Unlocking the power of Genes: Prediction of Breast Cancer Survival assisted by Machine Learning

Ayesha Sahar
Dept. of Biosciences
Comsats University Islamabad
Islamabad, Pakistan
fa20-bsi-020@isbstudent.comsats.edu.pk

M. Umar Khan
Dept. of Electrical and Computer Engineering
Comsats University Islamabad
Islamabad, Pakistan
umar_khan@comsats.edu.pk

Maleeha Azam
Dept. of Biosciences
Comsats University Islamabad
Islamabad, Pakistan
maleeha.azam@comsats.edu.pk

***Abstract*—The breast cancer prognosis plays a vital role in treatment planning and patient management. This study aims to investigate breast cancer prognosis using the molecular taxonomy of the breast cancer international consortium's (METABRIC) dataset for gene expression profiles. The considered dataset contains clinical and genomic data, providing a comprehensive prognostic assessment resource. Our study focuses on developing a predictive model using machine learning techniques. We preprocess the dataset using adaptive synthetic sampling (ADASYN), MinMax scaling, and feature reduction techniques. We assess the model's performance using accuracy, confusion matrix, and classification report metrics. We compare the performance of Logistic Regression and Random Forest classification techniques to determine efficacy in breast cancer prognosis. The results demonstrate the potential of gene expression profiles in accurately predicting breast cancer prognosis. The results show that the Logistic Regression achieves a certain level of performance while the Random Forest exhibits improved predictive capabilities.**

***Index Terms*—breast cancer prognosis, gene expression profiles, METABRIC dataset, machine learning, Logistic Regression, Random Forest.**

I. Introduction

Cancer is a diverse group of diseases characterized by the abnormal growth and spread of malignant cells, resulting from genetic alterations. It is a major global health concern and ranks as the second leading cause of death worldwide, surpassed only by cardiovascular diseases [1]. It is estimated that breast cancer in particular, accounts for a significant proportion of new cancer cases and cancer-related deaths among women, making it a major public health concern [2]. The death rate from breast cancer is expected to increase by 10% in developed countries, while the death rate in developing countries is expected to increase by 15% [3]. Accurate prognosis plays a crucial role in determining appropriate treatment strategies and optimizing patient outcomes. Traditionally, breast cancer prognosis has relied on clinical factors such as tumor size, lymph node involvement, and histological grade. However, recent advancements in genomics and high-throughput technologies have opened up new opportunities to explore the potential of gene expression profiles for more precise and personalized prognostic assessments. Over the years, significant progress has been made in understanding the complexity of breast cancer, leading to the identification of various prognostic factors and biomarkers. In recent years, advancements in genomics and high-throughput technologies have revolutionized the field of breast cancer prognosis.

Gene expression profiling has emerged as a promising approach to gain insights into the molecular characteristics of breast cancer and improve prognostic predictions [4] [5]. By analyzing the expression levels of thousands of genes, it is possible to unravel the underlying biological mechanisms and identify signatures associated with disease progression, recurrence, and patient survival. Moreover, Machine learning (ML) approaches have enabled combinations of multiple types of clinical and biological data to make accurate risk predictions [6]. In this study, we aim to investigate the utility of gene expression profiles for breast cancer prognosis using the Breast Cancer Gene Expression Profiles (METABRIC) dataset. The METABRIC dataset offers a comprehensive resource for studying breast cancer, encompassing both clinical and genomic data. By integrating these two data domains, we can gain valuable insights into the complex molecular landscape of breast cancer and its association with clinical outcomes. Our research focuses on developing a predictive model using machine learning techniques to leverage the rich information contained within the genomic data to develop robust prognostic models. The findings highlight the importance of incorporating genomic information alongside clinical data for accurate prognosis assessment. The proposed approach has implications for personalized treatment planning and patient stratification, ultimately improving breast cancer management and patient outcomes [7].

We begin with a thorough literature review, identifying gaps in existing studies related to breast cancer prognosis. Building upon the identified gaps, we then present our system model, outlining the methodology and approach adopted in this study. Subsequently, the results and discussion are presented, offering an in-depth analysis of the findings and their implications. Finally, we conclude the paper by summarizing the key findings and discussing future directions in breast cancer prognosis research.

II. Literature Review

Cancer survival prediction using gene expression profiles has gained significant attention in recent years. The advent of high-throughput technologies has enabled researchers to analyze large-scale genomic data and develop models for predicting cancer prognosis. In this section, we review several research papers that have focused on cancer survival prediction.

The study by Ferroni et al. [8] aimed to develop a breast cancer prognosis model using a machine learning approach. They utilized a dataset consisting of 454 patients, with the model built on data from 318 patients. The authors employed a random forest algorithm and achieved an accuracy of 85% for the model. However, certain gaps can be identified in the research. The dataset used was relatively small, which may limit the generalizability of the findings to the broader population of breast cancer patients. Additionally, the study solely relied on a single machine learning algorithm, which may restrict the exploration of other potential algorithms that could improve the model's performance. Furthermore, the authors did not discuss the impact of overfitting on the model's performance, which is an important consideration in machine learning models to ensure generalizability. Lastly, the genomic data used in the study focused solely on gene expression data from RNASeq analysis, potentially overlooking other important genetic markers or molecular features that could enhance the predictive accuracy of the model.

The study conducted by Sammut et al. [9] focused on developing a multi-omic machine learning predictor of breast cancer therapy response. However, one notable gap in their research was the utilization of a relatively small dataset. This limitation has the potential to impact the accuracy of the predictions made by the model. With a smaller dataset, there may be insufficient representation of the diverse patterns and variations within breast cancer patients, which can lead to less accurate predictions. A larger dataset would provide more comprehensive and diverse information for the model to learn from, enabling it to capture a wider range of patterns and enhance its predictive performance.

The study by Santos et al. [10] utilized the same dataset as us, the Breast Cancer Gene Expression profiles (METABRIC) dataset, to develop their model. However, the study had certain gaps that could have further enhanced its predictive performance. Firstly, the authors did not address the issue of data imbalance in the dataset. They did not employ techniques for oversampling or undersampling to mitigate the impact of imbalanced classes, which could lead to biased predictions. Furthermore, the study did not explore the linearity of the data, which could have provided insights into the relationship between different variables and potentially improved the model's interpretability. In addition, the authors did not perform feature selection specifically for the genomic part of the dataset, which may have resulted in the inclusion of irrelevant or redundant features, potentially leading to suboptimal model performance. These two aspects, data imbalance and feature selection, are crucial in building accurate and robust machine learning models for breast cancer prognosis. Lastly, the study employed XGBoost as the only classification algorithm, achieving an accuracy of only 0.779. This suggests the potential for further improvements in the model's predictive performance by exploring other advanced algorithms or optimizing the hyperparameters.

In the study conducted by Boeri et al. [11], there was an absence of key data types, which provide crucial insights into the underlying biological mechanisms and patient-specific characteristics of breast cancer. By omitting these data sources, the analysis may overlook important predictive factors and hinder the development of a robust prognosis model. Moreover, the limited number of patients included in the study was another significant gap. A small sample size as already mentioned before, reduces the representation of the findings and may lead to less reliable predictions.

In the study conducted by Liu et al. [12], 631 cases of BC were downloaded from TCGA breast cancer database (TCGA-Breast Invasive Carcinoma), which included 87 cases in the healthy control group and 544 cases in the cancer group. Overall, this data's size is small and they did not address the imbalance in this dataset.

In another study by Kalafi et al. [13], they used a considerably big dataset with data of 4902 patients, but they just focused in clinical attributes only and did not include the gene expression profiles of these patients. Repo et al. [14] developed a prognostic model based on cell-cycle control that predicts the outcome of breast cancer patients but the dataset only had data for just 1135 patients.

In response to these gaps, our study addresses these limitations by incorporating both genomic and clinical data in the analysis. We used the dataset, "Breast Cancer Gene Expression Profiles (METABRIC)". It had clinical features, m-RNA levels z-score, and genes mutations for 1904 patients. The dataset can be divided into two parts clinical and genomic. The clinical part consisted of 31 features whereas the genomic part consisted of m-RNA levels z-score for 331 genes, and mutation for 175 genes. By integrating genomic data, such as gene expression profiles obtained from RNASeq analysis along with gene mutations, our study aims to capture the molecular intricacies of breast cancer, enabling a more comprehensive understanding of the disease. Furthermore, the inclusion of clinical variables provides essential context and patient-specific information that enhances the accuracy and relevance of the prognosis model. Additionally, our study recognizes the importance of a larger patient cohort and aims to expand the dataset. By including a greater number of patients, the statistical power of the analysis increases, allowing for more robust and reliable predictions. The incorporation of a more extensive and diverse patient population improves the generalizability of the prognosis model and enhances its applicability in real-world scenarios. Our study has also included the "most relevant" features from the genomic part of our dataset via employing feature selection techniques. The comprehensive evaluation of multiple data types and a larger sample size enable a better understanding of the disease and improve the accuracy and reliability of the prognosis model.

III. System Model

In this section, we discuss the system model of our study to develop an accurate and robust breast cancer prognosis methodology, as presented in Figure.1. This system model encompasses several key components: data pre-processing, feature engineering, model training, and evaluation. The goal of the system model is to leverage high-throughput genomic data and clinical information to predict cancer survival outcomes for individual patients.

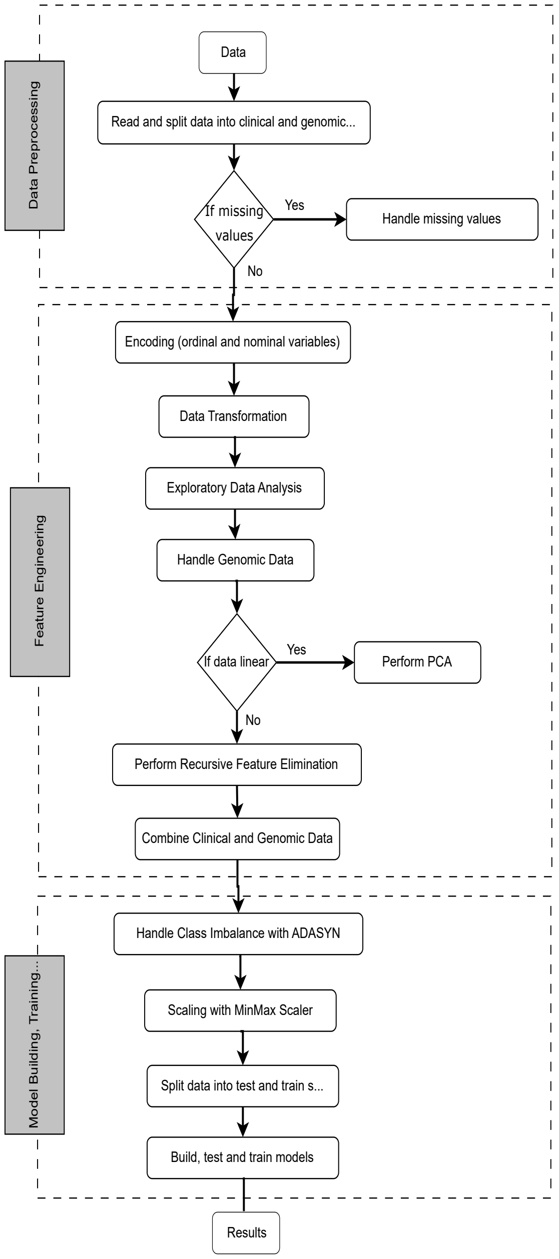

Data Preprocessing
Data
Read and split data into clinical and genomic...
If missing values
Yes
Handle missing values
No
Feature Engineering
Encoding (ordinal and nominal variables)
Data Transformation
Exploratory Data Analysis
Handle Genomic Data
If data linear
Yes
Perform PCA
No
Perform Recursive Feature Elimination
Combine Clinical and Genomic Data
Model Building, Training...
Handle Class Imbalance with ADASYN
Scaling with MinMax Scaler
Split data into test and train s...
Build, test and train models
Results

Fig. 1. Flow of the proposed methodology for beast cancer prognosis

A. *Data Preprocessing*

The first step in the system model is data preprocessing. The breast cancer gene expression profiles dataset (METABRIC) dataset, was utilized for this research. The dataset was loaded into a pandas DataFrame, and the clinical and genomic data were separated into two distinct subsets. Missing values in the clinical data were addressed by imputing them with the mode of each respective column, ensuring the preservation of valuable information.

B. *Feature Engineering*

Following data preprocessing, feature engineering techniques were employed to enhance the predictive power of the model. The clinical data underwent label encoding to convert categorical variables into numerical representations. This process facilitated the incorporation of clinical information into the model. Additionally, nominal features were one-hot encoded to account for their categorical nature, allowing for the utilization of these features in the prediction process effectively. The genomic data underwent a transformation to ensure its compatibility with the model. The gene expression values were binarized, representing the presence or absence of gene expression for each respective gene. Correlation analysis was conducted to identify highly correlated features and to see if the data was linear or not. We found that the genomic data was not linear, so in order to reduce the dimensionality of the data, we used the Recursive Feature Elimination Technique.

C. *Model Building, Training and Evaluation*

The processed clinical and genomic data were merged into a unified dataset for model training. To address the challenge of imbalanced classes in the dataset, the ADASYN oversampling technique was applied to generate synthetic samples for the minority class, thereby balancing the class distribution. The unified dataset was then split into training and testing sets using an 80:20 ratio, respectively. The model training phase involved the utilization of two machine learning algorithms: Logistic Regression and Random Forest Classifier. Logistic Regression, a linear classifier, and Random Forest Classifier, an ensemble method, were chosen due to their proven effectiveness in binary classification tasks. During the training phase, the model learned the patterns and relationships between the features and the target variable (i.e., death from cancer). The trained model was subsequently evaluated using various metrics, including a confusion matrix, classification report, and accuracy scores. These metrics provided insights into the model's performance, including its ability to correctly predict cancer survival outcomes and its generalization to unseen data.

IV. Results and Discussion

This section discusses the results and the corresponding analysis of the considered data. The performance metrics in regard to the machine learning algorithms are also evaluated. We divide the clinical and genomic data into distinct subsets. The missing values are catered through pre-processing

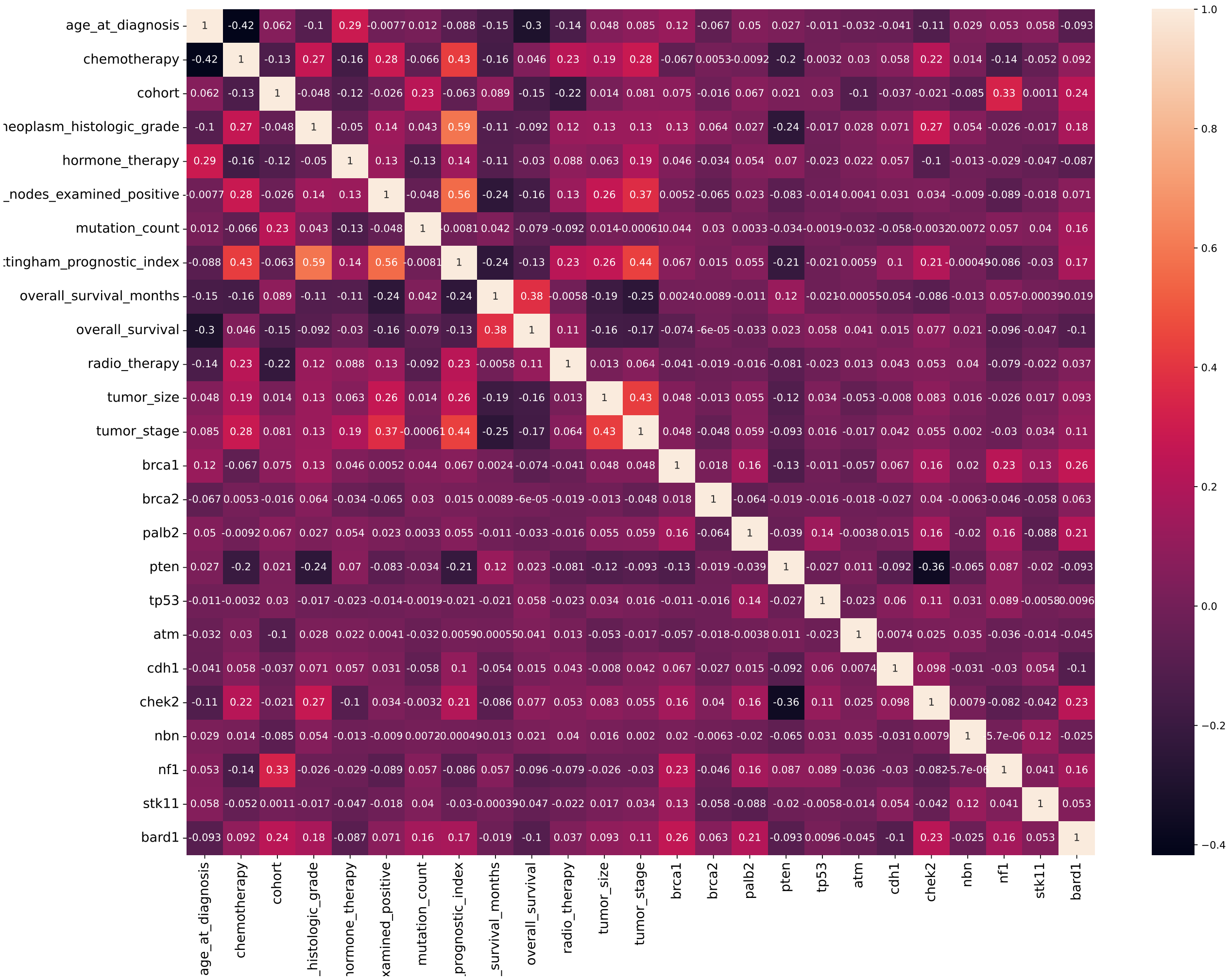


Fig. 2. Correlation Matrix Heatmap for Clinical Data

followed by encoding techniques. The correlation analysis to assess the relationships and dependencies between the clinical features can be seen in Figure.2.

The genomic part of our dataset contained information on m-RNA levels z-score for 331 genes and mutation status for 175 genes. This comprehensive genomic data encompassed a wide range of genetic variations and expression profiles, providing valuable insights into the molecular characteristics of breast cancer. However, due to the high dimensionality of the genomic data, it posed a challenge for our analysis. Applying traditional dimensionality reduction techniques such as Principal Component Analysis (PCA) was not feasible since the genomic data exhibited non-linear relationships. To address this issue, we employed the Recursive Feature Elimination (RFE) technique. RFE is a feature selection method that recursively eliminates less informative features while training a model, ultimately identifying a subset of the most relevant features for prediction. Via RFE, we reduced our genomic data features to the most relevant 100 features. Then for the target variable, we studied the samples per class. Since there was an imbalance in the dataset, we used ADASYN to handle this issue. In the table given below, the original counts and the counts after ADASYN can be seen.

TABLE I
CLASS COUNTS

Class	Original Count	Count after ADASYN
Died of Disease (0)	622	706
Died of other causes (1)	480	808
Living (2)	802	802

Then we normalized the dataset and split it into train (80%)

and test set (20%)

Finally, the prognosis model for breast cancer was evaluated using two machine learning algorithms: Logistic Regression and Random Forest Classification. The results obtained from both algorithms are presented below.

A. *Logistic Regression*

The Logistic Regression model achieved the following results:

- The model correctly classified 85 patients with the prognosis that they will die from cancer, and they actually did. However, for 50 patients, it incorrectly predicted that they will die from cancer, but they did not.
- The model correctly classified 120 patients with the prognosis that they will die from some other disease, not from cancer, and they did die from some other disease. However, for 37 patients, it incorrectly classified them as such.
- The model correctly classified 172 patients with the prognosis that they will not die from cancer, and they actually did not die from cancer.

The confusion matrix and the classification report can be seen in the Figures. 3 and 4.

The train score of the Logistic Regression model is 0.847, indicating that the model was able to correctly predict the class of 84.7% of the samples in the training set. The test score is 0.8125, indicating that the model was able to correctly predict the class of 81.25% of the samples in the test set.

B. *Random Forest Classification*

The Random Forest Classification model achieved the following results:

- The model correctly classified 96 patients with the prognosis that they will die from cancer, and they actually did. However, for 39 patients, it incorrectly predicted that they will die from cancer, but they did not.
- The model correctly classified 133 patients with the prognosis that they will die from some other disease, not from cancer, and they did die from some other disease. However, for 24 patients, it incorrectly classified them as such.
- The model correctly classified 172 patients with the prognosis that they will not die from cancer, and they actually did not die from cancer.

The confusion matrix and the classification report can be seen in Figures 5 and 6.

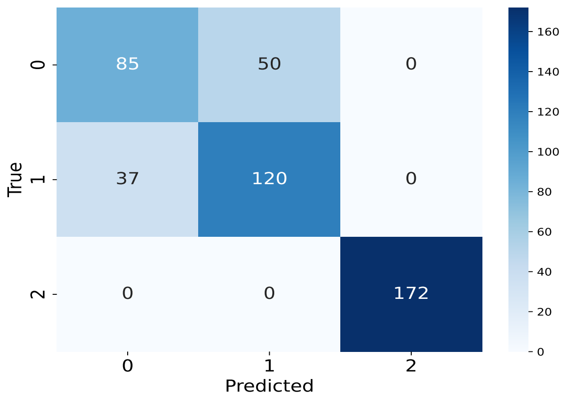
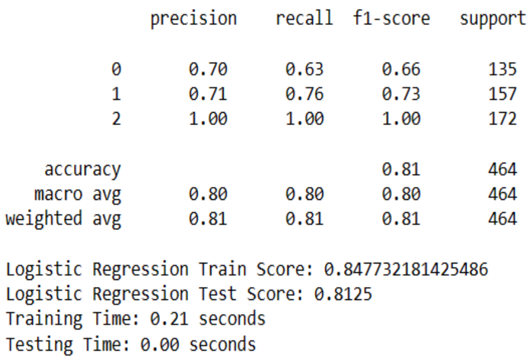


Fig. 3. Confusion Matrix of Logistic Regression



```
              precision    recall  f1-score   support

           0       0.70      0.63      0.66       135
           1       0.71      0.76      0.73       157
           2       1.00      1.00      1.00       172

    accuracy                           0.81       464
   macro avg       0.80      0.80      0.80       464
weighted avg       0.81      0.81      0.81       464

Logistic Regression Train Score: 0.847732181425486
Logistic Regression Test Score: 0.8125
Training Time: 0.21 seconds
Testing Time: 0.00 seconds
```

Fig. 4. Classification Report of Logistic Regression

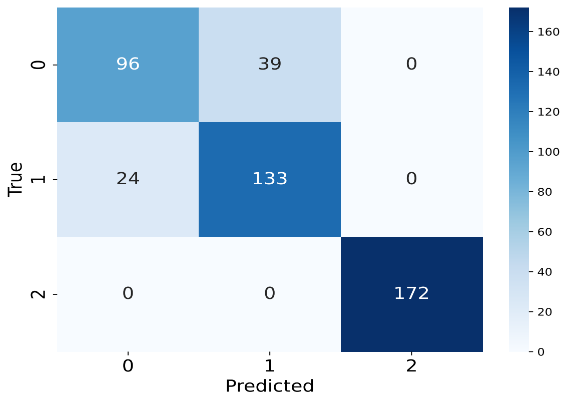
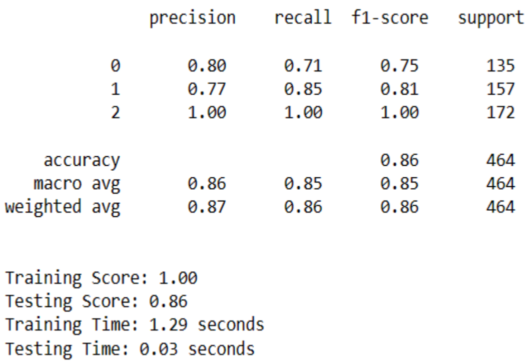


Fig. 5. Confusion Matrix of Random Forrest Classifier



```
              precision    recall  f1-score   support

           0       0.80      0.71      0.75       135
           1       0.77      0.85      0.81       157
           2       1.00      1.00      1.00       172

    accuracy                           0.86       464
   macro avg       0.86      0.85      0.85       464
weighted avg       0.87      0.86      0.86       464


Training Score: 1.00
Testing Score: 0.86
Training Time: 1.29 seconds
Testing Time: 0.03 seconds
```

Fig. 6. Classification Report of Random Forrest Classifier

The train score of the Random Forest Classification model is 1.00, indicating that the model was able to correctly predict the class of 100% of the samples in the training set. The test score is 0.86, indicating that the model was able to correctly predict the class of 86% of the samples in the test set.

C. Discussion

The results obtained from both algorithms provide insights into the performance of the breast cancer prognosis model. The Logistic Regression model demonstrated moderate predictive power, with an overall accuracy of 81.25% on the test set. It showed relatively better performance in predicting patients who will not die from cancer, achieving an accuracy of 84.7% on the training set. However, the model struggled to accurately predict patients who will die from cancer, with a significant number of false-positive predictions.

On the other hand, the Random Forest Classification model exhibited higher performance, achieving an accuracy of 86% on the test set and perfect accuracy on the training set. This was also greater than the accuracy of Santos et al. 's [10] best model, XGBoost, which had an accuracy of 77.9%. This indicates that the Random Forest model was able to learn the complex relationships between features and the target variable, resulting in improved predictive accuracy. It demonstrated better performance in predicting patients who will die from cancer, achieving a higher true-positive rate compared to the Logistic Regression model. The comparison of both algorithms can be seen in Table 2 below.

TABLE II
COMPARISON OF ALGORITHMS

Algorithm	Accuracy	Training Time (s)	Testing Time (s)
Logistic Regression	81.25%	0.21	0.01
Random Forest	86%	1.29	0.03

These results suggest that the Random Forest Classification model outperforms the Logistic Regression model and the XGBoost model as well in terms of predictive accuracy for breast cancer prognosis. The high accuracy achieved by the Random Forest model on the training set indicates its potential to generalize well to unseen data.

V. CONCLUSION

In conclusion, our study investigated the use of gene expression profiles for breast cancer prognosis using the molecular taxonomy of the breast cancer international consortium's (METABRIC) dataset. We developed a predictive model using machine learning techniques that integrated clinical and genomic data. The model can predict patient survival with an accuracy of 86% accurately.

The findings of our study demonstrate the value of gene expression profiling as a powerful tool in breast cancer prognosis. However, it is essential to acknowledge that the findings are based on the analysis of a specific dataset and may not be directly generalizable to all breast cancer populations. Additionally, while the predictive model achieved exceptional accuracy, there is room for further improvement. Incorporating additional clinical and genomic features and exploring more advanced machine learning algorithms could enhance the prognostic accuracy and refine patient stratification.

In future research, we aim to validate our findings using independent datasets and explore integrating other genomics data, such as proteomics and epigenomics, to improve prognostic models.

REFERENCES

[1] Miller, K.D.; Ortiz, A.P.; Pinheiro, P.S.; Bandi, P.; Minihan, A.; Fuchs, H.E.; Martinez Tyson, D.; Tortolero Luna, G.; Fedewa, S.A.; Jemal, A.M.; et al. Cancer Statistics for the US Hispanic/Latino Population, 2021. CA A Cancer J Clin 2021, 71, 466–487, doi:10.3322/caac.2169

[2] World Health Organization (WHO). (2020). Breast cancer. Retrieved from https://www.who.int/news-room/fact-sheets/detail/breast-cancer

[3] Xu, Y., Gong, M., Wang, Y. et al. Global trends and forecasts of breast cancer incidence and deaths. Sci Data 10, 334 (2023). https://doi.org/10.1038/s41597-023-02253-5

[4] Munkácsy, G.; Santarpia, L.; Győrffy, B. Gene Expression Profiling in Early Breast Cancer—Patient Stratification Based on Molecular and Tumor Microenvironment Features. Biomedicines 2022, 10, 248, doi:10.3390/biomedicines10020248.

[5] Brewczyński, A.; Jabłońska, B.; Mazurek, A.M.; Mrochem-Kwarciak, J.; Mrowiec, S.; Śnietura, M.; Kentnowski, M.; Kołosza, Z.; Składowski, K.; Rutkowski, T. Comparison of Selected Immune and Hematological Parameters and Their Impact on Survival in Patients with HPV-Related and HPV-Unrelated Oropharyngeal Cancer. Cancers 2021, 13, 3256, doi:10.3390/cancers13133256

[6] Zitnik, M. et al. Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. Information Fusion 50, 71–91 (2019)

[7] Ahmed, Z.; Mohamed, K.; Zeeshan, S.; Dong, X. Artificial Intelligence with Multi-Functional Machine Learning Platform Development for Better Healthcare and Precision Medicine. Database 2020, 2020, baaa010, doi:10.1093/database/baaa010.

[8] Ferroni, P., Zanzotto, F. M., Riondino, S., Scarpato, N., Guadagni, F., & Roselli, M. (2019). Breast cancer prognosis using a machine learning approach. Cancers, 11(3), 328. https://doi.org/10.3390/cancers11030328

[9] Sammut, SJ., Crispin-Ortuzar, M., Chin, SF. et al. Multi-omic machine learning predictor of breast cancer therapy response. Nature 601, 623–629 (2022). https://doi.org/10.1038/s41586-021-04278-5

[10] Santos, D. (2022). Breast Cancer Survival Prediction using Machine Learning and Gene Expression Profiles. medRxiv (Cold Spring Harbor Laboratory). https://doi.org/10.1101/2022.01.22.22269470

[11] Boeri C, Chiappa C, Galli F, De Berardinis V, Bardelli L, Carcano G, Rovera F. Machine Learning techniques in breast cancer prognosis prediction: A primary evaluation. Cancer Med. 2020 May;9(9):3234-3243. doi: 10.1002/cam4.2811. Epub 2020 Mar 10. PMID: 32154669; PMCID: PMC7196042.

[12] Liu L, Chen Z, Shi W, Liu H, Pang W. Breast cancer survival prediction using seven prognostic biomarker genes. Oncol Lett. 2019 Sep;18(3):2907-2916. doi: 10.3892/ol.2019.10635. Epub 2019 Jul 18. PMID: 31452771; PMCID: PMC6676410.

[13] Kalafi EY, Nor NAM, Taib NA, Ganggayah MD, Town C, Dhillon SK. Machine Learning and Deep Learning Approaches in Breast Cancer Survival Prediction Using Clinical Data. Folia Biol (Praha). 2019;65(5-6):212-220. PMID: 32362304.

[14] Repo, H., Löyttyniemi, E., Kurki, S. et al. A prognostic model based on cell-cycle control predicts outcome of breast cancer patients. BMC Cancer 20, 558 (2020). https://doi.org/10.1186/s12885-020-07045-3