Deciphering Breast Cancer Prognostics: An Analysis of Gene-Expression Signatures and Their Correlation to Survival in Breast Cancer

Apishan Thayananthan
University of Southampton
COMP3212
Student ID: 33383642
Email: at16g21@soton.ac.uk

1. Introduction

Breast cancer is one of the significant causes of death among women [1], making it a significant health concern. Exploring the correlation between gene expression and patient survival is pivotal, whilst previous studies have analysed gene expression profiles in various cancers, such as lung adenocarcinoma and metastatic kidney cancer. They have found that gene expressions correlate with patient survival rates [2], [3].

This study delves into whether these findings also apply to breast cancer. By employing computational techniques to analyse biological data, we aim to identify specific gene expressions that can serve as reliable prognostic markers. The methodology combines machine learning models and statistical analysis to evaluate the predictive power of these gene expression profiles.

This study's ultimate goal is to enhance understanding of the molecular underpinnings of breast cancer prognostics and significantly contribute to advancing personalised medicine approaches. This could revolutionise the way we approach breast cancer treatment, tailoring it to the genetic profile of individual tumours and potentially improving patient outcomes [4].

This report outlines the scientific principles underpinning our computational approaches, details our methodology, presents our findings through well-labelled graphs and tables, and concludes with a comprehensive evaluation of the results and potential directions for future research.

2. Relevant Background

Studying gene expression patterns has been instrumental in understanding breast cancer's complexity. Early research identified distinct expression profiles that could differentiate between tumour subclasses, greatly influencing patient management strategies. These foundational studies highlighted the significant variation in clinical outcomes among patients with similar histopathological features, underscoring the potential of gene-expression profiles in prognostic applications [5], [6].

The Cox proportional hazards (CPH) model, a fundamental tool in cancer prognostics, continues to be a game-changer. It empowers researchers to effectively analyse survival data and comprehend the influence of various covariates on patient outcomes. This model is widely appreciated for its interpretability and robustness, particularly in handling complex datasets, which is critical for survival analysis in cancer research [7], [2], [3]. While the Cox model does not predict the most significant genes, it provides a solid framework for survival analysis, setting the stage for more detailed investigations.

Advanced machine learning techniques are pivotal in identifying specific genes that significantly influence outcomes. These methods, including hierarchical clustering and algorithms like support vector machines (SVM), are used to analyse and interpret high-dimensional gene-expression data. Such analyses are instrumental in identifying clusters or groups of genes that behave similarly, thereby revealing potential biomarkers [8], [9].

The Kaplan-Meier fitter is also utilised to estimate survival functions from lifetime data [10]. These computational tools collectively enhance our understanding of breast cancer at the molecular level and support the development of targeted and personalised treatment strategies.

Integrating gene expression profiling into clinical practice has significantly contributed to the emergence of personalised medicine. Studies have begun demonstrating how these profiles can guide treatment decisions, leading to more tailored and effective medical care for individual patients [11]. Furthermore, gene expression profiling tests are already used in clinical applications, indicating the practicality and feasibility of incorporating this technology into personalised medicine [12]. This shift towards personalised medicine promises better outcomes and emphasises computational biology's role in modern healthcare.

3. Material and Methods

3.1. Dataset Acquisition and Processing

The study utilised a dataset from UCSC Xena [13], including microarray expression profiling of 135 early-stage breast cancer tumours. This dataset was selected due to its comprehensive demographic representation of breast cancer cases and previous validations in external studies [14].

Our study utilised two primary datasets: clinical and gene expression data relevant to breast cancer research. These datasets were meticulously processed to ensure compatibility and accuracy in subsequent analyses.

Data Importation: The clinical dataset included patient demographics and clinical outcomes, while the gene expression dataset provided quantitative expression levels for numerous genes across various samples. These datasets were systematically imported into our analysis environment.

Data Encoding: To prepare the data for computational analysis, categorical variables such as estrogen receptor status, survival status, and other clinical outcomes were transformed into numerical codes. This encoding process facilitated clinical data integration with gene expression profiles, allowing for a comprehensive dataset combining phenotypic and genotypic information.

Data Integration and Cleaning: The clinical data was indexed by unique sample identifiers, which allowed for precise merging with the gene expression data. This aligns each patient's clinical information with their corresponding molecular data. The gene expression data was carefully transposed to match this structure. This step was critical to ensure that subsequent analyses would be based on accurately matched data.

Feature Selection: After merging, we performed feature selection to isolate gene expression data for further analysis. This process involved filtering the data to include only those features relevant to identifying potential prognostic biomarkers, thus focusing on the most pertinent data for our study objectives.

Through these thorough data processing steps, we ensured that our dataset was robust and tailored to address the specific research questions about the prognostic potential of gene expression in breast cancer. This initial phase laid the groundwork for applying advanced statistical models and machine learning algorithms to uncover insights into the molecular drivers of breast cancer prognosis.

3.2. Exploratory Data Analysis

3.2.1. Correlation Analysis

To grasp the interrelationships between gene expressions, we utilised correlation heat maps. This method quantifies the linear relationship between pairs of genes across all samples. These heat maps have been previously used in other studies to analyse gene expression patterns in breast tumours [15]. For practical reasons, we generated heat maps from a subset of gene expressions to manage computational load and enhance clarity. By visualising these correlations, we could identify genes that behave similarly, which might be linked to biological functions or pathways. High correlation values suggest functional connections, whereas low correlation may indicate independent actions.

Insights from Correlation Heat maps. The correlation heat maps in Figure 1 revealed significant insights into the interrelationships between genes. High correlation coefficients typically occur when a gene expression already exists, where if one gene expression occurs, another will most likely occur. However, in most scenarios, the gene expressions had little impact on each other, as shown in Figure 9 from Appendix A. This information is crucial for identifying gene clusters that could be targeted collectively in therapeutic strategies. Conversely, genes with low correlations may function independently, which could also be critical for understanding diverse mechanisms in breast cancer pathology.

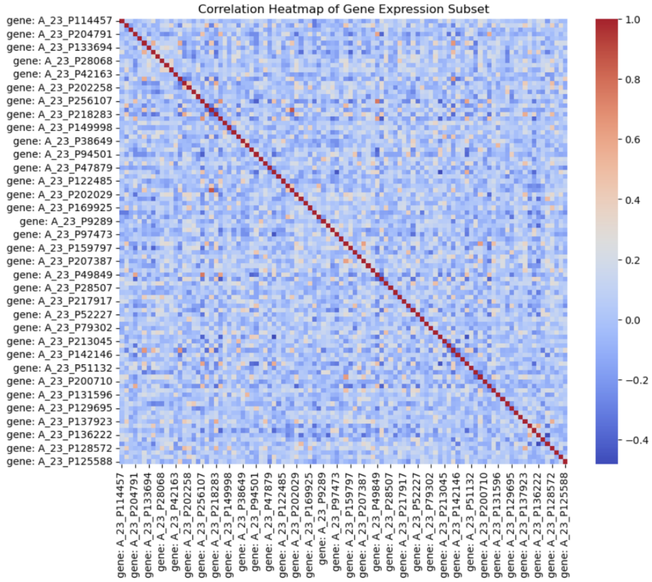

Figure 1: Correlation Heat map of Gene Expression Subset, of size 100

3.2.2. Histogram Visualisation

Histograms were also used to visualise the distribution of gene expression levels for selected genes. This approach helped examine the dataset's variability and central tendencies of gene expressions. By plotting histograms, we assessed whether the gene expression data followed a normal distribution or showed skewness, which could influence the application of specific statistical models. Each histogram plots the frequency of samples within various expression levels, providing insight into typical and atypical expression behaviours. An example of a histogram is shown in Figure 2.

Observations from Histogram Analysis The histograms of gene expression levels provided a detailed view of the distribution characteristics of gene expressions within the dataset. Several genes showed a normal distribution, as shown in Figure 2. In contrast, others displayed skewness to the left, as shown in Figure 3, or right, as shown in Figure 4, indicating over or under-expression relative to the median. These patterns are essential for selecting appropriate statistical tests and models that assume normality in later analyses. Moreover, outliers in some histograms suggest that specific genes exhibit extreme expressions in a subset of

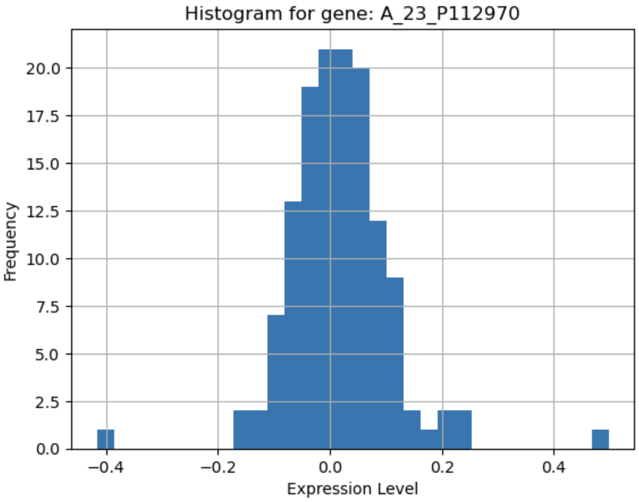

Figure 2: Histogram of gene *A_23_P112970* showing a symmetrical distribution of expression levels

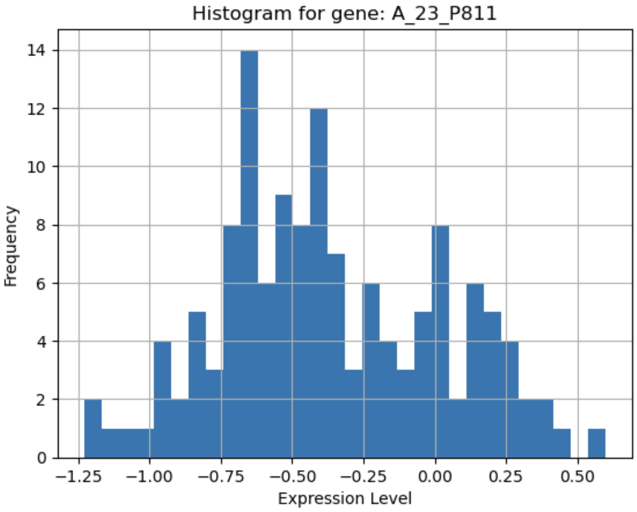

Figure 3: Histogram of gene *A_23_P138058* displaying a right-skewed distribution

patients, potentially signifying their crucial roles in specific cancer sub-types or responses to treatment.

3.3. Clustering Analysis

Clustering analysis is crucial for extracting meaningful insights from complex gene expression data. This method allows researchers to detect natural groupings and patterns within the data. For example, hierarchical clustering can reveal subgroups within cancer patients based on their gene expression patterns, potentially correlating to different prognoses or responses to treatment. The importance of this technique in the context of breast cancer research includes:

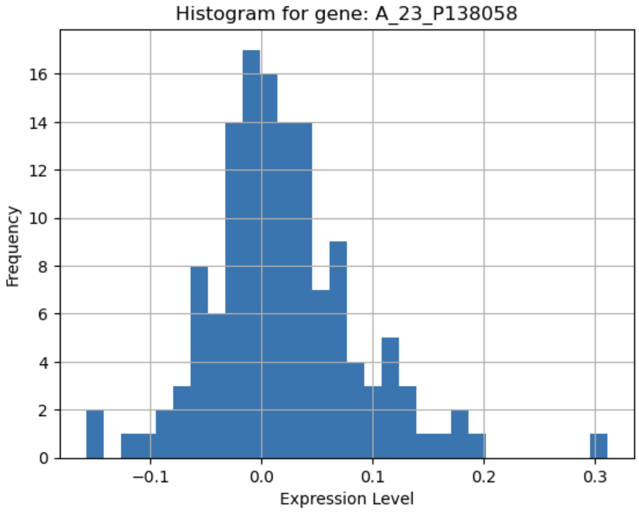

Figure 4: Histogram of gene *A_23_P811* exhibiting a left-skewed distribution

- **Identifying Molecular Subtypes:** Clustering, with its precision and specificity, can uncover distinct molecular subtypes of breast cancer. This differentiation is not just important; it is crucial for developing targeted therapies tailored to the specific biological pathways active in different tumour types.
- **Understanding Disease Mechanisms:** Genes with similar expression patterns are grouped by clustering, aiding in identifying biological pathways involved in disease progression.
- **Direct Impact on Patient Outcomes:** Hierarchical clustering of gene expression data refines diagnostic criteria significantly, enabling more precise categorization of disease states. This precision can directly impact patient outcomes, underlining the urgency and importance of its use.

Results from Hierarchical Clustering Analysis. The dendrogram produced by hierarchical clustering, as shown in Figure 5, visually represents the relationships among various genes based on their expression levels. The structure of the dendrogram helps in understanding the proximity and grouping of genes, where closely linked branches represent genes with similar expression patterns, suggesting functional similarities.

3.4. Statistical Modelling

In our statistical modelling, the CPH Model played a pivotal role in understanding the impact of individual gene expressions on patient survival times in breast cancer. Each gene was modelled separately to determine its influence on the survival outcomes, enabling us to estimate the effect sizes and show their statistical significance. This method helped identify which genes significantly affect survival, guiding further analyses.

Our focus on the top genes, identified through their p-values from the Cox model, has provided us with a deeper

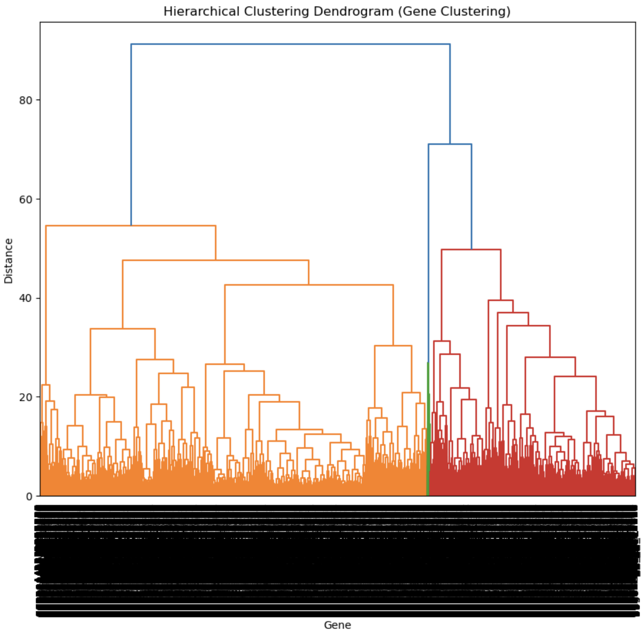

Figure 5: Hierarchical Clustering Dendrogram

Gene ID	P-value
A_23_P14621	3.40e-05
A_23_P22613	4.21e-05
A_23_P63896	7.59e-05
A_23_P69720	7.92e-05
A_23_P153301	8.65e-05
A_23_P2492	1.25e-04
A_23_P52569	1.40e-04
A_23_P98953	1.79e-04
A_23_P13554	1.89e-04
A_23_P99226	2.15e-04

TABLE 1: Top 10 Most Significant Genes by P-value

understanding. Table 1 presents the top 10 most significant genes per their corresponding p-values.

Each gene is listed with its unique identifier and the P-value obtained from the statistical analysis. The P-value indicates the likelihood of the observed correlation between gene expression and patient survival resulting from chance. A lower P-value indicates more vital evidence against the null hypothesis (P-value < 0.05), suggesting that the gene's expression level significantly affects survival outcomes. These results can provide valuable insights for developing more targeted therapies and understanding the biological mechanisms underlying breast cancer. Each gene represents a potential biomarker for prognosis, offering possibilities for personalised treatment plans based on genetic profiles.

Univariate Survival Analysis of Gene Expressions. Table 2 shows the top 10 genes being analysed separately to estimate their influence on survival outcomes, with their statistical significance indicated by p-values. A lower p-value suggests more robust evidence against the null hypothesis, indicating a significant effect of the gene's expression on survival.

Gene ID	KM Test Statistic	P-Value	-log2(p)
A_23_P13554	1.72	0.19	2.40
A_23_P14621	0.57	0.45	1.15
A_23_P153301	0.59	0.44	1.18
A_23_P22613	2.46	0.12	3.10
A_23_P2492	0.04	0.83	0.26
A_23_P52569	1.92	0.17	2.59
A_23_P63896	0.03	0.86	0.22
A_23_P69720	0.24	0.62	0.68
A_23_P98953	3.21	0.07	3.78
A_23_P99226	0.34	0.56	0.85

TABLE 2: Top 10 most significant genes identified using CPH Model, showcasing their KM test statistic, p-values, and the transformed logarithmic significance (-log2(p)).

- **KM Test Statistic:** This represents the value obtained from the Kaplan-Meier (KM) test, which assesses how well the gene predicts survival.
- **P-Value:** Indicates the probability that the observed difference in survival is due to chance. Values below 0.05 are typically considered statistically significant, although none of the genes here strictly meet this criterion, suggesting a cautious interpretation.
- **-log2(p):** This transformed measure helps visually identify the statistical significance, with higher values indicating greater significance.

In this study, while some genes show potential as prognostic markers due to lower p-values, such as *A_23_P98953*, *A_23_P22613*, and *A_23_P52569*, further research is required to validate these findings in more extensive cohorts. These genes could be targeted for therapeutic interventions, offering a pathway to personalised medicine.

Multivariate Cox Model Analysis. Following identifying potential prognostic genes through univariate analysis, the next step involved a deeper exploration via a Multivariate CPH Model. This approach evaluates the impact of gene expressions on survival outcomes while adjusting for other covariates, thus providing a comprehensive understanding of each gene's influence in the context of other known risk factors.

After identifying genes with potential significance in univariate models, multivariate analysis is crucial to ascertain their independent effect on survival outcomes. This analysis, shown in Table 3 from Appendix B, incorporates various clinical and demographic variables such as age, stage, grade, and estrogen receptor status, thus controlling for potential variables that might influence the survival predictions.

The p-values and confidence intervals associated with each coefficient assess the statistical robustness of the findings. In Appendix A, genes with significant p-values in this model, in *A_23_P22613* exhibit a notably high hazard ratio. This was similar to the findings from the Table 2, which indicate a strong and potentially clinically relevant impact on patient survival.

The analysis provides critical insights into which gene expressions significantly alter survival probabilities independently of other factors. For example, the high hazard ratio for *A_23_P22613* suggests it may play a crucial role in breast cancer.

Schoenfeld Residuals. The Schoenfeld residuals, a critical component of our analysis, were analysed to verify the proportional hazards assumption of the Cox model. These residuals are a powerful tool in determining if the effects of the covariates remain consistent over time. As seen in Figure 6, the residuals scatter around the zero line and show no clear trend over time, evident in both rank-transformed and km-transformed plots (p-values of 0.4601 and 0.4507, respectively). This lack of trend supports the model's validity, suggesting that the impact of gene *A_23_P14621* on survival is consistent throughout the study period. Such analyses are fundamental in confirming the reliability of our survival analysis results and reinforce the appropriateness of using the Cox model in our study.

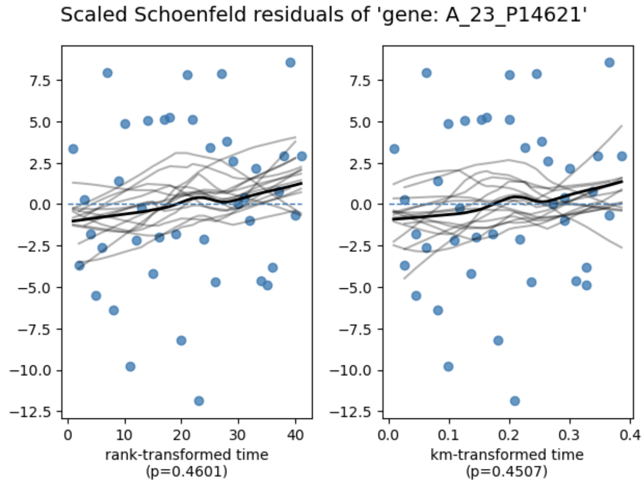

Figure 6: Schoenfeld Residuals of Gene *A_23_P14621*

Kaplan-Meier Survival Curves. The KM survival curves offer critical insights into the influence of gene expression on breast cancer survival. Analysis of gene *A_23_P22613*, shown in figure reveals significant survival differences, indicating that high expression correlates with lower survival rates, suggesting a negative impact on prognosis. Conversely, gene *A_23_P14621* shows a less noticeable effect on survival outcomes, implying a more moderate influence.

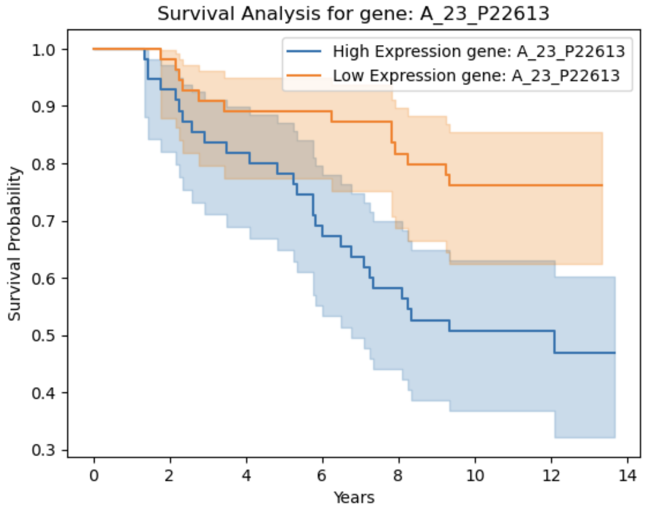

Figure 7: Kaplan-Meier plot for Gene *A_23_P22613*

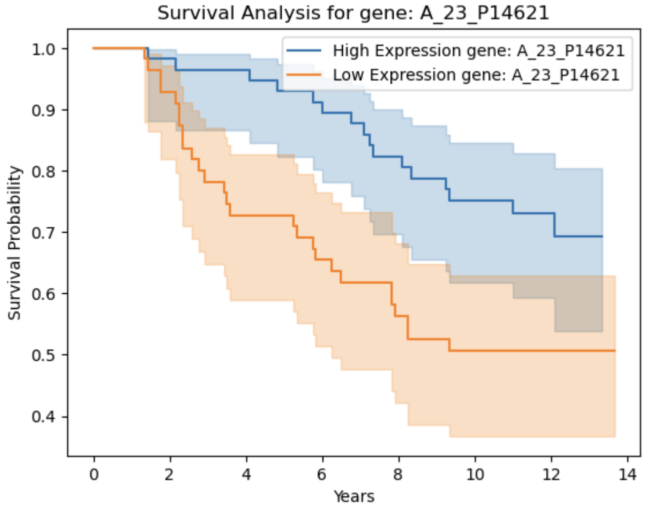

Figure 8: Kaplan-Meier plot for Gene *A_23_P14621*

3.5. Machine Learning Applications

Machine learning techniques were used to further analyse gene expression's impact on breast cancer survival rates. Machine learning provides robust tools for handling complex, high-dimensional data, allowing for more profound insight, and has predictive power beyond traditional statistical methods. This study used two models: SVM and Extreme Gradient Boosting (XGBoost).

SVM was selected for its effectiveness in achieving high classification accuracy in various contexts. Previous studies used SVMs to analyse and interpret high-dimensional gene-expression data, outperforming all other models [8]. The model excels in capturing intricate patterns without overfitting, thanks to its reliance on support vectors and the ability to optimise through kernel functions. These characteristics make SVM particularly suitable for datasets where the boundary between classes is not immediately apparent or is highly dimensional.

The novelty of this study stems from the incorporation of XGBoost, which was not used in similar studies [14], [8]. It was chosen due to its exceptional performance in dealing with structured data. This model is mainly known for its performance and speed in classification tasks. It is particularly adept at managing imbalanced datasets, a common issue in medical data analysis, through its advanced regularisation, which helps prevent overfitting. Several hyperparameters, including max depth, n_estimators, and learning rate, were optimised through grid search to enhance the model's ability to predict outcomes accurately. XGBoost has proven valuable in classification and identifying the most significant features contributing to the outcomes, thus providing insights into which genes are most influential in affecting patient survival rates.

SVM Results: The results from the SVM model show the best cross-validation accuracy of approximately 71.38%. While precision, recall, and F1-score values indicate moderate effectiveness, the balance between class predictions shows room for improvement. The top features influencing predictions display how specific genes negatively or positively impact the survival chances, thus directing focus towards potential biomarkers for further investigation. **XGBoost Results:** The XGBoost model demonstrated superior performance compared to SVM, achieving a best cross-validation accuracy of approximately 75.89%. The model effectively identified the top features that influence survival predictions, providing insights into potential biomarkers.

Both models played crucial roles in unveiling significant gene interactions and their implications on survival, paving the way for deeper biomedical investigations and inspiring new, personalised treatment approaches. They both provided top features influencing predictions that display how specific genes negatively or positively impact survival, thus directing focus towards potential biomarkers for further investigation. This is shown in Tables 4 and 5 from Appendix C. The Tables show that SVM and XGBoost outlined different genes as top predictors due to their distinct underlying algorithms, which interpret the data features and their relationships uniquely. This demonstrates that the machine learning approach may lead to variations in gene selection based on their modelling approaches.

4. Conclusion

This study has clarified the impact of gene expression on prognostic outcomes in breast cancer patients. Employing many analytical approaches,which is shown in Appendix D, the research integrated CPH models, SVM and XGBoost models, and hierarchical clustering analysis to examine the complex interrelations between gene expression patterns and patient survival probabilities. Deploying CPH models was instrumental in identifying specific genes that significantly affect survival outcomes, providing insights into their prognostic significance. Moreover, the application of SVM and XGBoost showcased the predictive accuracy of survival outcomes. Hierarchical clustering further contributed to our comprehension by outlining the inherent groupings within the gene expression data. Such classifications are pivotal to explaining disease progression's underlying biological mechanisms. Overall, this study reaffirms the critical role of gene expression profiling in breast cancer and highlights the benefits of integrating diverse computational techniques to refine the accuracy and personalisation of treatment modalities. The findings advocate for continued exploration into the molecular stratification of cancer to enhance therapeutic precision and patient care outcomes.

References

[1] K. P. Trayes and S. E. Cokenakes, "Breast cancer treatment," *American family physician*, vol. 104, no. 2, pp. 171–178, 2021.

[2] D. G. Beer, S. L. Kardia, C.-C. Huang, T. J. Giordano, A. M. Levin, D. E. Misek, L. Lin, G. Chen, T. G. Gharib, D. G. Thomas *et al.*, "Gene-expression profiles predict survival of patients with lung adenocarcinoma," *Nature medicine*, vol. 8, no. 8, pp. 816–824, 2002.

[3] J. R. Vasselli, J. H. Shih, S. R. Iyengar, J. Maranchie, J. Riss, R. Worrell, C. Torres-Cabala, R. Tabios, A. Mariotti, R. Stearman *et al.*, "Predicting survival in patients with metastatic kidney cancer by gene-expression profiling in the primary tumor," *Proceedings of the National Academy of Sciences*, vol. 100, no. 12, pp. 6958–6963, 2003.

[4] S. Jeibouei, M. E. Akbari, A. Kalbasi, A. R. Aref, M. Ajoudanian, A. Rezvani, and H. Zali, "Personalized medicine in breast cancer: pharmacogenomics approaches," *Pharmacogenomics and personalized medicine*, pp. 59–73, 2019.

[5] T. Sørlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. Van De Rijn, S. S. Jeffrey *et al.*, "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications," *Proceedings of the National Academy of Sciences*, vol. 98, no. 19, pp. 10 869–10 874, 2001.

[6] L. J. Van't Veer, H. Dai, M. J. Van De Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. Van Der Kooy, M. J. Marton, A. T. Witteveen *et al.*, "Gene expression profiling predicts clinical outcome of breast cancer," *nature*, vol. 415, no. 6871, pp. 530–536, 2002.

[7] J. Faradmal, A. Talebi, A. Rezaianzadeh, and H. Mahjub, "Survival analysis of breast cancer patients using cox and frailty models." 2012.

[8] X. Wei, J. Ai, Y. Deng, X. Guan, D. R. Johnson, C. Y. Ang, C. Zhang, and E. J. Perkins, "Identification of biomarkers that distinguish chemical contaminants based on gene expression profiles," *BMC genomics*, vol. 15, pp. 1–17, 2014.

[9] S. Cui, Q. Wu, J. West, and J. Bai, "Machine learning-based microarray analyses indicate low-expression genes might collectively influence pah disease," *PLOS Computational Biology*, vol. 15, no. 8, p. e1007264, 2019.

[10] B. Kearns, J. Stevens, S. Ren, and A. Brennan, "How uncertain is the survival extrapolation? a study of the impact of different parametric survival models on extrapolated uncertainty about hazard functions, lifetime mean survival and cost effectiveness," *Pharmacoeconomics*, vol. 38, pp. 193–204, 2020.

[11] L. J. Van't Veer and R. Bernards, "Enabling personalized cancer medicine through analysis of gene-expression patterns," *Nature*, vol. 452, no. 7187, pp. 564–570, 2008.

[12] A. Burska, K. Roget, M. Blits, L. Soto Gomez, F. Van De Loo, L. Hazelwood, C. Verweij, A. Rowe, G. Goulielmos, L. Van Baarsen *et al.*, "Gene expression analysis in ra: towards personalized medicine," *The pharmacogenomics journal*, vol. 14, no. 2, pp. 93–106, 2014.

[13] Caldas, "cohort: Breast cancer (caldas 2007)," Nov 2011. [Online]. Available: https://xenabrowser.net/datapages/?cohort=Breast+Cancer+%28Caldas+2007%29&removeHub=https%3A%2F%2Fxena.treehouse.gi.ucsc.edu%3A443

[14] A. Naderi, A. Teschendorff, N. Barbosa-Morais, S. Pinder, A. Green, D. Powe, J. Robertson, S. Aparicio, I. Ellis, J. Brenton *et al.*, "A gene-expression signature to predict survival in breast cancer across independent data sets," *Oncogene*, vol. 26, no. 10, pp. 1507–1516, 2007.

[15] C. M. Perou, T. Sørlie, M. B. Eisen, M. Van De Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen *et al.*, "Molecular portraits of human breast tumours," *nature*, vol. 406, no. 6797, pp. 747–752, 2000.