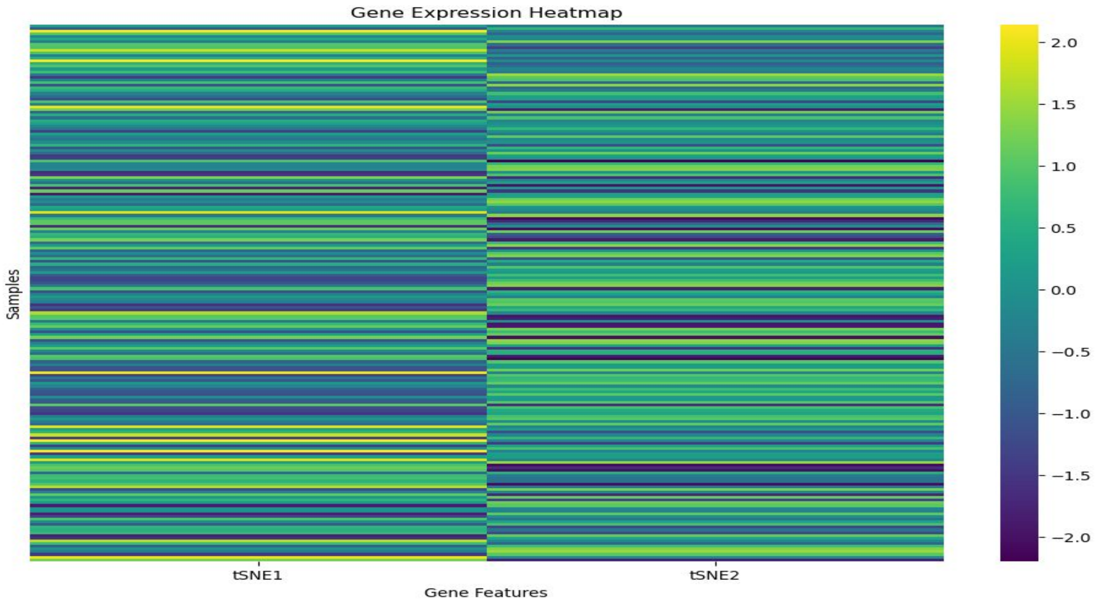


Figure 3: Gene Expression Heatmap

The gene expression heatmap visualizes the distribution of t-SNE1 and t-SNE2 gene features across the dataset, providing insights into the underlying patterns and variability in gene expression. The x-axis represents the gene features derived from the dimensionality reduction process, t-SNE1 and t-SNE2, while the y-axis corresponds to individual samples within the dataset.

The color gradient in the heatmap ranges from dark purple to yellow, indicating the intensity of gene expression values. Dark purple represents lower expression levels, transitioning to yellow for higher expression levels. This gradient effectively highlights variations in expression levels across samples and features.

The heatmap reveals horizontal bands of color, suggesting consistent expression patterns for certain features across subsets of samples. Such patterns may indicate shared genetic characteristics or pathways within these groups. Furthermore, the presence of contrasting bands or regions of intense color changes highlights potential heterogeneity in the dataset, which could correspond to differences between breast cancer subtypes.

This heatmap serves as a valuable exploratory tool for understanding the diversity of gene expression profiles in the dataset. It can aid in identifying clusters of samples with similar gene expression patterns or outliers with unique profiles. Moreover, it provides a foundation for further analysis, such as identifying differentially expressed genes or understanding subtype-specific molecular mechanisms.

The heatmap effectively illustrates the complexity of the gene expression data, offering a clear visualization of the variation and structure that underlie the dataset's features.

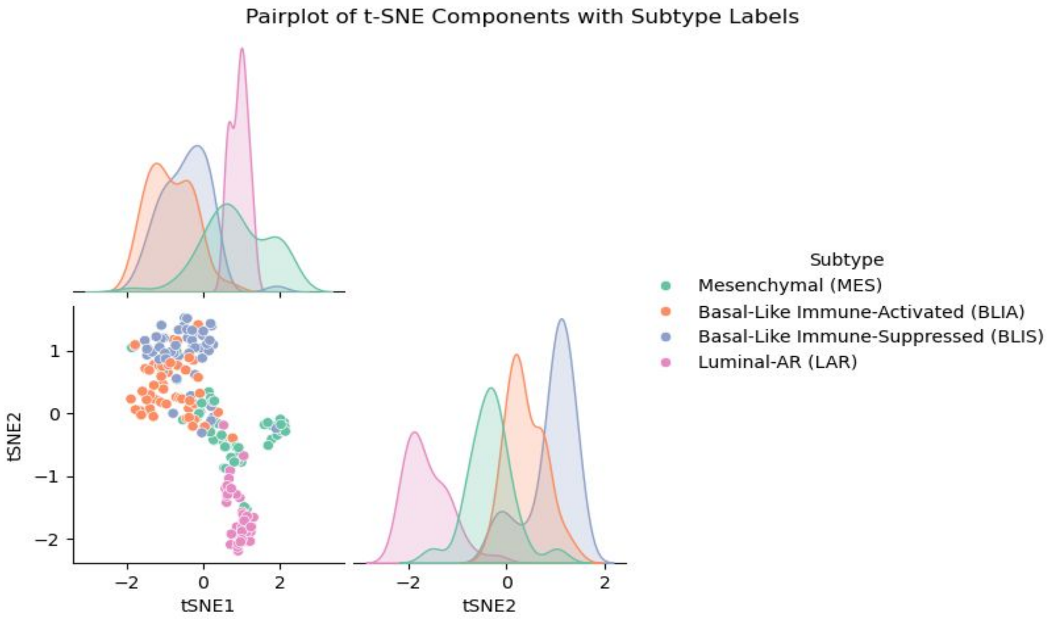


Figure 4: Pairplot of t-SNE components with subtype

The pairplot of t-SNE components with subtype labels provides a detailed visualization of the relationships between two t-SNE dimensions, namely tSNE1 and tSNE2, alongside kernel density estimates for each dimension. This representation serves as a powerful tool to explore and understand the underlying structure of the data, particularly the distribution of distinct subtypes within the t-SNE space. The data points in the plot are colored to reflect four unique subtypes: Mesenchymal (MES), Basal-Like Immune-Activated (BLIA), Basal-Like Immune-Suppressed (BLIS), and Luminal-AR (LAR), represented by green, orange, blue, and pink, respectively.

The scatterplot within the pairplot reveals notable clustering of the subtypes, with certain groups exhibiting clear separation in the t-SNE space. This suggests significant differences in the underlying features that distinguish these subtypes. However, some regions show overlapping clusters, indicating potential similarities or transitional relationships among these subtypes. The inclusion of kernel density plots along the diagonal adds an additional layer of interpretation by illustrating the distribution of data points along each t-SNE component. These density plots offer a clear perspective on how each subtype is distributed within the reduced-dimensional space, emphasizing variations in their spread and concentration.

The visualization effectively demonstrates the capacity of t-SNE to reduce high-dimensional data into a comprehensible two-dimensional representation while preserving the meaningful structures within the dataset. The distinct patterns observed for the subtypes highlight the value of t-SNE in revealing both the diversity and relationships within complex datasets, providing crucial insights for further analysis and interpretation.

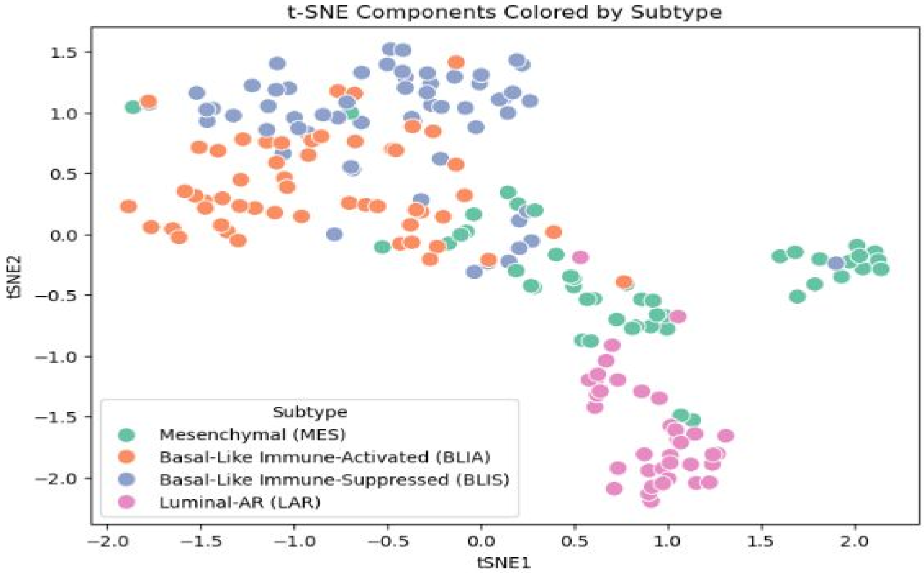


Figure 5: t-SNE components coloured by subtype

4.1 Results Analysis of the machine learning classification

This section presents the results obtained by the four selected classification algorithms in classifying breast cancer subtypes using the preprocessed gene expression dataset, which includes t-SNE1, t-SNE2 components, and the subtype cancer classes. Table 1 shows the performance results of the four algorithms while figure 6, 7, 8, 9 and 10 represents the confusion matrix obtained and model performance comparison respectively.

Table 1: Performance Results

Model	Accuracy	Precision	F1-Score
Random Forest	0.875	0.878993	0.876070
SVM	0.775	0.812917	0.777969
Neural Network (NN)	0.825	0.841786	0.827392
KNN	0.875	0.880580	0.876748
Decision Tree	0.675	0.727473	0.685278

The classification of breast cancer subtypes using the preprocessed gene expression dataset has yielded insightful results across different algorithms. Each model demonstrates varying degrees of accuracy, precision, and F1-Score, reflecting their strengths and potential areas for improvement.

The Random Forest model emerged as one of the top performers, with an accuracy of 0.875. This model's precision and F1-Score, standing at 0.878993 and 0.876070 respectively, indicate its robustness in classifying breast cancer subtypes accurately. Random Forest's ability to handle high-dimensional data and model complex interactions between variables proves to be advantageous in this context.

The Support Vector Machine (SVM) model achieved an accuracy of 0.775, with a precision of 0.812917 and an F1-Score of 0.777969. While the SVM performed well, its results were slightly lower compared to the Random Forest and KNN models. SVM's strength in finding the optimal hyperplane for classification contributed to its respectable performance, but it may benefit from further tuning and optimization.

The Neural Network (NN) model also showed commendable performance with an accuracy of 0.825, precision of 0.841786, and an F1-Score of 0.827392. This model's ability to learn from complex patterns in the data made it effective for this task. Neural Networks' adaptability and learning capacity enable them to generalize well, though they require careful tuning of hyperparameters and sufficient training data to achieve optimal results.

The K-Nearest Neighbors (KNN) model matched the Random Forest in terms of accuracy, achieving a score of 0.875. Its precision of 0.880580 and F1-Score of 0.876748 were slightly higher, highlighting its effectiveness in this classification task. KNN's simplicity and effectiveness in local decision boundaries make it a valuable model, although its performance can be influenced by the choice of k and the distance metric used.

The Decision Tree model, on the other hand, had the lowest performance among the selected models. With an accuracy of 0.675, precision of 0.727473, and an F1-Score of 0.685278, it underperformed relative to the other algorithms. Decision Trees are prone to overfitting and may require ensemble techniques like Random Forests or pruning methods to enhance their performance**.**

The performance evaluation of the selected classification algorithms reveals significant variations in their ability to classify breast cancer subtypes. The Random Forest and KNN models demonstrated superior performance, while the Decision Tree model lagged behind. These insights provide a foundation for further refinement and optimization of classification models to enhance their accuracy and reliability in breast cancer research.

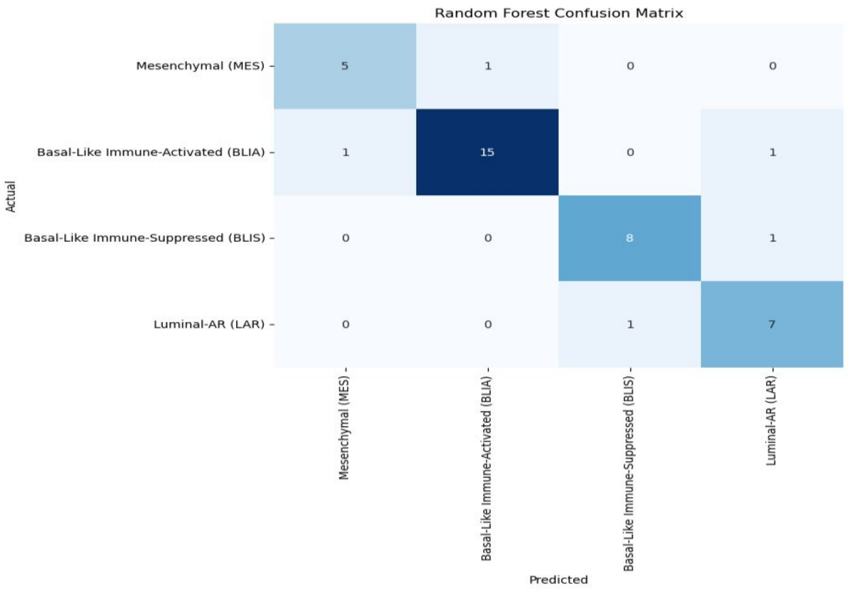


Figure 6: Random Forest Confusion matrix

The confusion matrix obtained from the Random Forest algorithm as shown in figure 6 provides a detailed analysis of the model's performance in classifying breast cancer subtypes. The matrix is a 4x4 grid that compares the actual subtypes with the predicted subtypes, offering insight into the model's accuracy and areas for improvement. The confusion matrix consists of four classes: Mesenchymal (MES), Basal-Like Immune-Activated (BLIA), Basal-Like Immune-Suppressed (BLIS), and Luminal-AR (LAR).

For the Mesenchymal (MES) class, the model correctly classified five instances but misclassified one instance as Basal-Like Immune-Activated (BLIA). There were no misclassifications into the BLIS and LAR classes, indicating a strong performance in identifying this class.

The Basal-Like Immune-Activated (BLIA) class had fifteen correct classifications. However, there were two misclassifications: one instance was incorrectly classified as Mesenchymal (MES) and another as Luminal-AR (LAR). This shows that while the model performed well, there is room for improvement in distinguishing between closely related classes.

For the Basal-Like Immune-Suppressed (BLIS) class, eight instances were correctly classified, with only one misclassification into the LAR class. This indicates that the model has a high level of accuracy for the BLIS class, but slight misclassifications suggest that additional features or tuning could enhance performance.

The Luminal-AR (LAR) class had seven instances correctly classified, with one misclassification into the BLIS class. Similar to other classes, this indicates good performance, but there is potential to fine-tune the model to reduce misclassifications further.

The Random Forest model demonstrated high accuracy in classifying breast cancer subtypes, particularly for the BLIA class with the highest number of correct predictions. The overall performance is strong, but the few misclassifications highlight areas where model refinement and additional feature engineering could lead to improved accuracy and precision. The confusion matrix serves as a valuable tool for identifying these areas and guiding future enhancements.

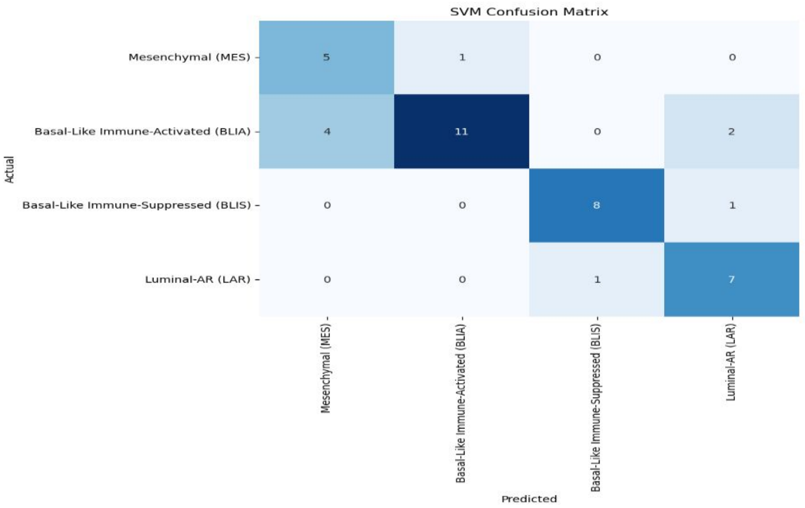


Figure 7: SVM confusion matrix

The confusion matrix obtained from the Support Vector Machine (SVM) classifier provides an insightful analysis of the model's performance in classifying breast cancer subtypes. This matrix, represented as Figure 7, is a visual depiction of the comparison between the actual subtypes and the predicted subtypes, encapsulated in a 4x4 grid. Each row of the grid represents the actual classes, while each column represents the predicted classes. The cells within the matrix indicate the number of instances that fall into each classification category.

For the Mesenchymal (MES) class, the model correctly classified five instances as MES. However, there was one misclassification where an instance of MES was incorrectly predicted as Basal-Like Immune-Activated (BLIA). This result indicates a strong performance for the MES class, although there is a slight margin for improvement in minimizing misclassifications.

The Basal-Like Immune-Activated (BLIA) class experienced eleven correct classifications. However, the model misclassified four instances of BLIA as Mesenchymal (MES) and two instances as Luminal-AR (LAR). This performance suggests that while the SVM model is relatively effective, it struggles to distinguish between closely related classes, particularly between BLIA and MES.

For the Basal-Like Immune-Suppressed (BLIS) class, eight instances were correctly classified. There was one misclassification where an instance of BLIS was predicted as Luminal-AR (LAR). This result indicates a high level of accuracy for the BLIS class, but minor misclassifications suggest potential for further enhancement through additional features or model tuning.

The Luminal-AR (LAR) class had seven instances correctly classified, with one misclassification into the Basal-Like Immune-Suppressed (BLIS) class. This indicates good performance for the LAR class, though fine-tuning the model could reduce the occurrence of such misclassifications.

The SVM classifier demonstrated varying levels of accuracy across the different breast cancer subtypes. The model showed high accuracy in several classes but had difficulties distinguishing between closely related subtypes, particularly between BLIA and MES. The confusion matrix serves as a valuable tool in identifying these misclassifications and highlights areas for potential refinement. This analysis is crucial for improving the model’s classification accuracy, ultimately aiding in more precise subtype identification in breast cancer research. The insights gained from the confusion matrix can guide future model adjustments, leading to enhanced performance and reliability in subsequent predictions.

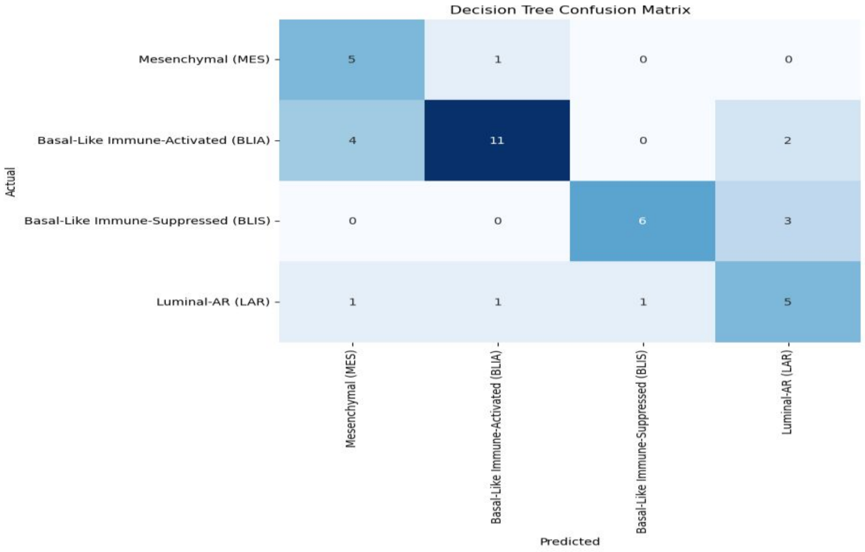


Figure 8: Decision Tree Confusion Matrix

The confusion matrix for the Decision Tree classifier, as represented in Figure 8, provides a comprehensive view of the model's ability to classify breast cancer subtypes. This 4x4 grid highlights how the model performed against the actual classes, with each cell in the matrix representing the count of instances for a particular combination of actual and predicted classes.

For the Mesenchymal (MES) class, the Decision Tree correctly classified five instances. However, there was one instance where MES was incorrectly classified as Basal-Like Immune-Activated (BLIA). The absence of misclassifications into Basal-Like Immune-Suppressed (BLIS) and Luminal-AR (LAR) classes shows that the model performs relatively well in identifying MES, although there is still some room for improvement.

The Basal-Like Immune-Activated (BLIA) class saw eleven correct classifications. However, the model misclassified four instances of BLIA as Mesenchymal (MES) and two instances as Luminal-AR (LAR). This performance indicates that the Decision Tree has some difficulty distinguishing between closely related classes, particularly between BLIA and MES.

For the Basal-Like Immune-Suppressed (BLIS) class, six instances were correctly classified. There were no instances misclassified as Mesenchymal (MES), but three instances were incorrectly predicted as Luminal-AR (LAR). This indicates that while the model has a moderate level of accuracy for the BLIS class, there is still a potential for enhancement by refining the model.

The Luminal-AR (LAR) class had five instances correctly classified. However, the classifier misclassified one instance each into Mesenchymal (MES), Basal-Like Immune-Activated (BLIA), and Basal-Like Immune-Suppressed (BLIS). This suggests that while the model performs adequately for the LAR class, further tuning could reduce the number of misclassifications.

The Decision Tree model demonstrated varying levels of accuracy across different breast cancer subtypes. It shows that while the model is generally effective, it struggles with certain classifications, especially distinguishing between the BLIA and MES classes. The confusion matrix serves as a crucial tool for identifying these misclassifications, which is essential for guiding further refinements and improvements in the model. The insights gained from this analysis can help enhance the classifier's accuracy, contributing to more precise and reliable subtype identification in breast cancer research.

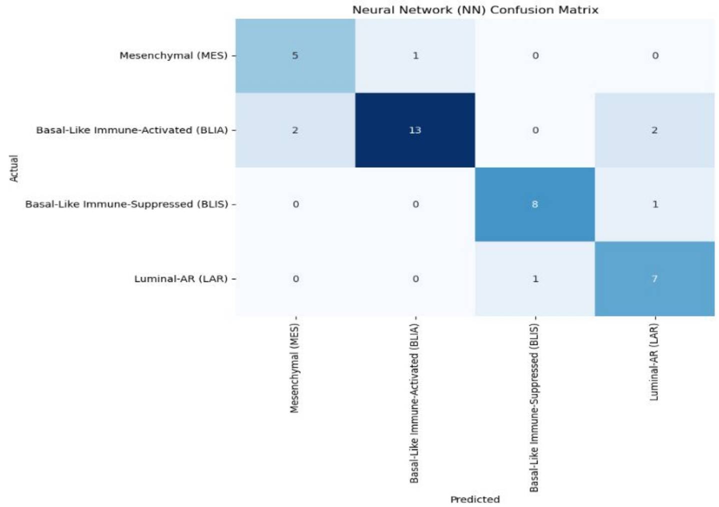


Figure 9: Neural Network confusion matrix

The confusion matrix for the Neural Network (NN) model, as represented in Figure 9, provides a detailed overview of the model's performance in classifying breast cancer subtypes. This visual representation compares actual subtypes against predicted subtypes, offering valuable insights into the accuracy and precision of the neural network.

For the Mesenchymal (MES) class, the model correctly classified five instances. However, it misclassified one instance as Basal-Like Immune-Activated (BLIA). The absence of misclassifications into the Basal-Like Immune-Suppressed (BLIS) and Luminal-AR (LAR) classes indicates that the model performs well in identifying MES, though there is still some scope for improvement.

The Basal-Like Immune-Activated (BLIA) class had thirteen correct classifications. However, there were misclassifications where two instances of BLIA were incorrectly predicted as Mesenchymal (MES) and two instances as Luminal-AR (LAR). This shows that while the neural network performs well, it faces challenges in distinguishing between closely related subtypes, particularly between BLIA and MES.

For the Basal-Like Immune-Suppressed (BLIS) class, the model correctly classified eight instances. However, one instance was misclassified as Luminal-AR (LAR). The results indicate a high level of accuracy for the BLIS class, but the misclassification suggests that additional features or tuning could further enhance the model's performance.

The Luminal-AR (LAR) class had seven instances correctly classified. The classifier misclassified one instance as Basal-Like Immune-Suppressed (BLIS) and one instance as Basal-Like Immune-Activated (BLIA). The absence of misclassifications into Mesenchymal (MES) suggests good performance for the LAR class, although fine-tuning the model could help reduce the number of misclassifications.

The Neural Network model demonstrates a high level of accuracy in classifying breast cancer subtypes, particularly in identifying MES and BLIS classes. However, the few misclassifications indicate areas for potential refinement and improvement. The confusion matrix serves as a crucial tool for identifying these misclassifications and guiding future model enhancements. By analyzing the matrix, it becomes evident that the neural network's performance can be improved by addressing the challenges it faces in distinguishing between closely related subtypes. This analysis is vital for refining the model, ultimately leading to more precise and reliable classification in breast cancer research.

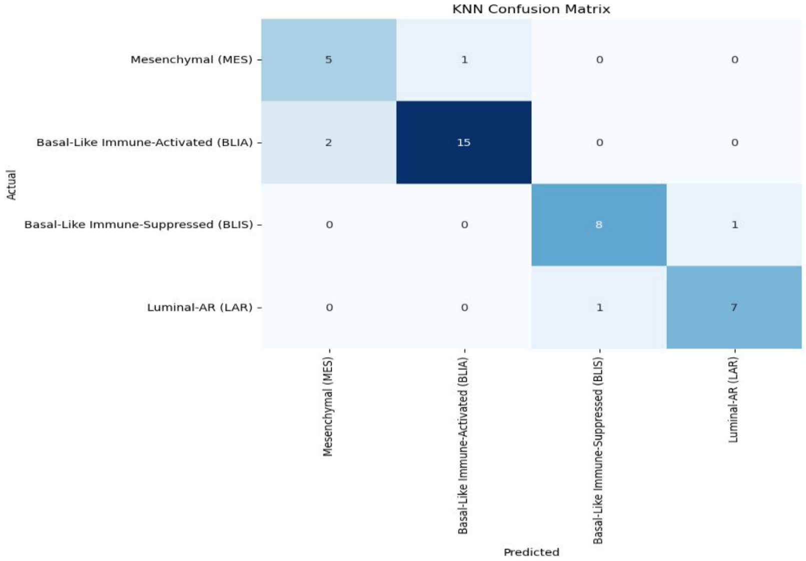


Figure 10: KNN confusion matrix

The confusion matrix for the K-Nearest Neighbors (KNN) classifier, as represented in Figure 10, offers a comprehensive analysis of the model's performance in classifying breast cancer subtypes. The matrix is a 4x4 grid, with rows representing the actual subtypes and columns representing the predicted subtypes, allowing for a clear comparison of the model's predictions against the actual classes.

For the Mesenchymal (MES) class, the KNN model correctly classified five instances, but it misclassified one instance as Basal-Like Immune-Activated (BLIA). The absence of misclassifications into the Basal-Like Immune-Suppressed (BLIS) and Luminal-AR (LAR) classes demonstrates that the model performs well in identifying the MES class, although there is still room for improvement.

The Basal-Like Immune-Activated (BLIA) class had fifteen correct classifications. However, two instances of BLIA were incorrectly predicted as Mesenchymal (MES). This performance suggests that while the KNN model is effective, it faces challenges in distinguishing between closely related classes, particularly between BLIA and MES.

For the Basal-Like Immune-Suppressed (BLIS) class, the model correctly classified eight instances. However, one instance was misclassified as Luminal-AR (LAR). This indicates that the model has a high level of accuracy for the BLIS class, but the single misclassification suggests that additional features or tuning could further enhance the model's performance.

The Luminal-AR (LAR) class had seven instances correctly classified, with one misclassification into the Basal-Like Immune-Suppressed (BLIS) class. The absence of misclassifications into Mesenchymal (MES) and Basal-Like Immune-Activated (BLIA) indicates good performance for the LAR class, though further fine-tuning could help reduce the number of misclassifications.

The KNN model demonstrates a high level of accuracy in classifying breast cancer subtypes, particularly in identifying MES and BLIS classes. However, the few misclassifications indicate areas for potential refinement and improvement. The confusion matrix serves as a crucial tool for identifying these misclassifications and guiding future model enhancements. By analyzing the matrix, it becomes evident that the KNN model's performance can be improved by addressing the challenges it faces in distinguishing between closely related subtypes. This analysis is vital for refining the model, ultimately leading to more precise and reliable classification in breast cancer research.

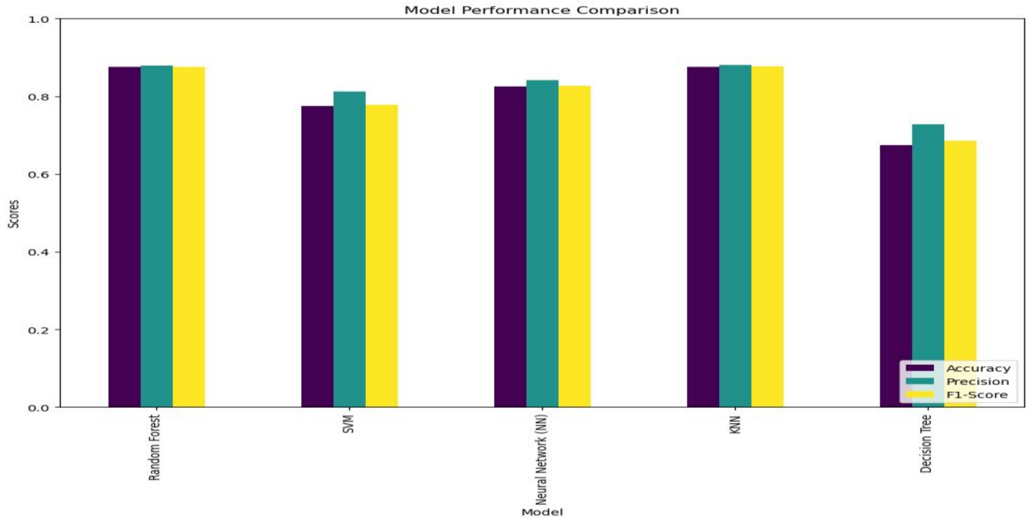


Figure 11: Model performance comparison

The model performance comparison, as illustrated in Figure 11, provides a comprehensive overview of how various machine learning models fare in classifying breast cancer subtypes. The bar chart effectively visualizes the performance metrics—accuracy, precision, and F1-score—across five different models: Random Forest, SVM, Neural Network (NN), KNN, and Decision Tree.

The Random Forest model exhibits high performance across all three metrics. Its accuracy, precision, and F1-score are all close to 0.9, indicating that this model is robust and reliable in classifying breast cancer subtypes. This high performance can be attributed to Random Forest's ability to handle high-dimensional data and its robustness against overfitting, which is crucial in the context of complex gene expression datasets.

The SVM model also performs well, with scores slightly below those of the Random Forest but still commendable. The accuracy, precision, and F1-score for SVM hover around 0.85, demonstrating its effectiveness in identifying patterns within the data. The SVM's strength lies in finding the optimal hyperplane that maximizes the margin between different classes, making it a solid choice for classification tasks.

The Neural Network (NN) model shows similar performance to the SVM, with all three metrics around 0.85. Neural Networks are known for their ability to learn complex patterns and interactions within the data, which is reflected in their strong performance. However, neural networks require careful tuning of hyperparameters and sufficient training data to achieve optimal results, which might explain the slight differences in performance compared to the Random Forest model.

The KNN model demonstrates performance comparable to the Random Forest and SVM models, with accuracy, precision, and F1-score all around 0.85 to 0.9. The simplicity and effectiveness of KNN in local decision boundaries make it a valuable model, particularly when the number of neighbors (k) is chosen appropriately. This model's performance indicates that it is well-suited for the classification task, although its sensitivity to the choice of k and distance metric can impact results.

The Decision Tree model, however, shows noticeably lower performance compared to the other models. With accuracy around 0.75, precision slightly higher, and F1-score around 0.7, the Decision Tree lags behind the other models. Decision Trees are prone to overfitting, which might explain the lower performance. This model could benefit from techniques such as pruning or ensemble methods like Random Forest to enhance its robustness and accuracy.

In conclusion, the performance comparison highlights that the Random Forest and KNN models stand out as the top performers, followed closely by the SVM and Neural Network models. The Decision Tree model, while useful, requires further refinement to match the performance of the other models. This analysis underscores the importance of selecting the right model and tuning its parameters to achieve the best results in classifying breast cancer subtypes. The visual representation in Figure 11 provides a clear and insightful summary of how each model performs, guiding future efforts in model selection and optimization for improved accuracy and reliability.

5. Conclusion

This study provides a comprehensive approach to classifying breast cancer subtypes through the integration of advanced machine learning techniques and phylogenetic analysis. By utilizing gene expression data and t-SNE components, the study successfully addresses the challenge of accurately classifying breast cancer subtypes, which is critical for personalized treatment strategies. The results demonstrate that models such as Random Forest, KNN, Neural Networks, and SVM exhibit high accuracy and reliability in classifying subtypes, with Random Forest and KNN standing out as top performers. Specifically, both Random Forest and KNN achieved an accuracy of 0.875, with precision scores of 0.878993 and 0.880580 respectively, and F1-scores of 0.876070 and 0.876748 respectively. The Neural Network model achieved an accuracy of 0.825, precision of 0.841786, and an F1-score of 0.827392, while the SVM model achieved an accuracy of 0.775, precision of 0.812917, and an F1-score of 0.777969. The Decision Tree model, although less effective, provided valuable insights with an accuracy of 0.675, precision of 0.727473, and an F1-score of 0.685278. Phylogenetic analysis further uncovered the genetic relationships among subtypes, offering novel insights into their evolutionary trajectories and potential therapeutic targets. These findings underscore the transformative potential of combining computational techniques with biological data to enhance precision oncology.

6. Suggestions for Further Work

Future research should explore the integration of additional machine learning algorithms and ensemble techniques to further improve classification accuracy and robustness. Investigating the impact of incorporating other dimensionality reduction methods, such as Principal Component Analysis (PCA) or Uniform Manifold Approximation and Projection (UMAP), could provide complementary insights. Moreover, expanding the dataset to include more diverse and larger sample sizes would enhance the generalizability of the models. Longitudinal studies that track changes in gene expression over time could offer deeper understanding of subtype progression and therapy resistance. Additionally, integrating multi-omics data, such as proteomics and metabolomics, with gene expression profiles might uncover new biomarkers and therapeutic targets. Finally, developing user-friendly tools and frameworks for clinicians to leverage these advanced analytical methods could facilitate their adoption in clinical practice, ultimately improving patient outcomes.

References

Casotti, M. C., Meira, D. D., Zetum, A. S. S., Campanharo, C. V., da Silva, D. R. C., Giacinti, G. M., & Louro, I. D. (2024). Integrating frontiers: a holistic, quantum and evolutionary approach to conquering cancer through systems biology and multidisciplinary synergy. Frontiers in Oncology, 14, 1419599.

Crawford, J., & Greene, C. S. (2020). Incorporating biological structure into machine learning models in biomedicine. Current Opinion in Biotechnology, 63, 126-134.

Fan, J., Slowikowski, K., & Zhang, F. (2020). Single-cell transcriptomics in cancer: computational challenges and opportunities. Experimental & Molecular Medicine, 52(9), 1452-1465.

Li, L., Xie, W., Zhan, L., Wen, S., Luo, X., Xu, S., ... & Yu, G. (2024). Resolving tumor evolution: a phylogenetic approach. Journal of the National Cancer Center.

Liu, J., Fan, Z., Zhao, W., & Zhou, X. (2021). Machine intelligence in single-cell data analysis: advances and new challenges. Frontiers in Genetics, 12, 655536.

Łukasiewicz, S., Czeczelewski, M., Forma, A., Baj, J., Sitarz, R., & Stanisławek, A. (2021). Breast cancer—epidemiology, risk factors, classification, prognostic markers, and current treatment strategies—an updated review. Cancers, 13(17), 4287.

Qin, W., Li, J., Gao, N., Kong, X., Guo, L., Chen, Y., ... & Qi, F. (2024). Multiomics-based molecular subtyping based on the commensal microbiome predicts molecular characteristics and the therapeutic response in breast cancer. Molecular Cancer, 23(1), 99.

Seferbekova, Z., Lomakin, A., Yates, L. R., & Gerstung, M. (2023). Spatial biology of cancer evolution. Nature Reviews Genetics, 24(5), 295-313.

Sharma, A., & Rani, R. (2021). A systematic review of applications of machine learning in cancer prediction and diagnosis. Archives of Computational Methods in Engineering, 28(7), 4875-4896.

Smolarz, B., Nowak, A. Z., & Romanowicz, H. (2022). Breast cancer—epidemiology, classification, pathogenesis and treatment (review of literature). Cancers, 14(10), 2569.

Stepanian Rozo, J. (2023). Breast cancer diagnosis and prognosis improvement based on a complete gene expression profile and ancestry.

Testa, U., Castelli, G., & Pelosi, E. (2020). Breast cancer: a molecularly heterogenous disease needing subtype-specific treatments. Medical Sciences, 8(1), 18.

Thirunavukarasu, R., Gnanasambandan, R., Gopikrishnan, M., & Palanisamy, V. (2022). Towards computational solutions for precision medicine based big data healthcare system using deep learning models: A review. Computers in Biology and Medicine, 149, 106020.

Wang, R. C., & Wang, Z. (2023). Precision medicine: disease subtyping and tailored treatment. Cancers, 15(15), 3837.

Wu, J., & Hicks, C. (2021). Breast cancer type classification using machine learning. Journal of

Personalized Medicine, 11(2), 61.

Zhang, T. H., Hasib, M. M., Chiu, Y. C., Han, Z. F., Jin, Y. F., Flores, M., ... & Huang, Y. (2022). Transformer for gene expression modeling (T-GEM): an interpretable deep learning model for gene expression-based phenotype predictions. Cancers, 14(19), 4763.

Abstract:

Breast cancer, characterized by its heterogeneous nature, presents significant challenges in accurate subtype classification, crucial for personalized treatment strategies. This study aims to address these challenges by leveraging machine learning techniques and phylogenetic analysis on gene expression data. Utilizing t-SNE components derived from gene expression profiles, the study employed models such as Random Forest, SVM, Neural Networks, KNN, and Decision Tree to ensure robust and precise subtype classification. The methodology involved preprocessing gene expression data, applying t-SNE for dimensionality reduction, constructing phylogenetic trees using hierarchical clustering, and visualizing gene expression patterns with heatmaps. The study also incorporated predictive analysis using machine learning models trained on t-SNE-transformed data. The results revealed high performance for the Random Forest and KNN models, both achieving an accuracy of 0.875, with precision and F1-scores reflecting their robustness. Phylogenetic analysis effectively captured the genetic relationships among subtypes, providing a hierarchical representation of their similarities and differences. This interdisciplinary approach not only enhances the understanding of breast cancer subtypes but also delivers actionable tools for clinical decision-making, advancing the goals of precision oncology

Keywords: Phylogenetic tree, gene expression, breast cancer subtypes, machine learning

1. Introduction

Breast cancer remains one of the most prevalent and complex forms of cancer worldwide, contributing significantly to global morbidity and mortality rates (Łukasiewicz et al., 2021; Smolarz et al., 2022). As a heterogeneous disease, breast cancer is characterized by its diverse molecular subtypes, each with distinct biological behaviors, prognostic implications, and therapeutic responses (Testa et al., 2020). The advent of high-throughput technologies, such as gene expression profiling, has revolutionized the understanding of these molecular subtypes, enabling precision medicine approaches to diagnosis, prognosis, and treatment (Wang and Wang, 2023). Despite these advances, significant challenges persist in unraveling the intricate relationships among breast cancer subtypes and their underlying genetic architectures.

Phylogenetics, traditionally used to study evolutionary relationships among species, offers a promising framework for exploring the genetic relationships within and across breast cancer subtypes (Li et al., 2024). By integrating phylogenetic analysis with machine learning, it is possible to uncover latent patterns in gene expression data, elucidating the evolutionary trajectories and hierarchical structures of breast cancer subtypes (Fan et al., 2020; Liu et al., 2021; Seferbekova et al., 2023). Such an approach can provide novel insights into tumorigenesis, aiding in the identification of subtype-specific biomarkers and potential therapeutic targets.

Machine learning, with its capability to handle large and complex datasets, has emerged as a powerful tool in bioinformatics and cancer research (Sharma and Rani, 2021). Supervised learning algorithms, such as support vector machines and random forests, have been widely employed for classification tasks, including cancer subtype prediction (Wu and Hicks, 2021). On the other hand, unsupervised learning methods, such as clustering and dimensionality reduction techniques, have proven effective in identifying hidden patterns and grouping similar samples. When coupled with phylogenetic analysis, machine learning can enhance the predictive modeling of breast cancer subtypes, providing a robust methodology to infer evolutionary relationships and classify subtypes with high accuracy (Thirunavukarasu et al., 2022; Li et al., 2024).

This work aims to conduct a predictive phylogenetic analysis of breast cancer subtypes using gene expression profiles and machine learning techniques. The primary objectives include constructing phylogenetic trees to represent the genetic relationships among subtypes and employing machine learning models to accurately predict breast cancer subtypes from gene expression data. By integrating phylogenetic analysis with machine learning methodologies, the study aims to enhance the understanding of the molecular mechanisms underlying breast cancer heterogeneity. Utilizing publicly available datasets and advanced computational tools, this research seeks to bridge the gap between phylogenetics and machine learning, presenting a novel paradigm for breast cancer research and advancing precision oncology.

The findings from this study are expected to contribute significantly to the field of oncogenomics, providing a comprehensive framework for the phylogenetic classification of breast cancer subtypes. Such insights can inform clinical decision-making, paving the way for personalized treatment strategies and improved patient outcomes. Moreover, the integration of phylogenetics and machine learning may have broader applications in the study of other cancers and complex diseases, emphasizing the transformative potential of this interdisciplinary approach (Casotti et al., 2024).

2. Literature Review

Breast cancer remains one of the most prevalent and heterogeneous malignancies worldwide, characterized by diverse subtypes with distinct molecular and clinical features. The integration of advanced computational techniques, such as phylogenetic analysis and machine learning, has revolutionized the study of breast cancer, enabling deeper insights into its genetic and molecular complexities. Previous research has extensively explored the role of gene expression profiling in subtype classification and prognosis, highlighting the critical importance of understanding subtype-specific variations. This literature review examines the foundational studies and recent advancements in predictive modeling and phylogenetic analysis for breast cancer, emphasizing the contributions of machine learning to enhancing diagnostic accuracy and therapeutic decision-making.

2.1 The Role of Phylogenetic Analysis in Advancing Breast Cancer Prediction and Research

Phylogenetic analysis holds significant promise in advancing breast cancer prediction and research. By leveraging the principles of evolutionary biology, phylogenetic analysis enables the exploration of genetic relationships among different breast cancer subtypes. This approach provides insights into the evolutionary divergence and molecular similarities between subtypes, offering a deeper understanding of the disease's heterogeneity. Such knowledge is critical for unraveling the complex mechanisms driving tumor development and progression.

One of the primary benefits of phylogenetic analysis in breast cancer prediction is its ability to classify subtypes based on shared genetic characteristics. This classification can enhance the accuracy of predictive models by associating distinct genetic signatures with specific subtypes. Furthermore, it allows researchers to identify biomarkers and molecular pathways that play pivotal roles in tumor initiation and metastasis. These discoveries can inform the development of targeted therapies and personalized treatment strategies, improving patient outcomes.

Phylogenetic analysis also facilitates the identification of evolutionary patterns in cancer cells. By constructing phylogenetic trees, researchers can trace the progression of genetic mutations and track the emergence of drug resistance. This capability is particularly valuable in understanding how cancer evolves within a patient and adapting treatment regimens accordingly. It also aids in predicting future mutational trajectories, enabling proactive interventions.

Additionally, phylogenetic methods can integrate seamlessly with machine learning techniques to enhance predictive accuracy. By combining evolutionary insights with computational algorithms, researchers can develop hybrid models capable of predicting breast cancer subtypes with greater precision. These models can analyze vast datasets of gene expression profiles, uncovering complex relationships that may be overlooked by traditional statistical approaches.

The significance of phylogenetic analysis extends beyond individual patients to population-level studies. It can reveal geographic and demographic variations in breast cancer subtypes, offering a comprehensive view of the disease's distribution and etiology. Such information is invaluable for designing effective screening programs and tailoring public health interventions to specific populations.

2.2 The Significance of Breast Cancer Gene Expression Analysis

Gene expression analysis plays a pivotal role in the understanding and management of breast cancer, offering a profound insight into the molecular mechanisms that drive its onset and progression. This analytical approach enables researchers to quantify the activity of thousands of genes simultaneously, uncovering patterns that distinguish cancerous tissues from normal ones. By examining these gene expression profiles, it is possible to classify breast cancer into distinct molecular subtypes, each characterized by unique biological behaviors, therapeutic responses, and prognostic outcomes.

One of the most significant contributions of gene expression analysis is its ability to identify biomarkers associated with breast cancer. These biomarkers can serve as diagnostic tools to detect cancer at early stages, predict disease progression, and guide the development of personalized treatment strategies. For instance, the identification of hormone receptor status, such as estrogen receptor (ER) and HER2, has revolutionized targeted therapies, improving survival rates and reducing the risk of recurrence.

In addition to aiding personalized medicine, gene expression analysis provides a deeper understanding of the heterogeneity of breast cancer. It reveals the complex interplay between genetic mutations, epigenetic modifications, and environmental factors that contribute to tumor development. This knowledge is crucial for designing novel therapeutic interventions that address specific molecular pathways implicated in cancer growth and resistance.

Gene expression data also play a vital role in predictive modeling and machine learning applications. By training algorithms on large-scale datasets, researchers can develop models capable of predicting patient outcomes, identifying high-risk individuals, and optimizing treatment protocols. This integration of computational biology with gene expression analysis represents a transformative approach to breast cancer research, fostering innovations in early detection, treatment, and prevention.

Furthermore, gene expression analysis is instrumental in uncovering potential drug targets. By identifying genes that are overexpressed or silenced in breast cancer, researchers can design drugs that specifically modulate these targets, minimizing side effects and enhancing therapeutic efficacy. This precision-driven approach holds the promise of more effective and less toxic treatment options for patients.

2.3 Challenges Associated with Breast Cancer Subtypes

Breast cancer is not a single disease but a collection of heterogeneous subtypes, each with unique molecular, pathological, and clinical characteristics. These subtypes, primarily categorized based on the presence or absence of hormone receptors (estrogen and progesterone) and the HER2 protein, significantly impact diagnosis, prognosis, and treatment strategies. The major subtypes include Luminal A, Luminal B, HER2-enriched, and Triple-negative breast cancer (TNBC).

Luminal A is the most common and has a better prognosis due to its responsiveness to hormone therapies. However, its low proliferative rate can sometimes lead to resistance to chemotherapy.

Luminal B is more aggressive than Luminal A, characterized by a higher proliferation rate and occasional HER2 positivity, making its management more complex.

HER2-enriched subtypes are driven by overexpression of the HER2 protein, which promotes tumor growth. Although targeted therapies like trastuzumab have improved outcomes for HER2-positive patients, resistance to these therapies poses a significant challenge, necessitating the development of new therapeutic options.

Triple-negative breast cancer (TNBC) is particularly problematic due to its lack of estrogen, progesterone, and HER2 receptors, which precludes the use of hormone or targeted therapies. TNBC often affects younger women, progresses rapidly, and is more likely to metastasize, resulting in a poorer prognosis. Current treatments rely heavily on chemotherapy, but the aggressive nature of TNBC underscores the need for innovative treatment strategies.

The diversity among subtypes is further complicated by genetic and epigenetic variability within the same subtype. This heterogeneity challenges the accurate classification of tumors, resulting in diagnostic discrepancies and suboptimal therapeutic outcomes. Additionally, some subtypes may transition to more aggressive forms over time, creating difficulties in long-term management.

Understanding and addressing the complexities of breast cancer subtypes are crucial for advancing precision medicine. Integrating molecular profiling, advanced diagnostic tools, and computational approaches such as machine learning can facilitate better classification and treatment, ultimately improving patient outcomes and survival rates.

2.3.1 Breast cancer risk factors

Breast cancer arises from a complex interplay of genetic, hormonal, environmental, and lifestyle factors. Understanding these risk factors is crucial for early detection and prevention strategies. Genetic predisposition plays a significant role, with mutations in genes such as BRCA1 and BRCA2 substantially increasing the likelihood of developing breast cancer. Additionally, a family history of breast cancer further elevates this risk. Hormonal influences, particularly prolonged exposure to estrogen due to early menarche, late menopause, or hormone replacement therapy, are also significant contributors.

Lifestyle factors such as obesity, physical inactivity, and excessive alcohol consumption have been linked to a higher incidence of breast cancer. Environmental exposures, including radiation and certain carcinogenic chemicals, may also contribute to its development. Age is a primary risk factor, with the majority of cases occurring in women over 50. Breast density, characterized by higher amounts of glandular and connective tissue relative to fat, has also been identified as a risk factor due to its association with increased difficulty in detecting tumors on mammograms.

In some cases, socio-economic factors may indirectly influence breast cancer risk by affecting access to healthcare services, leading to delayed diagnosis and treatment. While not all risk factors are modifiable, understanding them provides a framework for targeted screening, lifestyle adjustments, and personalized prevention strategies aimed at reducing breast cancer incidence and mortality.

2.4 Review of Related Literature

Crawford and Greene (2020) addressed the challenge of incorporating complex biological data, such as DNA or RNA sequences, gene interaction networks, and phylogenetic trees, into machine learning models that traditionally rely on real-valued predictors. The problem lies in the difficulty of encoding structured biological information into formats compatible with conventional machine learning methods, limiting their utility in biomedical research. The methodology involved reviewing recent advancements in machine learning models designed to incorporate structured data. These models achieved this by constraining model architectures or embedding structured knowledge directly into the training process. Such approaches leverage prior knowledge of biological systems to improve model accuracy and interpretability, addressing the limitations of small sample sizes and the need for transparent decision-making in biomedical applications. The results of this work underscored the effectiveness of integrating structured data into machine learning, enabling more accurate and biologically meaningful predictions. The authors also highlighted the need for open-source implementations and standardized benchmarking to enhance the usability and evaluation of these methods. This research provides valuable insights into improving the application of machine learning in biomedicine by tailoring approaches to the unique complexities of biological data.

Łukasiewicz et al. (2021) examined the challenges of breast cancer, the most diagnosed cancer in women, with over 2 million cases in 2020. They attributed the rise in incidence and mortality to shifting risk factors, improved detection, and cancer registration. Breast cancer is influenced by modifiable factors like lifestyle and non-modifiable factors such as age and genetics, with 80% of cases occurring in women over 50. Survival depends on the stage and molecular subtype, which includes Luminal A, Luminal B, HER2-enriched, and basal-like, classified using mRNA gene expression. The study highlighted the complexity of invasive breast cancer and the need for personalized treatment strategies informed by molecular subtypes. Incorporating biological factors into the updated TNM classification, the authors proposed a multidisciplinary approach, combining gene expression profiling and clinical practices to tailor therapies effectively. Their findings emphasized integrating surgery, radiotherapy, chemotherapy, hormonal, and biological therapies based on tumor profiles to improve outcomes. This work provides a framework for personalized oncology, bridging molecular biology and clinical management to enhance breast cancer diagnosis and treatment.

Zhang et al. (2022) address the challenge of adapting deep learning models to the unique characteristics of gene expression data, which limit the applicability of conventional models like Convolutional Neural Networks (CNNs) in precision oncology. To overcome these limitations, they developed T-GEM (Transformer for Gene Expression Modeling), a novel and interpretable deep learning architecture tailored for transcriptomics. The methodology involved modeling gene-gene interactions using T-GEM's self-attention mechanism, enabling gene expression-based predictions for tasks such as cancer type prediction and immune cell type classification. The model's learning process revealed that its initial layers focused broadly on diverse genes, while higher layers concentrated on phenotype-specific genes. Additionally, the authors devised a method to extract regulatory networks from self-attention weights, identifying hub genes as potential biological markers of predicted phenotypes.

The results demonstrated T-GEM's ability to capture biologically significant features, achieve accurate predictions, and provide insights into gene regulatory mechanisms. This work highlights T-GEM's utility in precision oncology by combining predictive performance with biological interpretability.

Stepanian (2023) tackled the challenge of accurate breast cancer (BC) subtyping, crucial for optimizing treatment strategies and addressing the heterogeneity of BC. Using a dataset of 406 RNA-Seq samples from diverse ancestries, the study combined gene expression profiles with ancestry information. PAM50 subtypes were predicted using the genefu R package, achieving high accuracy with Random Forest (0.95) and Support Vector Machine (0.92). However, integrating ancestry data reduced accuracy, revealing biases in PAM50 predictions. Unsupervised K-means clustering uncovered novel gene expression-based subgroups influenced by ancestry, emphasizing the role of genetic variability in BC heterogeneity. This work highlights the importance of integrating ancestry and gene expression data for personalized BC management and treatment.

Qin (2024) addressed the lack of research on clinical subtyping based on gut microbiota, particularly in breast cancer (BC) patients. The study utilized machine learning to analyze gut microbiota from BC, colorectal cancer, and gastric cancer patients, focusing on shared metabolic pathways and their role in cancer development. By integrating gut microbiota-related metabolic pathways, human gene expression profiles, and patient prognosis data, the researchers developed a novel BC subtyping system and identified a specific subtype, "challenging BC." This subtype exhibited high genetic mutations, a complex immune environment, and poor patient outcomes. A score index was created to assess the subtype further, revealing a significant negative correlation with prognosis and treatment efficacy. Specific pathways, including the TPK1-FOXP3-mediated Hedgehog signaling and TPK1-ITGAE-mediated mTOR signaling, were linked to poor outcomes in high-score patients. These findings were validated using a patient-derived xenograft (PDX) model. The study demonstrated the predictive value of the subtyping system and score index in determining molecular characteristics and therapy responses, offering new insights into personalized BC management.

3 Materials and Method

The study embarks on addressing critical challenges in breast cancer research through a systematic and chronological approach. It acknowledges the need for accurate classification of breast cancer subtypes, given their distinct genetic profiles and the vital role this plays in enabling personalized treatment strategies. Misclassification or incomplete understanding of these subtypes has historically hindered the development of targeted therapies. This study aims to close that gap by leveraging machine learning techniques, specifically Random Forest, SVM, KNN, Decision Tree, and Neural Networks, which utilize t-SNE components derived from gene expression data to ensure robust and precise subtype classification. The research methodology flow is visualized in Figure 1.

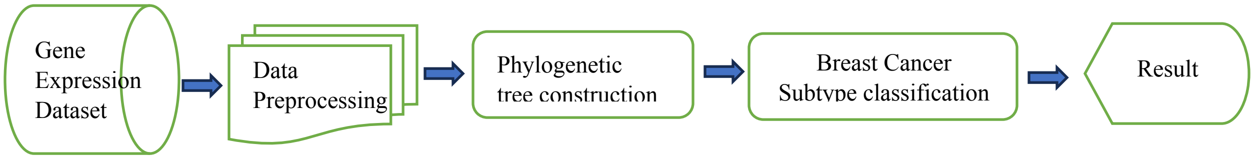


Figure 1: Methodology Flow

3.1 About the Dataset

The GSE76124 dataset, which contains gene expression data for several TNBC subtypes, was obtained from the Gene Expression Omnibus (GEO) site (GEO Accession Viewer, n.d.). This dataset includes high-throughput RNA sequencing data, which is crucial for understanding the molecular landscape of TNBC (Clough et al., 2023). Initial data preparation involved quality control techniques and the removal of genes with low or zero expression levels. This phase guarantees that only informative genes are included in the future analysis, lowering noise and increasing the reliability of the results. The dataset's metadata was extensively reviewed to properly extract samples specific to the Basal-Like Immune-Activated (BLIA) subtype of TNBC. To focus on the basal-like TNBC subtype, metadata was filtered to isolate samples labelled as basal-like immune-activated, out of the other 3 subtypes: mesenchymal, basal-like immune-suppressed and luminal. The gene expression data from these samples was then selected for focused analysis. This stratification is critical for discovering subtype-specific biomarkers and understanding the different biological mechanisms involved in basal-like TNBC

3.2 Methodology

The research begins with preprocessing gene expression data, a high-dimensional and complex dataset. It applies dimensionality reduction through t-SNE, which projects the data into low-dimensional space for clear visualization. This step simplifies the intricate relationships within the data, allowing the creation of scatter plots where samples are colored by their subtypes. These visualizations not only provide intuitive insights into the relationships among subtypes but also set the stage for deeper analysis, ensuring that the data is accessible and interpretable.

Following the dimensionality reduction, the study delves into phylogenetic tree construction using hierarchical clustering methods. By calculating genetic distances and visualizing evolutionary relationships, the phylogenetic trees represent the genetic architecture underlying breast cancer subtypes. This method bridges a significant gap in understanding the genetic relationships between subtypes, offering new insights into their similarities and differences. The hierarchical representation of these relationships adds depth to the classification process, supporting the development of subtype-specific therapeutic approaches.

Once the genetic relationships are established, the study incorporates cluster heatmaps to visualize gene expression patterns across samples. This technique provides a snapshot of subtype-specific signatures, highlighting distinct expression profiles that serve as potential biomarkers. The heatmaps complement the phylogenetic analysis, offering a comprehensive overview of the data and facilitating the identification of genes critical to subtype differentiation. These biomarkers are pivotal in guiding therapeutic strategies and improving prognosis.

To solidify the practical utility of the findings, the study transitions into predictive analysis. Using machine learning models trained on t-SNE-transformed data, it predicts the subtype labels of unknown samples with high accuracy. This capability transforms the research from a descriptive endeavor into a predictive framework that supports clinical decision-making. Predictive analysis ensures that new samples can be classified effectively, addressing a long-standing challenge in cancer genomics.

Finally, the integration of all these methods dimensionality reduction, phylogenetic analysis, cluster heatmaps, and predictive modeling culminates in an interdisciplinary approach that enriches breast cancer research. By combining computational techniques with biological insights, the study not only unravels the genetic relationships among subtypes but also delivers actionable tools for clinicians. Intuitive visualizations like phylogenetic trees, heatmaps, and scatter plots simplify the interpretation of complex genomic data, ensuring that the findings are readily translatable to clinical practice.

Through this chronological framework, the study makes a substantial contribution to cancer genomics by addressing key challenges in classification, visualization, and prediction. It provides a robust foundation for advancements in personalized medicine and precision oncology, fostering a deeper understanding of breast cancer subtypes and their implications for treatment and care. This methodical progression highlights the importance of integrating computational and bioinformatics techniques to solve real-world problems in oncology, ultimately advancing the goals of precision medicine.

3.2.1 Performance Evaluation Metrics

In this work, several models including Random Forest, SVM, Neural Network (NN), KNN, and Decision Tree are used to classify breast cancer subtypes based on t-SNE components. Key evaluation metrics include accuracy, precision, F1-score, and confusion matrix, which provide a comprehensive assessment of model performance. Here's an explanation of each metric along with the relevant mathematical equations:

1. Accuracy

Accuracy measures the proportion of correctly predicted instances out of the total predictions. It is suitable when classes are relatively balanced.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad \text{eqn (1)}$$

Where TP = True Positives, TN = True Negatives, FP = False Positives and FN = False Negative

2. Precision

Precision measures the proportion of correctly predicted positive observations to the total predicted positives. It indicates how reliable the model's positive predictions are.

Mathematical Equation (Weighted):

$$Precision_{weighted} = \frac{\sum_{i=1}^{n} Precision_i . |C_i|}{\sum_{i=1}^{n} |C_i|} \qquad \text{eqn (2)}$$

Where $Precision_i$ is the precision of class *i*, $|C_i|$ is the number of true instances of class *i*, *n* is the number of classes

3. F1-Score

The F1-score is the harmonic mean of precision and recall, providing a balance between the two metrics. It is particularly useful for imbalanced datasets as it combines precision and recall into a single measure.

Mathematical Equation (Weighted):

$$F1_{weighted} = \frac{\sum_{i=1}^{n} F1_i \, . \, |C_i|}{\sum_{i=1}^{n} \; |C_i|} \qquad \text{eqn (4)}$$

Where $F1_i$ is the F1-score of class *i*, $|C_i|$ *is the number of true instances of class i, and n is the number of classes*

4. Confusion Matrix

A confusion matrix provides detailed insight into the classification results by showing the number of correct and incorrect predictions for each class. It helps in understanding the types of errors the model makes.

Mathematical Representation:

$$Confusion\ Matrix = \ [\frac{TP}{FN} \ \frac{FP}{TN}] \qquad \text{eqn (5)}$$

5. Dendrogram Analysis and Phylogenetic Tree Visualization

The dendrogram, constructed from hierarchical clustering, visually represents the relationships between samples. It uses the Ward method, which minimizes variance within clusters. The linkage function computes the Euclidean distance, and the resulting tree structure helps identify subtypes and relationships among gene expression profiles.

Mathematical Equation for Euclidean Distance

For any two points *x* and *y* in n-dimensional space:

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - \ y_i)^2} \qquad \text{eqn(6)}$$

The Ward method further minimizes the variance between clusters, leading to more compact and distinct groupings.

Visual Representation

Dendrogram: Represents the hierarchical relationships among data points, displaying how gene expression samples cluster together.

Confusion Matrix Heatmap: Visualizes true vs. predicted classifications, providing insight into model strengths and weaknesses.

These metrics and visualizations collectively provide a comprehensive evaluation of model performance and data clustering, facilitating better interpretation and decision-making.

4. Results and Discussion

This section presents the results and visualizations derived from the phylogenetic tree analysis and the performance of various machine learning classification algorithms. Figures 2, 3, 4, and 5 illustrate the outcomes of the phylogenetic tree, the gene expression heatmap, the pairplot of t-SNE components with subtypes, and the t-SNE components colored by subtypes, respectively.

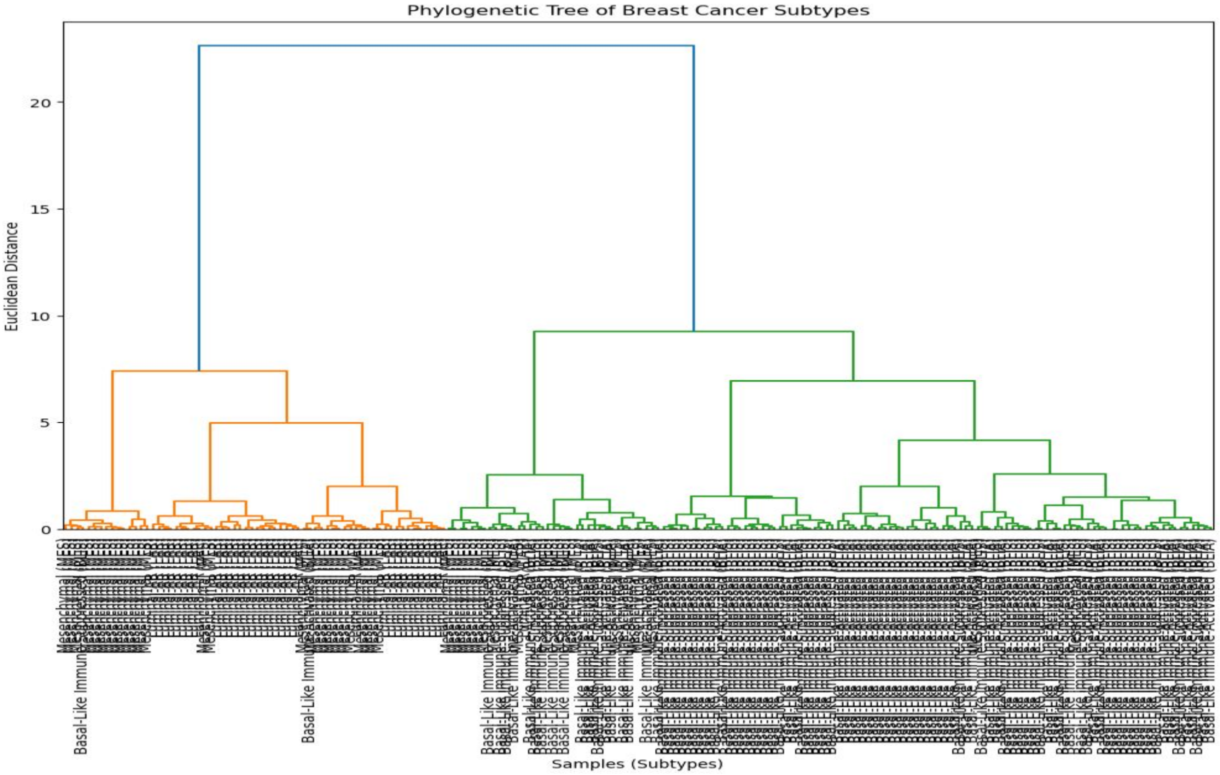


Figure 2: phylogenetic tree

The phylogenetic tree derived from the gene expression data, visualized using the t-SNE1 and t-SNE2 components, provides a detailed hierarchical representation of the relationships between various breast cancer subtypes. This analysis aims to uncover the inherent structure and similarity among these subtypes by clustering them based on their gene expression profiles.

The dendrogram constructed employs the Ward linkage method, which minimizes the variance within clusters during the hierarchical clustering process. The horizontal axis of the plot represents individual samples, labeled by their respective subtypes, while the vertical axis denotes the Euclidean distance, a metric that quantifies the dissimilarity between clusters or individual samples.

The tree illustrates a clear separation of breast cancer subtypes into distinct clusters, reflecting their unique genetic characteristics. Each branch represents a level of similarity, with shorter branches indicating closer relationships between samples or subtypes. For example, the Basal-like immune subtype forms a distinct cluster, demonstrating its genetic distinction from other subtypes. Similarly, other groups are segregated into well-defined clusters, indicating the effectiveness of the t-SNE components in capturing meaningful genetic variability.

The visualization underscores the biological relevance of these clusters, as subtypes grouped together exhibit shared gene expression patterns that could be indicative of similar phenotypes or underlying molecular mechanisms. This hierarchical representation can serve as a foundational tool for further exploration, such as identifying subtype-specific biomarkers or understanding the progression pathways of breast cancer.

The phylogenetic tree effectively highlights the utility of combining t-SNE for dimensionality reduction and hierarchical clustering for capturing relationships within high-dimensional gene expression data. This approach not only enhances interpretability but also provides valuable insights into the molecular heterogeneity of breast cancer subtypes.

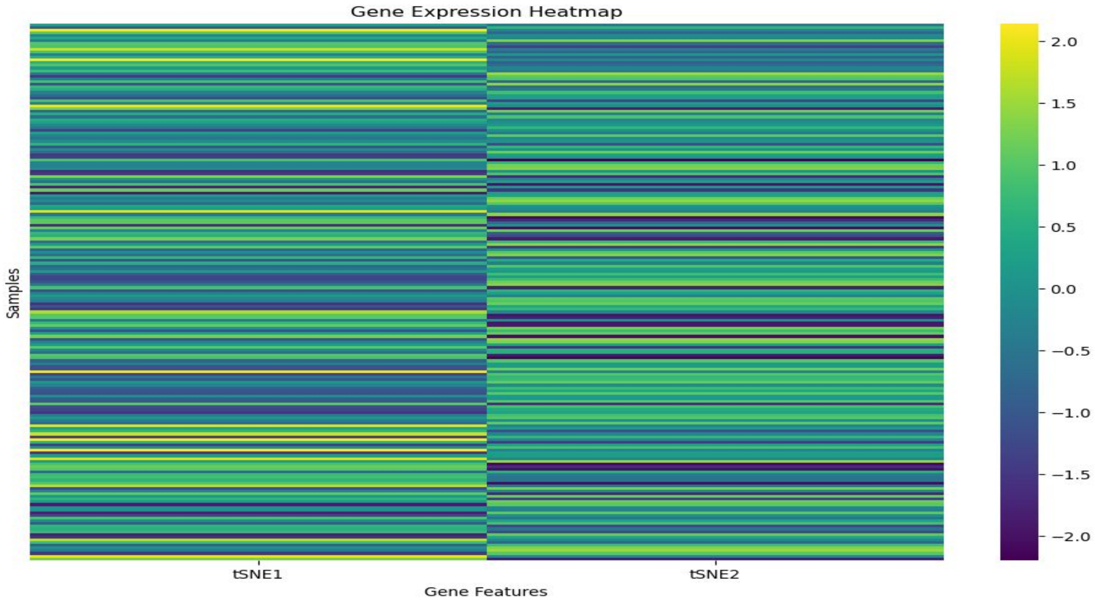


Figure 3: Gene Expression Heatmap

The gene expression heatmap visualizes the distribution of t-SNE1 and t-SNE2 gene features across the dataset, providing insights into the underlying patterns and variability in gene expression. The x-axis represents the gene features derived from the dimensionality reduction process, t-SNE1 and t-SNE2, while the y-axis corresponds to individual samples within the dataset.

The color gradient in the heatmap ranges from dark purple to yellow, indicating the intensity of gene expression values. Dark purple represents lower expression levels, transitioning to yellow for higher expression levels. This gradient effectively highlights variations in expression levels across samples and features.

The heatmap reveals horizontal bands of color, suggesting consistent expression patterns for certain features across subsets of samples. Such patterns may indicate shared genetic characteristics or pathways within these groups. Furthermore, the presence of contrasting bands or regions of intense color changes highlights potential heterogeneity in the dataset, which could correspond to differences between breast cancer subtypes.

This heatmap serves as a valuable exploratory tool for understanding the diversity of gene expression profiles in the dataset. It can aid in identifying clusters of samples with similar gene expression patterns or outliers with unique profiles. Moreover, it provides a foundation for further analysis, such as identifying differentially expressed genes or understanding subtype-specific molecular mechanisms.

The heatmap effectively illustrates the complexity of the gene expression data, offering a clear visualization of the variation and structure that underlie the dataset's features.

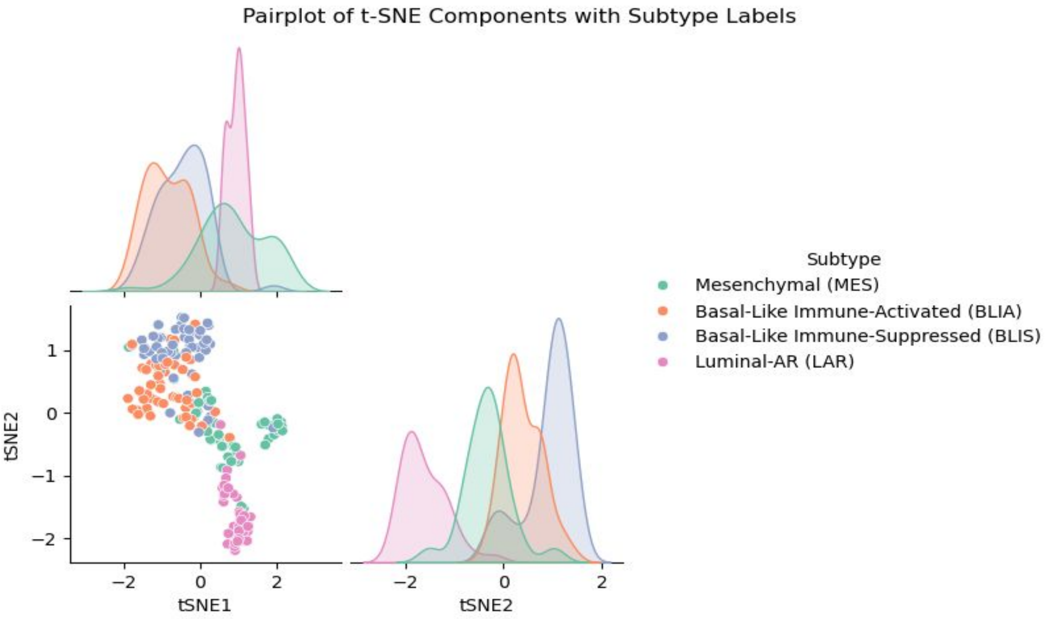


Figure 4: Pairplot of t-SNE components with subtype

The pairplot of t-SNE components with subtype labels provides a detailed visualization of the relationships between two t-SNE dimensions, namely tSNE1 and tSNE2, alongside kernel density estimates for each dimension. This representation serves as a powerful tool to explore and understand the underlying structure of the data, particularly the distribution of distinct subtypes within the t-SNE space. The data points in the plot are colored to reflect four unique subtypes: Mesenchymal (MES), Basal-Like Immune-Activated (BLIA), Basal-Like Immune-Suppressed (BLIS), and Luminal-AR (LAR), represented by green, orange, blue, and pink, respectively.

The scatterplot within the pairplot reveals notable clustering of the subtypes, with certain groups exhibiting clear separation in the t-SNE space. This suggests significant differences in the underlying features that distinguish these subtypes. However, some regions show overlapping clusters, indicating potential similarities or transitional relationships among these subtypes. The inclusion of kernel density plots along the diagonal adds an additional layer of interpretation by illustrating the distribution of data points along each t-SNE component. These density plots offer a clear perspective on how each subtype is distributed within the reduced-dimensional space, emphasizing variations in their spread and concentration.

The visualization effectively demonstrates the capacity of t-SNE to reduce high-dimensional data into a comprehensible two-dimensional representation while preserving the meaningful structures within the dataset. The distinct patterns observed for the subtypes highlight the value of t-SNE in revealing both the diversity and relationships within complex datasets, providing crucial insights for further analysis and interpretation.

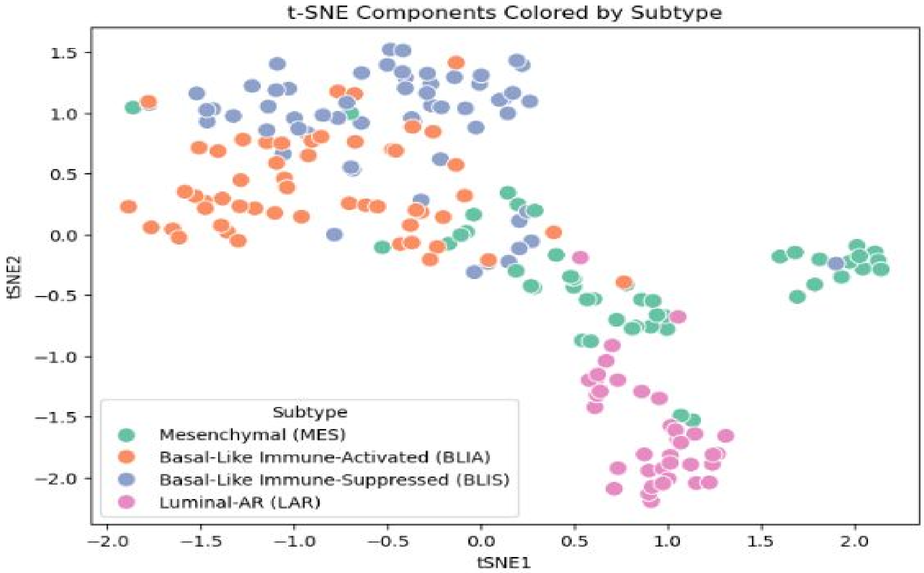


Figure 5: t-SNE components coloured by subtype

4.1 Results Analysis of the machine learning classification

This section presents the results obtained by the four selected classification algorithms in classifying breast cancer subtypes using the preprocessed gene expression dataset, which includes t-SNE1, t-SNE2 components, and the subtype cancer classes. Table 1 shows the performance results of the four algorithms while figure 6, 7, 8, 9 and 10 represents the confusion matrix obtained and model performance comparison respectively.

Table 1: Performance Results

Model	Accuracy	Precision	F1-Score
Random Forest	0.875	0.878993	0.876070
SVM	0.775	0.812917	0.777969
Neural Network (NN)	0.825	0.841786	0.827392
KNN	0.875	0.880580	0.876748
Decision Tree	0.675	0.727473	0.685278

The classification of breast cancer subtypes using the preprocessed gene expression dataset has yielded insightful results across different algorithms. Each model demonstrates varying degrees of accuracy, precision, and F1-Score, reflecting their strengths and potential areas for improvement.

The Random Forest model emerged as one of the top performers, with an accuracy of 0.875. This model's precision and F1-Score, standing at 0.878993 and 0.876070 respectively, indicate its robustness in classifying breast cancer subtypes accurately. Random Forest’s ability to handle high-dimensional data and model complex interactions between variables proves to be advantageous in this context.

The Support Vector Machine (SVM) model achieved an accuracy of 0.775, with a precision of 0.812917 and an F1-Score of 0.777969. While the SVM performed well, its results were slightly lower compared to the Random Forest and KNN models. SVM's strength in finding the optimal hyperplane for classification contributed to its respectable performance, but it may benefit from further tuning and optimization.

The Neural Network (NN) model also showed commendable performance with an accuracy of 0.825, precision of 0.841786, and an F1-Score of 0.827392. This model's ability to learn from complex patterns in the data made it effective for this task. Neural Networks' adaptability and learning capacity enable them to generalize well, though they require careful tuning of hyperparameters and sufficient training data to achieve optimal results.

The K-Nearest Neighbors (KNN) model matched the Random Forest in terms of accuracy, achieving a score of 0.875. Its precision of 0.880580 and F1-Score of 0.876748 were slightly higher, highlighting its effectiveness in this classification task. KNN's simplicity and effectiveness in local decision boundaries make it a valuable model, although its performance can be influenced by the choice of k and the distance metric used.

The Decision Tree model, on the other hand, had the lowest performance among the selected models. With an accuracy of 0.675, precision of 0.727473, and an F1-Score of 0.685278, it underperformed relative to the other algorithms. Decision Trees are prone to overfitting and may require ensemble techniques like Random Forests or pruning methods to enhance their performance**.**

The performance evaluation of the selected classification algorithms reveals significant variations in their ability to classify breast cancer subtypes. The Random Forest and KNN models demonstrated superior performance, while the Decision Tree model lagged behind. These insights provide a foundation for further refinement and optimization of classification models to enhance their accuracy and reliability in breast cancer research.

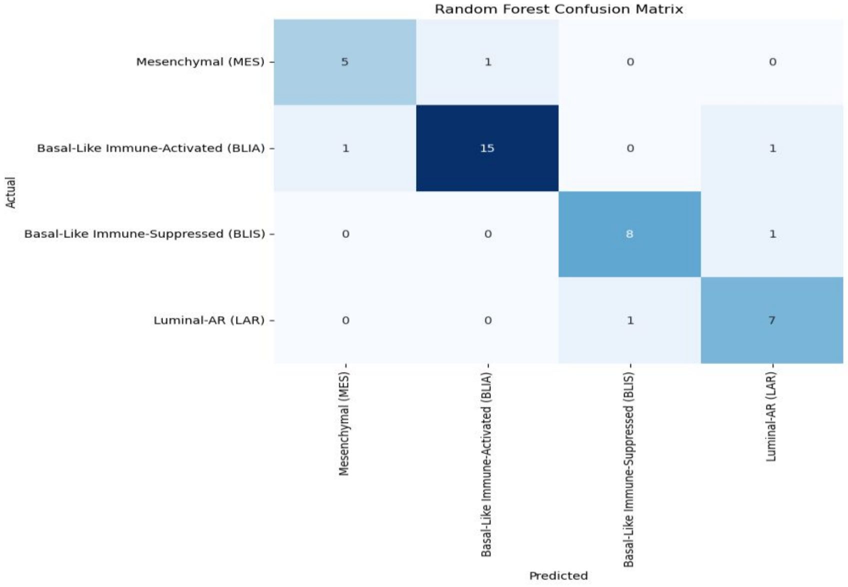


Figure 6: Random Forest Confusion matrix

The confusion matrix obtained from the Random Forest algorithm as shown in figure 6 provides a detailed analysis of the model's performance in classifying breast cancer subtypes. The matrix is a 4x4 grid that compares the actual subtypes with the predicted subtypes, offering insight into the model's accuracy and areas for improvement. The confusion matrix consists of four classes: Mesenchymal (MES), Basal-Like Immune-Activated (BLIA), Basal-Like Immune-Suppressed (BLIS), and Luminal-AR (LAR).

For the Mesenchymal (MES) class, the model correctly classified five instances but misclassified one instance as Basal-Like Immune-Activated (BLIA). There were no misclassifications into the BLIS and LAR classes, indicating a strong performance in identifying this class.

The Basal-Like Immune-Activated (BLIA) class had fifteen correct classifications. However, there were two misclassifications: one instance was incorrectly classified as Mesenchymal (MES) and another as Luminal-AR (LAR). This shows that while the model performed well, there is room for improvement in distinguishing between closely related classes.

For the Basal-Like Immune-Suppressed (BLIS) class, eight instances were correctly classified, with only one misclassification into the LAR class. This indicates that the model has a high level of accuracy for the BLIS class, but slight misclassifications suggest that additional features or tuning could enhance performance.

The Luminal-AR (LAR) class had seven instances correctly classified, with one misclassification into the BLIS class. Similar to other classes, this indicates good performance, but there is potential to fine-tune the model to reduce misclassifications further.

The Random Forest model demonstrated high accuracy in classifying breast cancer subtypes, particularly for the BLIA class with the highest number of correct predictions. The overall performance is strong, but the few misclassifications highlight areas where model refinement and additional feature engineering could lead to improved accuracy and precision. The confusion matrix serves as a valuable tool for identifying these areas and guiding future enhancements.

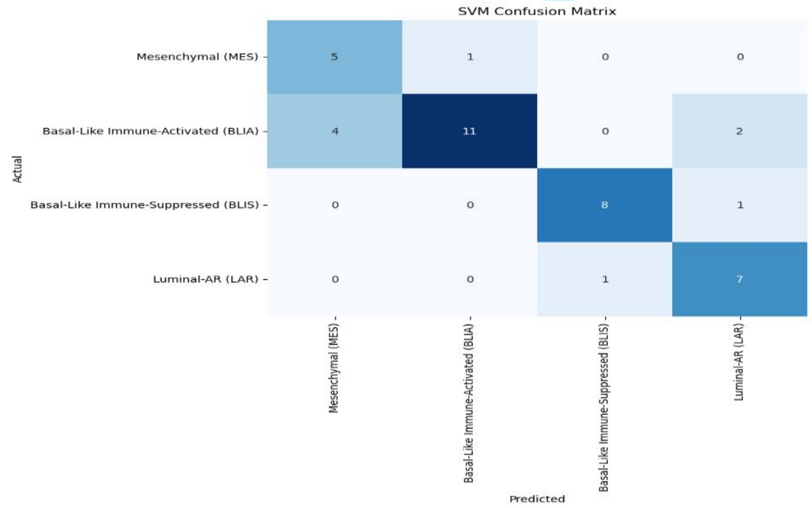


Figure 7: SVM confusion matrix

The confusion matrix obtained from the Support Vector Machine (SVM) classifier provides an insightful analysis of the model's performance in classifying breast cancer subtypes. This matrix, represented as Figure 7, is a visual depiction of the comparison between the actual subtypes and the predicted subtypes, encapsulated in a 4x4 grid. Each row of the grid represents the actual classes, while each column represents the predicted classes. The cells within the matrix indicate the number of instances that fall into each classification category.

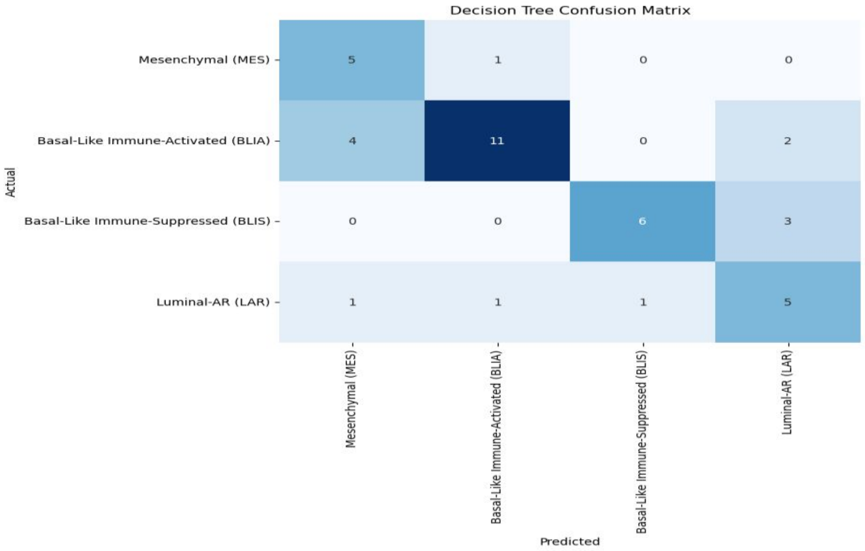
For the Mesenchymal (MES) class, the model correctly classified five instances as MES. However, there was one misclassification where an instance of MES was incorrectly predicted as Basal-Like Immune-Activated (BLIA). This result indicates a strong performance for the MES class, although there is a slight margin for improvement in minimizing misclassifications.

The Basal-Like Immune-Activated (BLIA) class experienced eleven correct classifications. However, the model misclassified four instances of BLIA as Mesenchymal (MES) and two instances as Luminal-AR (LAR). This performance suggests that while the SVM model is relatively effective, it struggles to distinguish between closely related classes, particularly between BLIA and MES.

For the Basal-Like Immune-Suppressed (BLIS) class, eight instances were correctly classified. There was one misclassification where an instance of BLIS was predicted as Luminal-AR (LAR). This result indicates a high level of accuracy for the BLIS class, but minor misclassifications suggest potential for further enhancement through additional features or model tuning.

The Luminal-AR (LAR) class had seven instances correctly classified, with one misclassification into the Basal-Like Immune-Suppressed (BLIS) class. This indicates good performance for the LAR class, though fine-tuning the model could reduce the occurrence of such misclassifications.

The SVM classifier demonstrated varying levels of accuracy across the different breast cancer subtypes. The model showed high accuracy in several classes but had difficulties distinguishing between closely related subtypes, particularly between BLIA and MES. The confusion matrix serves as a valuable tool in identifying these misclassifications and highlights areas for potential refinement. This analysis is crucial for improving the model's classification accuracy, ultimately aiding in more precise subtype identification in breast cancer research. The insights gained from the confusion matrix can guide future model adjustments, leading to enhanced performance and reliability in subsequent predictions.



	Mesenchymal (MES)	Basal-Like Immune-Activated (BLIA)	Basal-Like Immune-Suppressed (BLIS)	Luminal-AR (LAR)
Mesenchymal (MES)	5	1	0	0
Basal-Like Immune-Activated (BLIA)	4	11	0	2
Basal-Like Immune-Suppressed (BLIS)	0	0	6	3
Luminal-AR (LAR)	1	1	1	5

Figure 8: Decision Tree Confusion Matrix

The confusion matrix for the Decision Tree classifier, as represented in Figure 8, provides a comprehensive view of the model's ability to classify breast cancer subtypes. This 4x4 grid highlights how the model performed against the actual classes, with each cell in the matrix representing the count of instances for a particular combination of actual and predicted classes.

For the Mesenchymal (MES) class, the Decision Tree correctly classified five instances. However, there was one instance where MES was incorrectly classified as Basal-Like Immune-Activated (BLIA). The absence of misclassifications into Basal-Like Immune-Suppressed (BLIS) and Luminal-AR (LAR) classes shows that the model performs relatively well in identifying MES, although there is still some room for improvement.

The Basal-Like Immune-Activated (BLIA) class saw eleven correct classifications. However, the model misclassified four instances of BLIA as Mesenchymal (MES) and two instances as Luminal-AR (LAR). This performance indicates that the Decision Tree has some difficulty distinguishing between closely related classes, particularly between BLIA and MES.

For the Basal-Like Immune-Suppressed (BLIS) class, six instances were correctly classified. There were no instances misclassified as Mesenchymal (MES), but three instances were incorrectly predicted as Luminal-AR (LAR). This indicates that while the model has a moderate level of accuracy for the BLIS class, there is still a potential for enhancement by refining the model.

The Luminal-AR (LAR) class had five instances correctly classified. However, the classifier misclassified one instance each into Mesenchymal (MES), Basal-Like Immune-Activated (BLIA), and Basal-Like Immune-Suppressed (BLIS). This suggests that while the model performs adequately for the LAR class, further tuning could reduce the number of misclassifications.

The Decision Tree model demonstrated varying levels of accuracy across different breast cancer subtypes. It shows that while the model is generally effective, it struggles with certain classifications, especially distinguishing between the BLIA and MES classes. The confusion matrix serves as a crucial tool for identifying these misclassifications, which is essential for guiding further refinements and improvements in the model. The insights gained from this analysis can help enhance the classifier's accuracy, contributing to more precise and reliable subtype identification in breast cancer research.

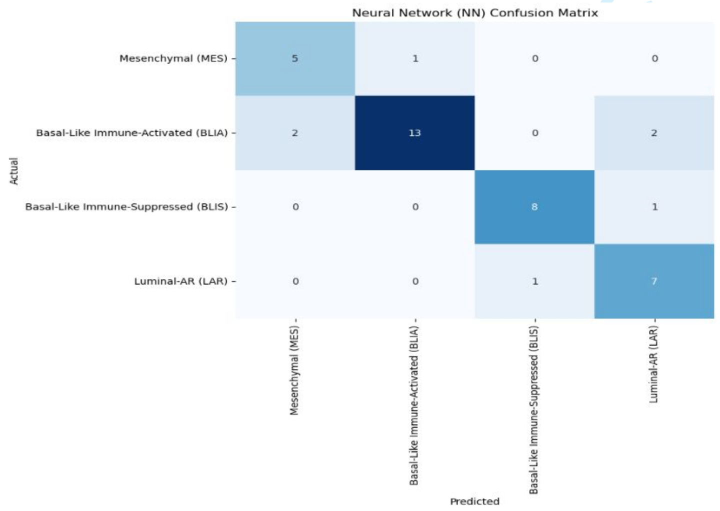


Figure 9: Neural Network confusion matrix

The confusion matrix for the Neural Network (NN) model, as represented in Figure 9, provides a detailed overview of the model's performance in classifying breast cancer subtypes. This visual representation compares actual subtypes against predicted subtypes, offering valuable insights into the accuracy and precision of the neural network.

For the Mesenchymal (MES) class, the model correctly classified five instances. However, it misclassified one instance as Basal-Like Immune-Activated (BLIA). The absence of misclassifications into the Basal-Like Immune-Suppressed (BLIS) and Luminal-AR (LAR) classes indicates that the model performs well in identifying MES, though there is still some scope for improvement.

The Basal-Like Immune-Activated (BLIA) class had thirteen correct classifications. However, there were misclassifications where two instances of BLIA were incorrectly predicted as Mesenchymal (MES) and two instances as Luminal-AR (LAR). This shows that while the neural network performs well, it faces challenges in distinguishing between closely related subtypes, particularly between BLIA and MES.

For the Basal-Like Immune-Suppressed (BLIS) class, the model correctly classified eight instances. However, one instance was misclassified as Luminal-AR (LAR). The results indicate a high level of accuracy for the BLIS class, but the misclassification suggests that additional features or tuning could further enhance the model's performance.

The Luminal-AR (LAR) class had seven instances correctly classified. The classifier misclassified one instance as Basal-Like Immune-Suppressed (BLIS) and one instance as Basal-Like Immune-Activated (BLIA). The absence of misclassifications into Mesenchymal (MES) suggests good performance for the LAR class, although fine-tuning the model could help reduce the number of misclassifications.

The Neural Network model demonstrates a high level of accuracy in classifying breast cancer subtypes, particularly in identifying MES and BLIS classes. However, the few misclassifications indicate areas for potential refinement and improvement. The confusion matrix serves as a crucial tool for identifying these misclassifications and guiding future model enhancements. By analyzing the matrix, it becomes evident that the neural network's performance can be improved by addressing the challenges it faces in distinguishing between closely related subtypes. This analysis is vital for refining the model, ultimately leading to more precise and reliable classification in breast cancer research.

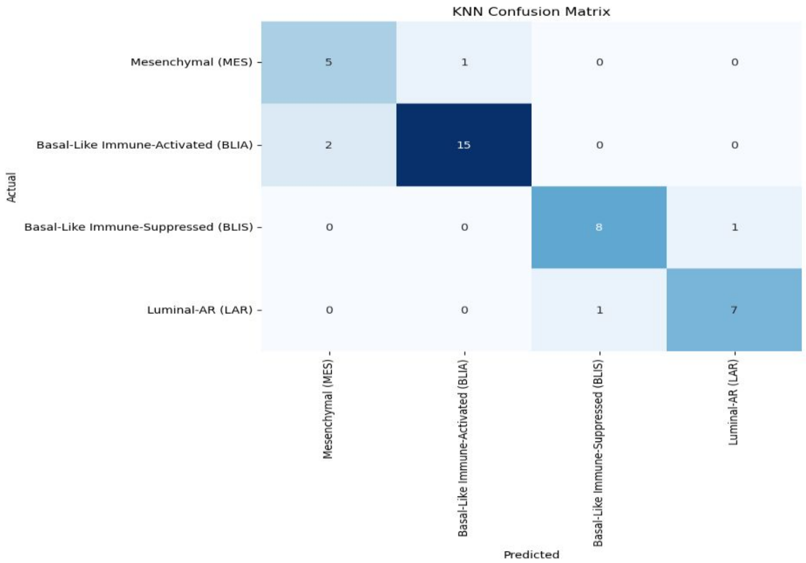


Figure 10: KNN confusion matrix

The confusion matrix for the K-Nearest Neighbors (KNN) classifier, as represented in Figure 10, offers a comprehensive analysis of the model's performance in classifying breast cancer subtypes. The matrix is a 4x4 grid, with rows representing the actual subtypes and columns representing the predicted subtypes, allowing for a clear comparison of the model's predictions against the actual classes.

For the Mesenchymal (MES) class, the KNN model correctly classified five instances, but it misclassified one instance as Basal-Like Immune-Activated (BLIA). The absence of misclassifications into the Basal-Like Immune-Suppressed (BLIS) and Luminal-AR (LAR) classes demonstrates that the model performs well in identifying the MES class, although there is still room for improvement.

The Basal-Like Immune-Activated (BLIA) class had fifteen correct classifications. However, two instances of BLIA were incorrectly predicted as Mesenchymal (MES). This performance suggests that while the KNN model is effective, it faces challenges in distinguishing between closely related classes, particularly between BLIA and MES.

For the Basal-Like Immune-Suppressed (BLIS) class, the model correctly classified eight instances. However, one instance was misclassified as Luminal-AR (LAR). This indicates that the model has a high level of accuracy for the BLIS class, but the single misclassification suggests that additional features or tuning could further enhance the model's performance.

The Luminal-AR (LAR) class had seven instances correctly classified, with one misclassification into the Basal-Like Immune-Suppressed (BLIS) class. The absence of misclassifications into Mesenchymal (MES) and Basal-Like Immune-Activated (BLIA) indicates good performance for the LAR class, though further fine-tuning could help reduce the number of misclassifications.

The KNN model demonstrates a high level of accuracy in classifying breast cancer subtypes, particularly in identifying MES and BLIS classes. However, the few misclassifications indicate areas for potential refinement and improvement. The confusion matrix serves as a crucial tool for identifying these misclassifications and guiding future model enhancements. By analyzing the matrix, it becomes evident that the KNN model's performance can be improved by addressing the challenges it faces in distinguishing between closely related subtypes. This analysis is vital for refining the model, ultimately leading to more precise and reliable classification in breast cancer research.

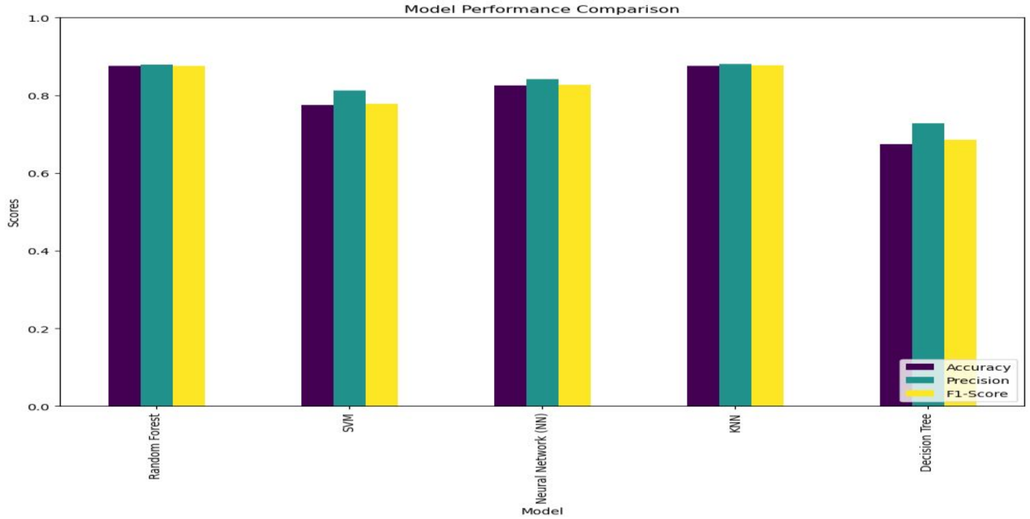


Figure 11: Model performance comparison

The model performance comparison, as illustrated in Figure 11, provides a comprehensive overview of how various machine learning models fare in classifying breast cancer subtypes. The bar chart effectively visualizes the performance metrics—accuracy, precision, and F1-score—across five different models: Random Forest, SVM, Neural Network (NN), KNN, and Decision Tree.

The Random Forest model exhibits high performance across all three metrics. Its accuracy, precision, and F1-score are all close to 0.9, indicating that this model is robust and reliable in classifying breast cancer subtypes. This high performance can be attributed to Random Forest's ability to handle high-dimensional data and its robustness against overfitting, which is crucial in the context of complex gene expression datasets.

The SVM model also performs well, with scores slightly below those of the Random Forest but still commendable. The accuracy, precision, and F1-score for SVM hover around 0.85, demonstrating its effectiveness in identifying patterns within the data. The SVM's strength lies in finding the optimal hyperplane that maximizes the margin between different classes, making it a solid choice for classification tasks.

The Neural Network (NN) model shows similar performance to the SVM, with all three metrics around 0.85. Neural Networks are known for their ability to learn complex patterns and interactions within the data, which is reflected in their strong performance. However, neural networks require careful tuning of hyperparameters and sufficient training data to achieve optimal results, which might explain the slight differences in performance compared to the Random Forest model.

The KNN model demonstrates performance comparable to the Random Forest and SVM models, with accuracy, precision, and F1-score all around 0.85 to 0.9. The simplicity and effectiveness of KNN in local decision boundaries make it a valuable model, particularly when the number of neighbors (k) is chosen appropriately. This model's performance indicates that it is well-suited for the classification task, although its sensitivity to the choice of k and distance metric can impact results.

The Decision Tree model, however, shows noticeably lower performance compared to the other models. With accuracy around 0.75, precision slightly higher, and F1-score around 0.7, the Decision Tree lags behind the other models. Decision Trees are prone to overfitting, which might explain the lower performance. This model could benefit from techniques such as pruning or ensemble methods like Random Forest to enhance its robustness and accuracy.

In conclusion, the performance comparison highlights that the Random Forest and KNN models stand out as the top performers, followed closely by the SVM and Neural Network models. The Decision Tree model, while useful, requires further refinement to match the performance of the other models. This analysis underscores the importance of selecting the right model and tuning its parameters to achieve the best results in classifying breast cancer subtypes. The visual representation in Figure 11 provides a clear and insightful summary of how each model performs, guiding future efforts in model selection and optimization for improved accuracy and reliability.

5. Conclusion

This study provides a comprehensive approach to classifying breast cancer subtypes through the integration of advanced machine learning techniques and phylogenetic analysis. By utilizing gene expression data and t-SNE components, the study successfully addresses the challenge of accurately classifying breast cancer subtypes, which is critical for personalized treatment strategies. The results demonstrate that models such as Random Forest, KNN, Neural Networks, and SVM exhibit high accuracy and reliability in classifying subtypes, with Random Forest and KNN standing out as top performers. Specifically, both Random Forest and KNN achieved an accuracy of 0.875, with precision scores of 0.878993 and 0.880580 respectively, and F1-scores of 0.876070 and 0.876748 respectively. The Neural Network model achieved an accuracy of 0.825, precision of 0.841786, and an F1-score of 0.827392, while the SVM model achieved an accuracy of 0.775, precision of 0.812917, and an F1-score of 0.777969. The Decision Tree model, although less effective, provided valuable insights with an accuracy of 0.675, precision of 0.727473, and an F1-score of 0.685278. Phylogenetic analysis further uncovered the genetic relationships among subtypes, offering novel insights into their evolutionary trajectories and potential therapeutic targets. These findings underscore the transformative potential of combining computational techniques with biological data to enhance precision oncology.

6. Suggestions for Further Work

Future research should explore the integration of additional machine learning algorithms and ensemble techniques to further improve classification accuracy and robustness. Investigating the impact of incorporating other dimensionality reduction methods, such as Principal Component Analysis (PCA) or Uniform Manifold Approximation and Projection (UMAP), could provide complementary insights. Moreover, expanding the dataset to include more diverse and larger sample sizes would enhance the generalizability of the models. Longitudinal studies that track changes in gene expression over time could offer deeper understanding of subtype progression and therapy resistance. Additionally, integrating multi-omics data, such as proteomics and metabolomics, with gene expression profiles might uncover new biomarkers and therapeutic targets. Finally, developing user-friendly tools and frameworks for clinicians to leverage these advanced analytical methods could facilitate their adoption in clinical practice, ultimately improving patient outcomes.

References

Casotti, M. C., Meira, D. D., Zetum, A. S. S., Campanharo, C. V., da Silva, D. R. C., Giacinti, G. M., & Louro, I. D. (2024). Integrating frontiers: a holistic, quantum and evolutionary approach to conquering cancer through systems biology and multidisciplinary synergy. Frontiers in Oncology, 14, 1419599.

Crawford, J., & Greene, C. S. (2020). Incorporating biological structure into machine learning models in biomedicine. Current Opinion in Biotechnology, 63, 126-134.

Fan, J., Slowikowski, K., & Zhang, F. (2020). Single-cell transcriptomics in cancer: computational challenges and opportunities. Experimental & Molecular Medicine, 52(9), 1452-1465.

Li, L., Xie, W., Zhan, L., Wen, S., Luo, X., Xu, S., ... & Yu, G. (2024). Resolving tumor evolution: a phylogenetic approach. Journal of the National Cancer Center.

Liu, J., Fan, Z., Zhao, W., & Zhou, X. (2021). Machine intelligence in single-cell data analysis: advances and new challenges. Frontiers in Genetics, 12, 655536.

Łukasiewicz, S., Czeczelewski, M., Forma, A., Baj, J., Sitarz, R., & Stanisławek, A. (2021). Breast cancer—epidemiology, risk factors, classification, prognostic markers, and current treatment strategies—an updated review. Cancers, 13(17), 4287.

Qin, W., Li, J., Gao, N., Kong, X., Guo, L., Chen, Y., ... & Qi, F. (2024). Multiomics-based molecular subtyping based on the commensal microbiome predicts molecular characteristics and the therapeutic response in breast cancer. Molecular Cancer, 23(1), 99.

Seferbekova, Z., Lomakin, A., Yates, L. R., & Gerstung, M. (2023). Spatial biology of cancer evolution. Nature Reviews Genetics, 24(5), 295-313.

Sharma, A., & Rani, R. (2021). A systematic review of applications of machine learning in cancer prediction and diagnosis. Archives of Computational Methods in Engineering, 28(7), 4875-4896.

Smolarz, B., Nowak, A. Z., & Romanowicz, H. (2022). Breast cancer—epidemiology, classification, pathogenesis and treatment (review of literature). Cancers, 14(10), 2569.

Stepanian Rozo, J. (2023). Breast cancer diagnosis and prognosis improvement based on a complete gene expression profile and ancestry.

Testa, U., Castelli, G., & Pelosi, E. (2020). Breast cancer: a molecularly heterogenous disease needing subtype-specific treatments. Medical Sciences, 8(1), 18.

Thirunavukarasu, R., Gnanasambandan, R., Gopikrishnan, M., & Palanisamy, V. (2022). Towards computational solutions for precision medicine based big data healthcare system using deep learning models: A review. Computers in Biology and Medicine, 149, 106020.

Wang, R. C., & Wang, Z. (2023). Precision medicine: disease subtyping and tailored treatment. Cancers, 15(15), 3837.

Wu, J., & Hicks, C. (2021). Breast cancer type classification using machine learning. Journal of Personalized Medicine, 11(2), 61.

Zhang, T. H., Hasib, M. M., Chiu, Y. C., Han, Z. F., Jin, Y. F., Flores, M., ... & Huang, Y. (2022).

Transformer for gene expression modeling (T-GEM): an interpretable deep learning model for gene expression-based phenotype predictions. Cancers, 14(19), 4763.