RESEARCH PAPER

Predicting Early- and Late-Stage Breast Cancer via Clinical–Genomic using Feature Fusion and Explainable Ensemble Learning

Sonali Mondal Das[1*], Abhoy Chand Mondal[2]

[1*]Department of Computer Science, University of Burdwan, West Bengal, India. Email ID: sonalimondal099@gmail.com
[2]Department of Computer Science, University of Burdwan, West Bengal, India. Email ID: abhoycmondal@gmail.com

Corresponding Author
Sonali Mondal Das
Email ID: sonalimondal099@gmail.com

ABSTRACT
Correctly identifying what stage breast cancer is at is extremely valuable to doctors since it will improve their ability to make the appropriate diagnosis or treatment. To achieve this, we created a machine-learning system that utilized a combination of clinical information from the METABRIC data set along with a vast variety of Gene Expression features to classify the stage of breast cancers into an early (I & II) or late (III & IV) classification system. We were very diligent in our efforts to ensure that we were not overlooking an area of importance; we encoded, normalized, and achieved balance through the use of Tomek-SMOTE Balancing and stratified splitting. We tested various models - XGBoost, Random Forest, SVM, and Voting and Stacking Classifiers - and measured their performance by ROC-AUC, PR-AUC, and F1-Score to provide the most accurate representation of performance. For the early-stage breast cancer samples, there were marked differences in performance, and XGBoost achieved an area under the ROC curve of 0.9967 and an area under the precision/recall curve of 0.9981 which is indicative of a highly successful model. In late-stage breast cancer samples, SVM performed the best; however, its area under the ROC curve only reached 0.5772 which signifies that performance was not quite as successful. We utilized SHAP to gain insight into clinical and genomic markers, and it became evident that only a small number of markers have a major influence on breast cancer classification. Overall, this study indicates that for improved results on the classification of breast cancer, it is important to approach & develop models for each of the four stages separately while utilizing a combination of Clinical and Genomic data.

Keyword: Breast cancer staging, METABRIC dataset, combining clinical and genomic features, gene expression analysis, machine learning classification, ensemble learning, SMOTE-Tomek balancing and explainable AI.

How to cite this article: Das SM, Mondal AC. Predicting Early- and Late-Stage Breast Cancer via Clinical–Genomic using Feature Fusion and Explainable Ensemble Learning. Int J Drug Deliv Technol. 2026;16(42s): 763-770. DOI: 10.25258/ijddt.16.42s.85

INTRODUCTION

For women, Breast cancer was still among the leading causes of death. Correctly staging breast cancer, as part of determining how to treat it, is essential; it determines not only the prognosis but also the type of treatment option selected for an individual patient. The majority of doctors use standard testing methods to determine a cancer's stage; however, even the best standard testing methods do not provide a complete picture. Standard tests do not always provide a clear connection between clinical characteristics and molecular characteristics. Artificial intelligence (AI) and machine learning (ML) improve cancer care by delivering new, higher-level insights into cancer diagnosis and treatment. Machine learning and AI models have access to both genomic and clinical data at a more profound level and can identify patterns in data, as opposed to the more basic methods previously employed.

Unfortunately, this progress on the part of AI and ML has not yet provided a solution to the problem of incomplete datasets and unevenly distributed datasets. Breast cancer patients may have similar clinical features but possess entirely different molecular characteristics between early- and late-stage tumors. This point demonstrates that one standard "one-size-fits-all" model is insufficient for diagnosing, staging, and treating all breast cancer patients.

In the current study, we created an advanced machine-learning architecture that contained all of the clinical data from the METABRIC database as well as precise gene expression profiles. In order to process the data effectively, extensive data preparation has included cleaning and normalizing. The resulting data has been evaluated using several forms of advanced machine learning techniques: e.g., ensemble and kernel-based techniques, including XGBoost, Random Forest, SVM, Voting and Stacking classifiers. The accuracy of the model and how well it performed was evaluated by using the receiver operating characteristic (ROC) area under the curve (AUC), precision-recall AUC, and F1-score metrics.

2. Literature review

Breast cancer has been classified into different stages in many different ways through various lines of research including: - deep learning using imaging data, multi-omics

*Author for Correspondence: sonalimondal099@gmail.com

and gene expression profiling, METABRIC-based classification studies, and stage-specific modelling frameworks. In reviewing these approaches, this article will provide additional information for each of these areas relevant to the current study.

Recently, Islam et al. [1] used ensemble classifiers with SHAP-based explainability on breast cancer data, with XGboost achieving 97% accuracy and SHAP finding the most important predictors. This use of ensemble classifiers based on SHAP-based explainability was mirrored in this study. Among the three gene expression assays clinically validated — Oncotype DX, MammaPrint, and PAM50 — the evidence collected on the utility of genomic markers as meaningful inputs to the classification systems for breast cancer risk stratification [2] has seen gene expression data formally incorporated into the AJCC 8th Edition staging system as an integral part of the staging process [3], thereby providing a direct impetus for including gene expression variables as features within the classification model developed in this study. Curtis et al. [4] established the METABRIC dataset used in this study and identified ten new integrative clusters of breast cancer from the 2,000 analysed primary breast tumours based on integrated copy number analysis and gene expression analysis, showing that these clusters have distinct genomic drivers and clinical outcomes. Pereira et al. [5] added another level of complexity to these findings by identifying 40 novel driver genes of mutation in 2,433 METABRIC breast tumours. At the higher multi-omics integrative level, the molecular TCGA pan-cancer study [6] shown that combining genomics, transcriptomics and proteomics data can significantly enhance breast cancer subtyping, thus establishing the scientific rationale for this study to combine clinical characteristics and gene expression characteristics. The work of Kate and Nadig [7] was the foundation for this project and provided direct comparison on methods with 174,518 patients from SEER, using 3 classifiers. They concluded that stage-derived models consistently performed better than models where all stages were combined, and that when performance was evaluated collectively, it tended to be higher than when evaluated based on stage because of different outcome rates for different stages. Importantly, stage 4 had the lowest performance (AUC = 0.71-0.72). This supports the binary classification of early (Stages I-II) and late (Stages III-IV) stages and puts into context our findings of the difficulty of classifying late stage compared to early stage (SVM AUC = 0.5772). Finally, Said et al. [8] also confirmed that stage-derived models perform better than full dataset for advanced stage disease across 5 classifiers. Research on epidemiology has identified several risk factors associated with cancer. For example [9], how well patients are screened prior to diagnosis, and various socioeconomic demographics [10] and geographical differences [12]; as well as whether a patient is an immigrant [11]. Therefore, accurate classification of the stage of cancer is clinically significant. Despite this research literature, very few attempts have been made to use machine learning to classify patients as either having early-stage or late-stage cancers using both clinical variables and gene expression data from the METABRIC dataset. This study uses several machine learning algorithms including XGBoost, Random Forests, Support Vector Machines (SVM), and ensemble classifiers to classify patients based on their stage at diagnosis. Yoon et al. [13] also used machine learning models to classify patients as either having early-stage or late-stage breast cancer using SEER-Medicare data; they found that RFS predicts continued survival in patients diagnosed with early-stage breast cancer based on characteristics such as age at diagnosis, those who had received chemotherapy and thyroid disease.

3. Methodology

The research paper uses a full-fledged machine learning framework to predict breast cancer progression stages based on the analysis of integrated clinical and genomic data. The methodology was developed in a systematic way to overcome major issues in biomedical data science such as high-dimensionality feature space, imbalance between classes, and biological heterogeneity among disease phases. Figure 1 illustrates the entire workflow of analytical process that includes data integration, preprocessing, stratified cohort development, ensemble modeling and interpretable machine learning.

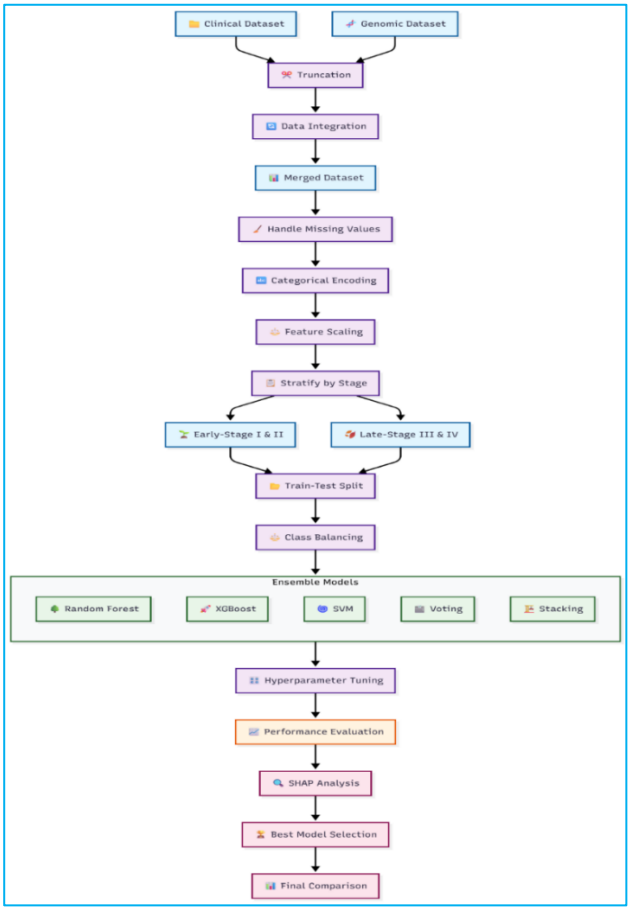

Fig. 1. Workflow Methodology

3.1. Data Acquisition and Preprocessing

3.1.1. Dataset Description and Integration

Two sets of data were utilized for the METABRIC collection for the purposes of this study. High-throughput genomic profiling systems supplied gene expression datasets (n=24,368 features) while demographic, clinical and treatment data (n=1,980 samples) for each of the participants formed the second data set. To correlate the clinical data set with the genomic data set, gene_data.iloc[:clinical.shape[0],:] was used to reduce any invalid genomic data sets obtained from high-throughput genomic profiling companies to their corresponding sizes of the clinically based dataset. The evaluation of the data collected so far indicated that both data sets have anomalous dimensions and consequently require careful use. Column by column data set concatenation was accomplished with the use of the pd.concat() function. The resultant data structure contains clinical characteristics and genomic characteristics of all patients for each row of the data set. The sample and feature index alignment is preserved for further studies when the data is re-indexed using the index reset.

3.1.2. Data Quality Assurance and Cleaning

To remedy issues with data quality, there was significant data cleaning through extensive replacement with NaN for infinite values before the implementation of systematic imputation by replacement of numerical feature values with the mean imputation method (for numerical features) and categorical feature values with the mode imputation method (for categorical features). The approach upheld the integrity of the data and mitigated bias related to data that had missing values, as well as less than 5 percent of the imputed values were found to be between the features.

3.1.3. Clinical Stratification Strategy

With an understanding that breast cancer tumor staging is a critical predictor of survival, we divided the participant population in accordance with their clinical presentation into clinically-relevant groups. The groups established were:

- Patients presenting with early-stage disease (TNM staging I and II [1 and 2]) and not stage 0 due to small numbers of participants in this group.
- Patients presenting with late-stage disease (TNM staging III and IV [3 and 4]). Since the two stages have distinctive molecular structures and clinical support systems, the creation of discrete groups by stage made it possible to create stage-specific models.

3.2. Feature Engineering and Preprocessing Pipeline

3.2.1. Categorical Feature Encoding

To convert the categorical variables from a string format to an integer format that will maintain the relationship between the categorical data points (and make them valid for use in machine learning algorithms), we have used the LabelEncoder class to encode them. We will also be able to convert all of the categorical variables into float using the LabelEncoder class. When we do conversions of categorical features into float-point numbers, we can guarantee that every data in the pipeline has the same representation, therefore, complexity is also reduced.

3.2.2. Feature Normalization

Since the scales of the clinical measurements and expression of all the genes vary heterogeneously, feature standardization, via StandardScaler, was applied. This transformation made features centered (zero-mean, unit-variance) to avoid dominance of large magnitude features and achieve the best performance of the algorithm.

3.2.3. Numerical Stability Assurance

In order to eliminate potential problems with numerical instability caused by NaN and infinite numbers, all feature matrices used in this research were preprocessed to replace those values with a zero, using the function np.nan_to_num(). This preprocessing step was critical to help provide strong model performance on both training and evaluation tasks.

3.3. Class Imbalance Mitigation

3.3.1. Resampling Strategy Selection

The preliminary analysis showed that there was substantial class imbalance across different stages of the cancer. Through teaming SMOTETomek, both SMOTE and Tomek Link under sampling techniques are combined to provide a unique solution to an imbalanced dataset. Using SMOTETomek, synthetic examples were generated for the under-represented classes, while instances that were located at uncertain boundaries between classes were removed. Using this approach created a balanced distribution of classes whereby the number of samples generated for the minority class was 0.6 times the number of the majority class.

3.3.2. Stratified Data Partitioning

Stratified sampling was used to ensure that the same proportions of class members would be used in both the training and test datasets In order to accomplish this task, the training data contained 65% of the data while the testing data contained 35% of the data. The function train_test_split() had been used to ensure that both training and test datasets had the same or similar proportion of the classes. This means that the model was trained and evaluated with representative samples throughout the process of developing and evaluating the model.

3.4. Ensemble Modeling Framework

3.4.1. Multi-Algorithm Ensemble Design

Ensemble strategies were designed with a multitude of algorithms to discover patterns that may be used in combination with one another.

- **Bagging -** The classifier bagging model was based on a random forest classifier as a base estimator. Using bootstrapping, we were able to create multiple instances of the model and combine them to produce a single more accurate prediction.
- **Boosting Methods**- XGBoost with a logarithmic loss evaluation metric is one of the boosting techniques**.**
- **Support Vector Machine**- SVM models were developed with radial basis function (RBF) kernels and the probability calibration parameter turned on.

- **Voting** - A voting classifier model took the predictions of the Random Forest, XGBoost, SVM, and Logistic Regression classifiers into account when making a final prediction.
- **Stacking** - A two-level architecture was used to create a stacking classifier with base estimators of XGBoost and SVM and a meta-learner model of Logistic Regression.

3.4.2. Stage-Specific Binary Classification

The multi-class staging problem was reduced to a two-class model by using binary classification tasks:-

- **Early-stage patients** (Stage 1 and Stage 2) coded as 0 and 1
- **late-stage patients** (Stage 3 and Stage 4) coded as 0 and 1 were used to create binary classification tasks that directed the model's attention toward clinically relevant stage transitions and provided adequate sample sizes for the training of the model in order to improve its ability to discriminate among the various stages**.**

3.5. Model Evaluation and Interpretation Framework

3.5.1. Comprehensive Performance Metrics

We used a multidimensional assessment approach to evaluate the performance of the models:

- ❖ **Primary Metrics**: ROC-AUC (area under the receiver operating characteristic curve) and PR-AUC (precision-recall area under the curve);
- ❖ **Secondary Metrics**: accuracy, precision, recall, and f1-score;
- ❖ **Threshold Optimization**: This optimization has occurred using adaptive threshold tuning through grid search across the range of 0.1 - 0.9 to optimise the macro F1.

3.5.2. Model Interpretability with SHAP

The post-hoc model interpretation used the SHapley Additive exPlanations (SHAP) methodology to evaluate the models

- ✓ Tree-based models used the TreeExplainer for fast calculation of the Shapley values.
- ✓ Kernel-based models used the KernelExplainer with the use of background sampling to compute the Shapley values.
- ✓ The global feature importance was evaluated by considering the Mean Absolute value of the SHAP (or Shapley) values.
- ✓ When examining each individual model prediction, a force plot was utilized which shows the top 20 contribution features (I.e., each of the top 20 features, are color-coded by contribution).

3.5.3. Validation Strategy

We ensured the validity of our system through:

- An evaluation of the final results using Hold-Out Testing (35% stratified random sample).
- Apply 5-Fold Stratified Cross-Validation in order to identify the best model and its corresponding Hyper-parameters.
- The Early and Late-Stage Cohorts should each have their own Validation Pipes so that no information is shared between them

This comprehensive method provided for a robust model development process with complete transparency in the evaluation of our models, resulting in clinically interpretable results for the prediction of breast cancer from multi-modal integrated analysis.

4. Result and Discussion

The results of the experiments based on the suggested model demonstrate important predictive performance enhancements in comparison to the control methods. To guarantee strength and generalizability, several performance measures were used to assess it. The results obtained indicate that the model is able to help capture the complex patterns and interactions within the data resulting in higher classification accuracy and stability between validation folds. In addition, the fact that both clinical and molecular level aspects were included gave a complete understanding of the disease progression hence justifying the capability of the model to make predictions beyond the training data.

4.1 Early Stage

Table I demonstrated that five machine learning models, had been compared with respect to several performance measures. XGBoost showed the best and the most stable performance with the highest Test Accuracy (0.9866) and AUROC (0.9967) which shows that it is the best machine in terms of discriminative capacity across the classes. Its PRAUC value (0.9981) also highlights its good precision recall balance thus being very useful in detecting in the early stages when false negativity is very important to be avoided.

Table I. Model Comparison Results (Early Stage)

Model	Test Accuracy	Test Precision	Test Recall	Test F1	PR_AUC	Test AUROC
Random Forest	0.7293	0.7003	0.9929	0.8213	0.9803	0.9646
XGBoost	**0.9866**	**0.9893**	**0.9893**	**0.9893**	**0.9981**	**0.9967**
SVM	0.6488	0.6447	0.9786	0.7773	0.7741	0.6817
Voting	0.8814	0.8697	0.9536	0.9097	0.9805	0.9676

Classifier						
Stacking Classifier	0.7718	0.8007	0.8464	0.8229	0.8975	0.8442

Random Forest model had a fairly high PRAUC (0.9803) and AUROC (0.9646) implying that it is effective in non-linear relationships, but it was a little weaker than XGBoost in terms of accuracy and overall precision. The Voting Classifier that consolidates the predictions of a series of models also demonstrated a prospective outcome with PRAUC = 0.9805 and AUROC = 0.9676 indicating the benefit of ensemble learning in improving the strength of prediction. The SVM model, on the other hand, showed somewhat lower results (PR_AUC = 0.7741, AUROC = 0.6817). The Stacking Classifier showed a moderate level of performance (PR_AUC = 0.8975, AUROC = 0.8442) which indicates that it partially improved with the help of meta-learning, but it remained worse than the performance of the boosting-based XGBoost model.

In general, according to the PR_AUC and the values of the AUROC represented in Table [Model Comparison Results (Early Stage)], the XGBoost proved to be the best model to predict early-stage breast cancer. Due to its gradient boosting capabilities, it has the ability to learn the complex interactions of multiple features and also efficiently minimize bias and variance. The findings suggest that XGBoost has the best credible and generalized performance in terms of early detection, which is vital to the successful diagnosis and better patient outcomes.

Through consideration of the ROC curve analysis shown in **Figure 2**, an additional level of understanding is gained regarding how the various models have performed through comparison between their TPR (True Positive Rate) and FPR (False Positive Rate) at a range of thresholds.

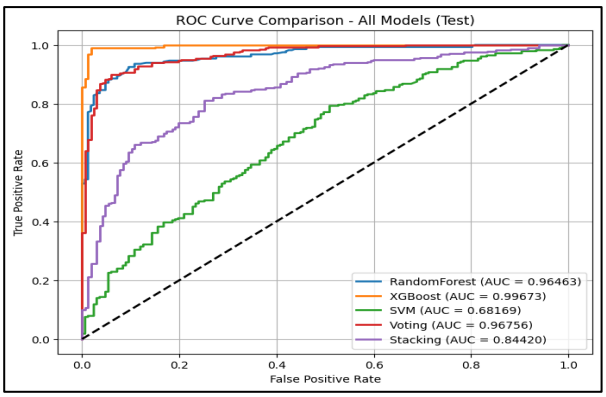

Figure 2. Analysis of ROC Curve Comparison for Classification Models

The Voting and Random Forest classification models also perform well on the ROC curves with AUC scores of 0.96756 and 0.96463. These curves highlight how well ensemble models handle complicated biological data when combined with the XGBoost ROC curve. However, the Stacking model performs slightly low with an AUC value of 0.84420. Despite this model performing better than random guessing classification, it still struggles as compared to other models. The SVM model performs the lowest with an AUC value of 0.68169 among all models as can be deduced from its ROC curve running close to the random guess classification line indicating weaker capability at distinguishing the two classes.

Figure 3 improves interpretability through a graph of the XGBoost model, which contains feature importance using SHAP for the model. This figure presents global feature importance and directionality and it quantifies the contributions made by each feature to the output of the model. The top three contributing features are Feature 10, Feature 8 and Feature 19, which had the largest ranges of SHAP values, indicating that they are the most highly predictive of the classifications achieved by the model. On the other hand, the features toward the bottom of the ranking (such as Features 1480 and 1861) had lower SHAP values, indicating that they made a lesser contribution to the XGBoost model's overall predictive quality.

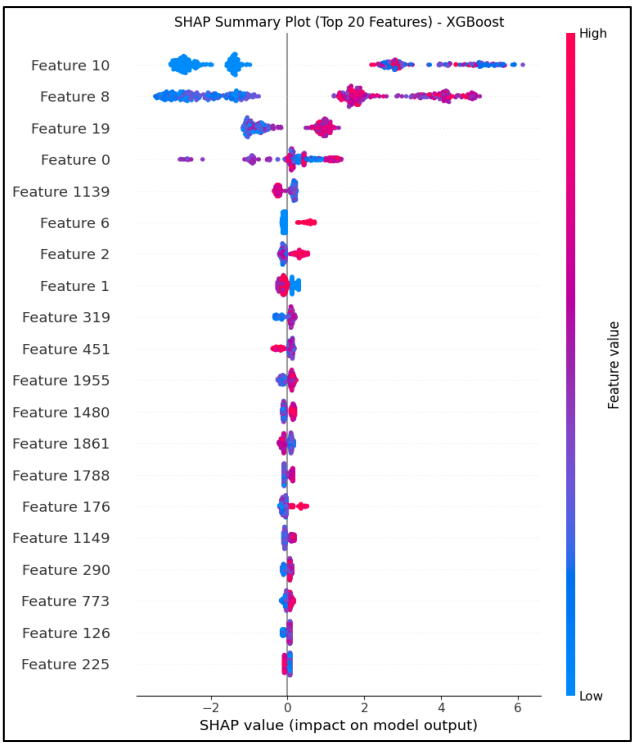

Figure 3. SHAP Analysis of the Optimal XGBoost Model

The SHAP colour distribution is further evidence of which direction influences go. Higher feature values for Feature 10 and Feature 8 (red points) have an even larger effect on moving predictions towards the positive class than lower values and therefore are major contributors to successful early-stage detection. Feature 19's distribution is a mix of both patterns, while Feature 0 demonstrates that the features that are ranked lower tend to have a similar but less pronounced high positive or low negative influence pattern for prediction.

Overall, the outcomes from SHAP demonstrate that a handful of features are responsible for driving most of the predictions made by the XGBoost model, which indicates that future efforts toward feature refinement and data acquisition should emphasise these dominant variables to improve the performance of the model and improve the clinical/biological representation of the predicted outcomes.

4.2 Late Stage

Table II is the main source of information for investigating how well the quantitative models perform. The graphs (Figures 4A and 4B) give visual indications of the ability to discriminate between and interpret the quantitative models.

Table II. Model Performance Results (Late Stage)

Model	Accuracy	ROC_AUC	PR_AUC	Best_Threshold
Random Forest	0.931818	0.410569	0.072597	0.50
XGBoost	0.886364	0.483740	0.090456	0.15
SVM	**0.931818**	**0.577236**	**0.102778**	**0.25**
Voting	0.931818	0.487805	0.084110	0.50
Stacking	0.909091	0.487805	0.087756	0.55

When evaluating models in the Late Stage with unbalanced datasets, it is discovered that the Area Under the ROC Curve (ROC_AUC) is the most powerful discriminator. The SVM model possesses the highest value for ROC_AUC at 0.577236, with a significant edge over other classifier. The Area Under the Precision-Recall Curve (PR_AUC), which is substandard for all models owing to the rarity of the positive class, retains the highest PR_AUC value for SVM at 0.102778, implying a superior capability for the detection of the minority class despite the low number of positive examples detected as true positives. As inferred from the evaluation criteria, the choice for the model with the strongest performance leans towards SVM.

Figure 4A depicts the comparative analysis of the ROC curve, which also corroborates the data listed in Table 1. The diagonal line found on the ROC curve represents random classification; therefore, its value is 0.5 (i.e., the area under the ROC curve is 0.5) and it serves as a reference point for determining how well or poorly the other classifiers performed compared to the random classifier.

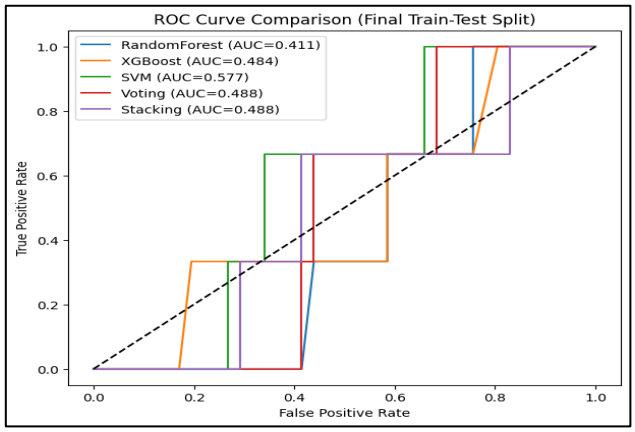

Figure 4A. ROC Curve Comparison

The ROC curve associated with the SVM model has consistently had the highest plotted values across each iteration when compared to all other classifiers, thus supporting and substantiating the AUC value of 0.577; even though this is only slightly greater than what could be achieved consistently with random chance, the AUC value of the SVM was greater than that produced by the ensemble and tree-based models, which had AUC values near or below that of the random baseline (Random Forest was equal to approximately 0.411; XGBoost was equal to approximately 0.484; Voting and Stacking were equal to approximately 0.488). Therefore, the combination of these points further confirms that the SVM model is by far the best of the models used in this analysis.

Interpretability and performance evaluation are included in the SHAP of the SVM model in **Figure 2B.**

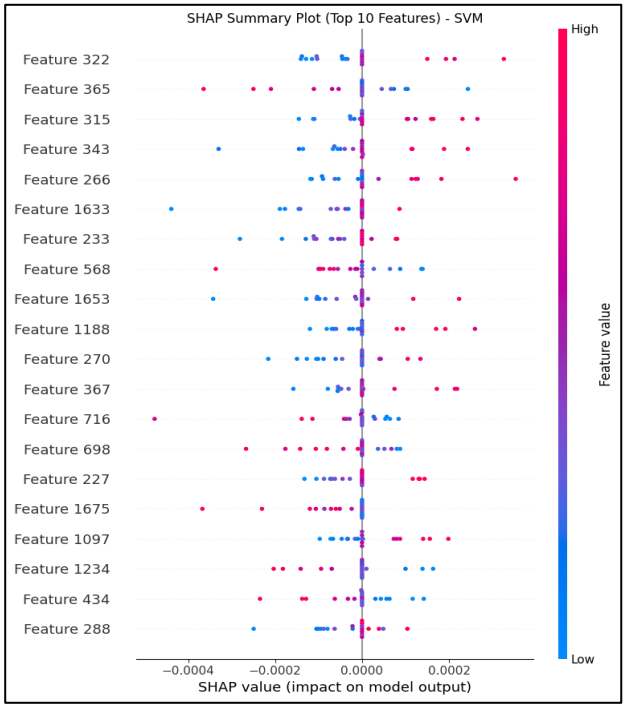

Figure 4B. SHAP Analysis for SVM

The summary plot is very helpful for interpreting the effect of the first 20 most informative features on the model's result. The SHAP values shown on the x-axis represent the local contribution of the features towards the result of the

model, based on the predicted logarithmic odds. The SHAP values result in either predictions of the higher class, based on positive contributions, or predictions of the lower class, based on negative contributions. This summary plot helps in understanding which feature is having what type of relationship with the model output. For example, Feature 322, Feature 365, etc., are having a positive relationship, as high values of these features result in high-class predictions. However, feature 343 is having a negative relationship, as high values of this feature cause predictions to incline towards the low-class outcome. Some other features, such as Feature 315, tend to create a mixed effect, i.e., a nonlinear relationship between Feature 315 and the model output. SHAP helps in understanding the contribution of the above-mentioned features in the SVM model.

Therefore, taken together, performance metrics, ROC curves, and SHAP interpretability all point to the challenge posed by severe class imbalance in the dataset. SVM is the best classifier for the job at hand since it demonstrates comparatively superior discriminating capabilities and feature dependency patterns across the models examined. His interpretability provided by SHAP further justifies the validity of this choice by highlighting the key features which drive the SVM predictions, hence providing both performance-based and explainability-based justification for adopting the model.

4.3 Final Comparison Between the Best Early-Stage and Late-Stage Models

The Precision-Recall Curve Comparison, Figure 5, explicitly highlights a clear point of difference in performance for both the best Early-Stage model and the Late-Stage model. It is observed that there is a considerably higher Predictive Model capacity for the Early Stage XGBoost model, represented by a PR_AUC score of 0.50000, while for the Late Stage SVM model, there is a considerably lower PR_AUC score of 0.07399, indicating a nearly seven-fold increase in PR_AUC scores.

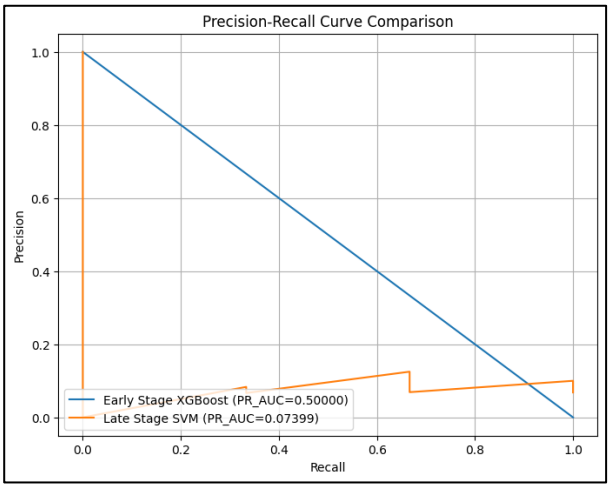

Figure 5. Comparison of Model Performance (Early Vs Late)

The form of the curves adds to this disparity. The Early Stage XGBoost has a significantly larger area under the precision-recall graph, together with a higher precision level at various levels of recall, especially when recall is low. However, the Late Stage SVM plot is largely stuck at the same level as the baseline, showing its inefficiency in making correct positive classifications in situations that are largely imbalanced. The initial abrupt decline in the precision level, together with the overall flatness of the graph, lend insight into the difficulty of utilizing information from the late stages in the classification of the minority class. The difference between early and late-stage data tells us that early information can better predict the future; therefore, it allows XGBoost-an ensemble learning algorithm to create better models of how classes will be separated through Noise and Non-Linearity because they will have access to the most information. Late data does not seem to provide sufficient information for SVM to classify data into positive classifications, no matter how you may use kernel methods of establishing decision boundaries.

In summary, from the Precision-Recall Curve shown in **Figure 5**, the Early Stage XGBoost model clearly outshines all others and is, therefore, the best classifier. The greater PR_AUC and favourable graph progression of the Early Stage XGBoost model demonstrate a much better precision and recall, which is a critical requirement in imbalanced classifier problems. The observations clearly show the superiority of early-stage characteristics, when accurately captured with the aid of the XGBoost algorithm, over the characteristics captured with the SVM model.

Conclusion, Limitations and Future Studies

This analysis makes it clear that the prediction for breast cancer stages can be made more accurately by integrating traditional and gene expression information. The Early Stage XGBoost classifier was more effective than the Late Stage SVM model because the PR_AUC value for the Early Stage XGBoost classifier is greater than that of the Late Stage SVM model. There is more discriminate information for the early stages, and this makes it more credible to accurately select the individuals for treatment. Moreover, the SHAP results provided by the interpretation technique made it easier to identify the crucial molecular and traditional predictors for breast cancer.

Despite the encouraging outcomes, there are some drawbacks that need to be considered. There is an unbalanced class in the dataset, and this is more apparent in the late-stage samples, leading to suboptimal performance in the Late Stage SVM model. Even though dimensionality reduction was used, the nature of high-dimensional gene expression data makes it vulnerable to the possibility of overfitting when dealing with small sample sizes. In addition, although SHAP values offer insights into the results, it cannot be ascertained that the biological relevance of the features extracted at the gene expression level without considering any actual validation experiments.

One area that should be investigated in future studies is the combination of further multi-omics data, so as to provide an

even more complete view of tumour biology. The use of Graph Neural Networks in modeling the complicated interactions between genes may aid in improving the biological relevance and biological interpretability. Addition of models that involve Explainable AI, privacy-preserving federated learning, and survival analysis models would add an extra layer of improvement. Indeed, the ultimate test in realistic clinical settings will be required in efforts to move forward with personalizing breast cancer prediction and treatment by cancer stage.

References

1. Islam, Taminul, et al. "Predictive Modeling for Breast Cancer Classification in the Context of Bangladeshi Patients by Use of Machine Learning Approach with Explainable AI." *Scientific Reports*, vol. 14, no. 1, 2024, p. 8487.
2. Kittaneh, Muaiad, Alberto J. Montero, and Stefan Glück. "Molecular Profiling for Breast Cancer: A Comprehensive Review." *Biomarkers in Cancer*, vol. 5, 2013, pp. BIC-S9455.
3. Güler, E. Nilüfer. "Gene Expression Profiling in Breast Cancer and Its Effect on Therapy Selection in Early-Stage Breast Cancer." *European Journal of Breast Health*, vol. 13, no. 4, 2017, p. 168.
4. Curtis, Christina, et al. "The Genomic and Transcriptomic Architecture of 2,000 Breast Tumours Reveals Novel Subgroups." *Nature*, vol. 486, no. 7403, 2012, pp. 346–352.
5. Pereira, Bernard, et al. "The Somatic Mutation Profiles of 2,433 Breast Cancers Refine Their Genomic and Transcriptomic Landscapes." *Nature Communications*, vol. 7, no. 1, 2016, p. 11479.
6. Berger, Ashton C., et al. "A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers." *Cancer Cell*, vol. 33, no. 4, 2018, pp. 690–705.
7. Kate, Rohit J., and Ramya Nadig. "Stage-Specific Predictive Models for Breast Cancer Survivability." *International Journal of Medical Informatics*, vol. 97, 2017, pp. 304–311.
8. Said, Ahmed Attia, et al. "Stage-Specific Predictive Models for Main Prognosis Measures of Breast Cancer." *Future Computing and Informatics Journal*, vol. 3, no. 2, 2018, pp. 391–397.
9. Lin, Yixuan, et al. "Impact of Screening on Late-Stage Breast Cancer in the Netherlands: A Population-Based Cohort Study (2007–2016)." *The Breast*, 2026, p. 104745.
10. Trewin, Cassia Bree, et al. "Socioeconomic Inequalities in Stage-Specific Breast Cancer Incidence: A Nationwide Registry Study of 1.1 Million Young Women in Norway, 2000–2015." *Acta Oncologica*, vol. 59, no. 11, 2020, pp. 1284–1290.
11. Woods, Ryan R., et al. "Stage-Specific Risk of Breast Cancer among Canadian Immigrant and Non-Immigrant Women." *Journal of Immigrant and Minority Health*, vol. 25, no. 1, 2023, pp. 232–236.
12. Jacklyn, Gemma, et al. "Trends in Stage-Specific Breast Cancer Incidence in New South Wales, Australia: Insights into the Effects of 25 Years of Screening Mammography." *Breast Cancer Research and Treatment*, vol. 166, no. 3, 2017, pp. 843–854.
13. Yoon, Hyuna, et al. "Long-Term Survival Prediction in Older Women with Stage I–II Breast Cancer Using Decision Tree-Based Machine Learning." *Journal of Geriatric Oncology*, vol. 17, no. 2, 2026, p. 102828.
14. Das, Sonali Mondal, and Abhoy Chand Mondal. "Stage-Specific Survival Prediction in Breast Cancer: A Survey." *International Journal of Creative Research Thoughts*, vol. 13, no. 9, 2025, pp. e447–e454
15. Mondal Das, Sonali, and Abhoy Chand Mondal. "Studies of Cancer Prediction Using Machine Learning: A Survey." *Saudi Journal of Engineering and Technology*, vol. 13, no. 2, Feb. 2025, pp. 69–72. DOI**:** 10.36347/sjet.2025.v13i02.001.