An Efficient and Interpretable Machine Learning Model for Classifying Breast Cancer Subtypes Using Gene Expression Profiles

Tareque Mohmud Chowdhury
Computer Science and Engineering Department, Islamic University of Technology, Dhaka, Bangladesh
tareque@iut-dhaka.edu (corresponding author)

Abu Raihan Mostofa Kamal
Computer Science and Engineering Department, Islamic University of Technology, Dhaka, Bangladesh
raihan.kamal@iut-dhaka.edu

Received: 28 March 2025 | Revised: 26 April 2025 and 5 May 2025 | Accepted: 8 May 2025

 DOI: https://doi.org/10.48084/etasr.11179

ABSTRACT

Breast Cancer (BRCA) is a complex and heterogeneous disease. This heterogeneity has been shown to affect gene expression patterns and molecular activity of different subtypes in different ways. BRCA subtype identification is of critical importance in the context of prognosis and treatment decisions for the disease. Advances in transcriptomic profiling and Machine Learning (ML) models have enabled the classification of BRCA subtypes with higher accuracy, yet the majority of classification models lack interpretability, thereby limiting their clinical applicability. In this study, an interpretable ML framework for classifying BRCA subtypes is proposed using high-dimensional RNA-sequencing data. The framework was evaluated using a publicly available TCGA transcriptomic dataset, by applying dimensionality reduction techniques and optimizing ML models through grid search tuning. Shapley Additive Explanations (SHAP) values are used to find important transcriptomic markers that facilitate the classification of subtypes. This approach provides insights into the gene sets associated with the molecular mechanisms of each subtype. The experimental results demonstrate that the proposed method exhibits superior performance in terms of accuracy, precision, F1-score, and interpretability when compared to existing works. Finally, the gene set enrichment analysis highlights key pathways associated with BRCA and its subtypes.

Keywords-Breast Cancer (BRCA); BRCA subtypes; interpretable AI; ML; subtype classification

I. INTRODUCTION

Breast Cancer (BRCA), a heterogeneous disease, is the most prevalent cancer among women in 157 out of 185 countries and the second leading cause of cancer-related deaths among women worldwide according to World Health Organization (WHO). It poses a significant global health burden. BRCAs are traditionally diagnosed using Machine Learning (ML) methods and histological data [1, 2]. However, with the advent of high-throughput RNA-sequencing technology, the diagnosis and prognosis of BRCA, along with molecular subtyping, have significantly improved. This leads to better personalized treatment for patients. An accurate, interpretable ML classifier is essential for identifying the key genes responsible for BRCA subtype regulation. This remains a challenge due to the complex, high-dimensional nature of the transcriptomic data.

ML has revolutionized cancer research by enabling the analysis of high-dimensional genomic datasets to reveal patterns that drive cancer development, progression, and response to treatment. Gene expression data, generated via RNA-sequencing, capture the activity of thousands of genes, offering a molecular snapshot of tumor biology. However, the complexity, noise, and dimensionality of these datasets pose significant challenges that ML models are uniquely equipped to address.

PAM50 [3] is a 50-protein coding gene (mRNA) signature that classifies BRCA into five molecular intrinsic subtypes: basal-like, HER2-enriched, luminal A, luminal B, and normal-like. PAM50 utilizes protein-coding gene expression data (mRNA) generated by old-fashioned microarray technology. Interestingly, at the cellular level, mRNA constitutes only one-third of the total transcriptomic volume. MicroRNA (miRNA), long noncoding RNA (lncRNA), and many other types of RNA play crucial roles in regular and irregular (cancerous) cellular activities. Hence, analyzing only mRNA (single omics data) does not perceive the full spectrum of the cellular functionality at different levels of a disease state, which may prevent comprehensive insight into the biological processes of BRCA subtypes. Several studies have reported this issue with the

PAM50 method [4, 5]. Researchers are now using RNA-sequencing full transcriptomic data, which is more accurate and provides more insight about intrinsic cellular activities. In addition, DNA-specific information, such as DNA methylation (DNAm), Copy Number Alteration (CNA), and Copy Number Variation (CNV) also affects the expression level of genes. These DNA-specific data have also helped many researchers to understand disease-derived changes in cellular activities.

Some studies have used single omics data to identify and diagnose BRCA subtypes. Authors in [6] implemented a DNAm-based neural network approach and reported an average accuracy of 84.20%. They used semi-supervised learning on two datasets: labeled TCGA and unlabeled GEO. Authors in [7] proposed a modification of PAM50, naming it MPAM50. The proposed method classifies many samples into different subtypes compared to PMA50. For a combination of 19 datasets, the MPAM50 and PAM50 methods exhibited 79.20% similarity in classification. Authors in [8] proposed an mRNA-based deep learning model, called Cascade Flexible Neural Forest (CFNForest) and achieved an average accuracy of 94.40%. Though they achieved a very high accuracy, they only considered four subtypes of BRCA. In [9], a regularized Logistic Regression (LR) model was proposed using mRNA. The authors trained their model using four datasets (TCGA, GSE81538, GSE96058, and PanCA) and achieved an average accuracy of 90%.

Some studies also use a combination of two-omics data. Authors in [10] proposed a correlation analysis-based Deep Neural Network (DNN) classification model that uses mRNA expression and DNA methylation data. They achieved an average accuracy of 90.60% and an F1-score of 92.20%. In [11], a DNN model was developed to be trained using CNA and mRNA expression data. This study achieved 79.20% accuracy with the full set of features and 76.6% with the top 500 contributing features selected from the trained model. Authors in [12] developed a 2-norm Support Vector Machine (SVM) model and trained it with mRNA and lncRNA data downloaded from TCGA. They trained the model in three different scenarios: using mRNA alone, using lncRNA alone, and using mRNA and lncRNA combined. Their experiments achieved average accuracies of 87.6%, 87.8%, and 88.5%, respectively.

Some researchers combine three-omics data for BRCA subtype classification. Authors in [13] implemented a number of ML models by integrating mRNA, CNV, and DNAm data from the TCGA repository. Their experiments showed that SVM performed best in terms of accuracy. They also used Shapley Additive Explanations (SHAP) to interpret the model to identify key features related to classification tasks. Different deep learning models were proposed in [14-16] using mRNA, miRNA, and DNA methylation data downloaded from TCGA, achieving accuracies of 89.1%, 82.9%, and 84.6%, respectively. Authors in [17] proposed a deep learning model using mRNA, CNV, and DNA methylation data to classify BRCA subtypes with 78.2% accuracy.

The existing work demonstrated an accuracy of approximately 90% for the classification of BRCA subtypes. The objective of this study is to develop a robust integrative framework to increase the accuracy of BRCA subtype classification using multi-omics gene expression data. Furthermore, the objective is to establish a balance between classification accuracy and biological understanding in the context of BRCA subtyping. By leveraging the interpretability of ML models, this study demonstrates that the proposed classification model improves upon the accuracy of existing methods while offering clear, clinically useful explanations of the reasons behind its predictions. The selected biomarkers are systematically verified against known published literature and pathway analysis, supporting their involvement in subtype-specific oncogenesis. The integration of strict computational methods and systematic understanding contributes to the utilization of genomic information in clinical decision support systems, which could facilitate the improvement of BRCA diagnosis and treatment precision.

II. MATERIAL AND METHODS

A. Dataset Preparation and Preprocessing

The RNA-sequencing transcriptome data (FPKM normalized) for BRCA were obtained from the publicly available TCGA [18] repository using the TCGAbiolinks [19] package in R [20]. Samples for which the sample type was not "Primary Tumor" were removed, leaving a total of 1,099 tumor samples with 60,660 features. Features with very low expression value (average value < 0.04) were removed. This step reduces the feature count to 34,394. Given the sparsity of the expression values in the prepared dataset, the $\log_2$ transformation was applied to render them in a more linear form. To prevent the occurrence of NaN errors during the process of log transformation, numeric 1 was added to each value before applying the log. Based on the PAM50 metadata value of the samples, the intrinsic subtype was assigned to each sample. The summary of the prepared dataset is presented in Table I.

TABLE I. NUMBER OF SAMPLES PER SUBTYPE AND FEATURE DISTRIBUTION IN THE DATASET.

BRCA tumor subtypes					
Basal	**HER2**	**Lum A**	**Lum B**	**Normal**	**Total**
197	82	571	209	40	1,099
Feature space					
mRNA	**miRNA**	**lncRNA**	**otherRNA**	**Total**	
16,967	624	8,291	8,516	34,294	

B. Feature Selection

RNA-sequencing gene expression data contain thousands of features (genes), compared to a very low sample count. A successful feature selection process reduces dimensionality by selecting the most informative genes or eliminating irrelevant or noisy features. This process improves computational efficiency and enhances the performance of the model, which can otherwise hinder classification accuracy and generalization. Additionally, reducing the number of features helps prevent overfitting, a common issue with ML models working with high-dimensional data. For this study, a multi-view feature selection method [21] is used, which was specifically proposed for transcriptomic data. The sequence of data processing steps is illustrated in the experiment workflow depicted in Figure 1.

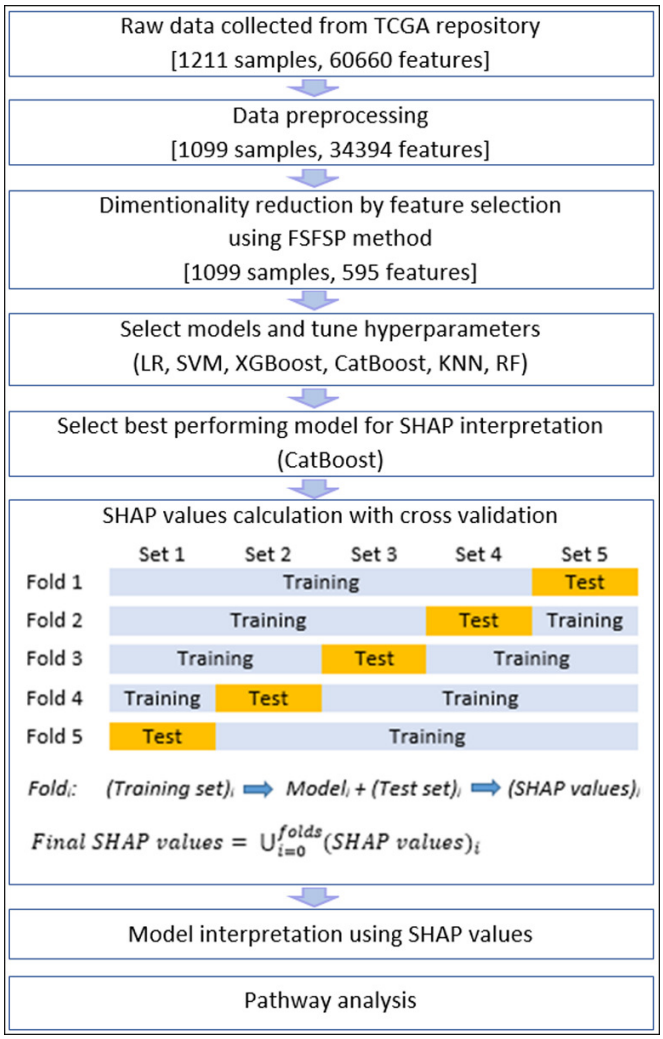

Fig. 1. Detailed workflow of the experiment performed in this study.

C. Classifier Models

Six state-of-the-art ML classification models were selected and implemented to classify the BRCA subtype samples. These models are LR [22], SVM [23], Extreme Gradient Boosting (XGBoost) [24], CatBoost [25], K-Nearest Neighbors (KNN) [26], and Random Forest (RF) [27].

LR is a supervised ML algorithm that is widely used for classification tasks. LR has proven effective in classifying biological data. SVM identifies the optimal boundary that distinguishes different classes of samples within the data. Thus, it maximizes the distance between classes. Because of this, SVM is ideal for solving complicated classification problems. SVM is a very efficient model that constructs many small decision trees, each of which is incrementally refined through previous errors. This makes SVM fast and accurate with large amounts of data. CatBoost is another boosting algorithm that excels at handling different types of data. CatBoost is easy to use and requires less tuning. KNN classifies data by examining the closest examples. It compares new data to nearby classes and selects the most common class. RF builds on the strategy of creating many decision trees and combining their results. It takes a vote from all the sub-trees to make the final decision. This makes RF models more stable and less likely to make mistakes.

D. Hyperparameters

Hyperparameter tuning is a crucial step in adapting an ML model to a specific task. The random search approach was applied to fine-tune the hyperparameters and enhance model performance for dataset classification. Table II presents the hyperparameter values obtained for the utilized models used in this study.

E. Models Performance Evaluation Metrics

The classification results of the models are evaluated using the confusion matrix and a number of other statistical measures derived from the confusion matrix, including accuracy, precision recall, and F1-score. The following equations are employed to calculate the metrics:

$$Accuracy = \frac{TN+TP}{TN+TP+FN+FP} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1-score = \frac{2*Precision*Recall}{Precision+Recall} \quad (4)$$

where TP signifies true positive, TN signifies true negative, FP signifies false positive, and FN signifies false negative.

F. Model Interpretation using Shapley Additive Explanations

SHAP [28] is an Explainable AI (XAI) methodology based on game theory that offers interpretations of ML models to elucidate the decision-making process. The SHAP value explanation method incorporates a number of intriguing properties, including local accuracy, handling of missing data, and consistency. This approach involves decomposing the prediction into a linear collection of feature contributions, as outlined in (5).

$$g(z') = \emptyset_0 + \sum_{i=0}^{F} \emptyset_i z_i \quad (5)$$

where g denotes the explanation formula, F is the number of features, z' is a vector that indicates the presence or absence of a feature, and $\emptyset_i$ is the SHAP value of feature i, which can be estimated using (6). This requires the evaluation of all possible subsets of features, both with and without the i-th feature. The $val(\cdot)$ function returns the prediction of each subset.

$$\emptyset_i = \sum_{S \in F\{i\}} \frac{|S|!(f-|S|-1)!}{f!} [val(S \cup \{i\}) - val(S)] \quad (6)$$

To effectively measure the performance of the model used in this cross-validated experiment, SHAP values were calculated for each fold. Subsequently, the SHAP values were aggregated through a union operation to have SHAP values for every sample of the dataset. The calculation process of SHAP values is outlined in (7) and illustrated in the workflow of the experiment (Figure 1).

$$Final\ SHAP\ values = \cup_{i=0}^{folds} (SHAP\ values)_i \quad (7)$$

TABLE II. OPTIMIZED HYPERPARAMETERS FOR TRAINING THE MODELS

Parameter name	Models					
	CatBoost	LR	KNN	RF	SVM	XGBoost
boosting type	'plain'	-	-	-	-	-
bootstrap	-	-	-	false	-	-
criterion	-	-	-	'entropy'	-	-
C	-	20	-	-	1	-
depth	6	-	-	-	-	-
kernel	-	-	-	-	'linear'	-
l2_leaf_reg	3	-	-	-	-	-
learning_rate	0.04	-	-	-	-	0.1
max_depth	-	-	-	12		4
max_features	-	-	-	'sqrt'	-	-
min_samples_leaf	-	-	-	1		-
min_samples_split	-	-	-	2		-
n_estimators	-	-	-	100	-	1000
n_neighbors	-	-	7	-	-	-
penalty	-	'l2'	-	-	-	-
solver	-	'lbfgs'	-	-	-	-
depth	-	-	-	-	1e-5	-

III. RESULTS AND DISCUSSION

A. Dimensionality Reduction

The applied feature selection method significantly reduces the dimensionality of the dataset, selecting only 595 informative and valuable features from a large set of 34,394. Among the selected features, there are 495 mRNA, 1 miRNA, 98 lncRNA, and 21 otherRNA.

B. Classification Results

In a 5-fold cross-validation experiment using the selected features, the CatBoost model outperformed the other five ML models. The evaluation metrics for each model are calculated as an average of the respective metrics of the folds.

Table III presents the precision, recall, and F1-score of the ML models. It is evident that the CatBoost model significantly outperforms the other models. The F1-score is a favorable metric for evaluating classification models on imbalanced datasets because it balances precision and recall, offering a more robust measure of performance than accuracy alone.

TABLE III. COMPARISON OF PERFORMANCE MEASURES OF VARIOUS ML MODELS

Model-wise overall classification performance (%)			
Models	**Precision**	**Recall**	**F1-score**
LR	91.23	90.99	90.97
SVM	91.26	90.90	90.87
XGBoost	90.99	91.45	90.87
CatBoost	92.55	92.54	92.33
KNN	89.00	88.81	88.25
RF	91.04	90.99	90.25

The class-wise performance of the CatBoost model is illustrated in Table IV. The performance of the normal-like subtype is lower than that of the other subtypes. This study identifies two reasons for this outcome. First, the normal-like subtype has fewer samples in the dataset than other subtypes, with only 40 samples, as shown in Table I. Second, the model's confusion matrix, shown in Figure 2, confirms that the normal-like subtype is mostly misclassified as the luminal-A subtype. This phenomenon has been reported previously in [29].

TABLE IV. EVALUATION METRICS FOR EACH CLASS FOR THE BEST PERFORMING MODEL.

Subtype classification performance by CatBoost (%)			
Subtypes	**Precision**	**Recall**	**F1-score**
Basal	98.48	98.48	98.48
HER2	93.24	84.15	88.46
Luminal A	92.89	96.15	94.49
Luminal B	86.60	86.60	86.60
Normal-like	85.71	60.00	70.59

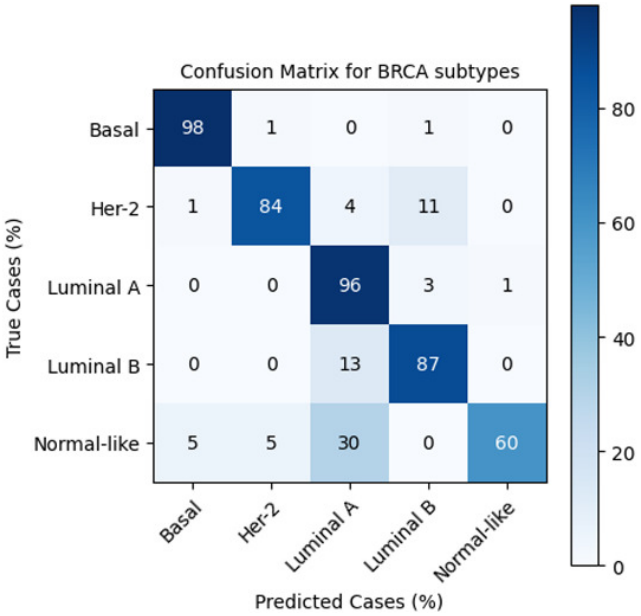

Fig. 2. Confusion matrix for CatBoost model results.

C. Comparative Study

In the literature review, 12 research works [6-17] focusing on the classification of BRCA subtypes using different omics

data were analyzed. The results obtained from this study are compared with those from the other studies and are shown in Table V. This comparative study confirms the effectiveness and importance of the proposed workflow for accurately identifying and diagnosing different BRCA subtypes.

TABLE V. PERFORMANCE COMPARISON OF THE RESULTS

Ref. and dataset	Omics data used	Accuracy (%)	F1-score (%)
[6] TCGA+GEO	DNAm	84.2	83.1
[7] 19 GEO datasets	mRNA	79.2	-
[8] TCGA	mRNA	94.4	-
[9] TCGA+GEO	mRNA	90.0	-
[10] TCGA	mRNA, DNAm	90.6	92.2
[11] METABRIC	CNA, mRNA	79.2	-
[12] TCGA	mRNA, lncRNA	87.6	87.8
[13] TCGA	mRNA, CNV, DNAm	83.5	83.1
[14] TCGA	mRNA, miRNA, DNAm	89.1	88.7
[15] TCGA	mRNA, miRNA, DNAm	82.9	-
[16] TCGA	mRNA, miRNA, DNAm	84.6	-
[17] TCGA	mRNA, CNV, DNAm	78.2	-
[Proposed] TCGA	mRNA, miRNA, lncRNA, otherRNA	92.6	92.3

a. In [8], 4 subtypes of BRCA were used in the experiment. The normal-like subtype, which is difficult to classify, was excluded.

D. GO Pathway Enrichment Analysis

To further validate the 495 protein-coding mRNAs selected in this study, a gene set enrichment analysis was performed using the ShinyGo [30] tool for the KEGG [31] pathway search, setting the FDR cut-off value to 0.05. The top enriched pathways identified are presented in Table VI in order of decreasing fold enrichment value. The associated references suggest that these pathways are highly related to cancer, BRCA, and different BRCA subtypes.

TABLE VI. TOP TEN ENRICHED PATHWAYS IDENTIFIED BY SHINYGO USING SELECTED MRNAS.

Sl.	Pathways
1	Anaphase-promoting complex binding [32]
2	Plus-end-directed microtubule motor activity [33]
3	Histone kinase activity [34]
4	DNA replication origin binding [35]
5	Microtubule motor activity [36]
6	Cyclin-dependent protein serine/ threonine kinase regulator activity [37]
7	Microtubule binding [38]
8	Cytoskeletal motor activity [39]
9	Tubulin binding [40]
10	Kinase binding [41]

E. Model Interpretation using Shapley Additive Explanations Values

SHAP values offer an interpretation of an ML model by assigning values to the individual features that contribute to the prediction probability of each class. In this study, the SHAP values were accumulated for each cross-validation fold using (7) for all samples in the dataset. Figures 3-7 illustrate the subtype-specific features listed on the y-axis in descending order by the mean SHAP value across samples. Figure 8 shows the top contributing features for overall classification. The gene responsible for coding the Estrogen Receptor (ER) is known as ESR1. This protein plays a crucial role in the development and progression of BRCA. According to Shapley's interpretation, ESR1 is among the top 30 genes for overall classification and among the top 10 genes in all five subtypes of BRCA.

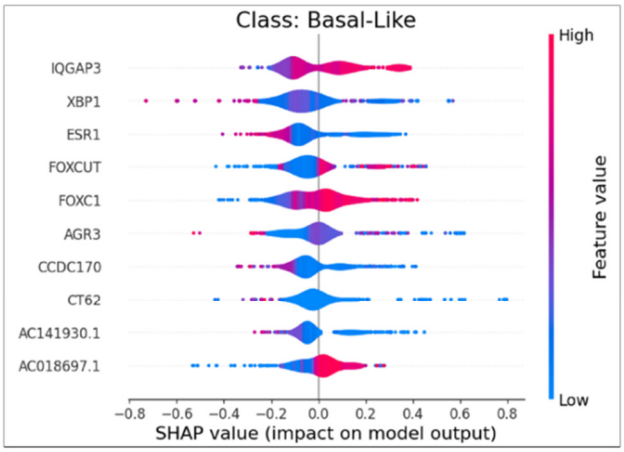

Fig. 3. SHAP violin plot of the Basal-Like BRCA class.

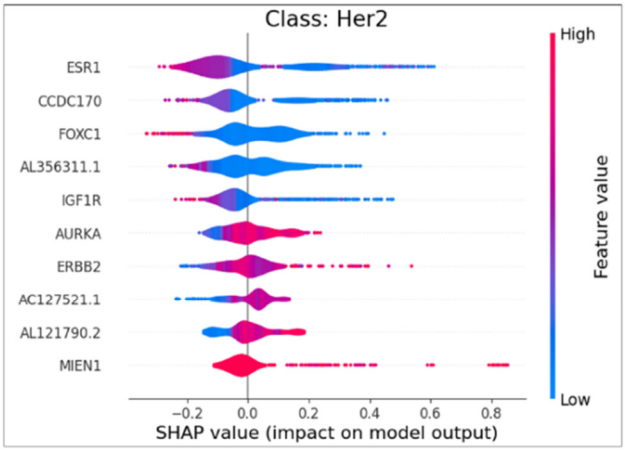

Fig. 4. SHAP violin plot of the Her2 BRCA class.

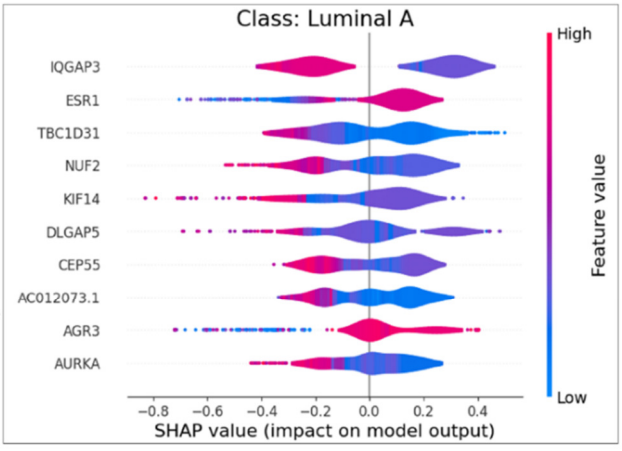

Fig. 5. SHAP violin plot of the LuminalA BRCA class.

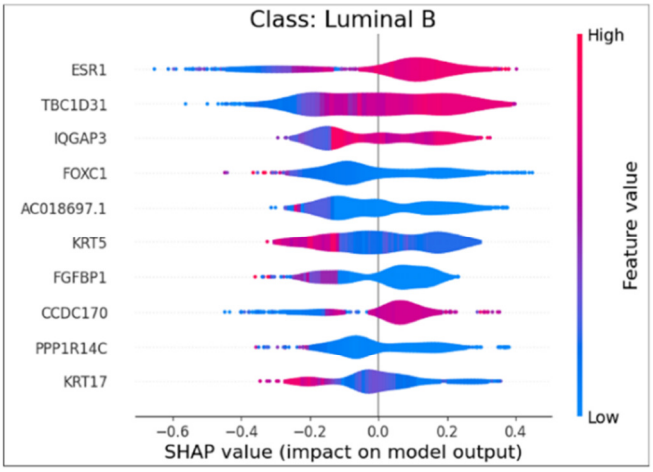

Fig. 6. SHAP violin plot of the LuminalB BRCA class.

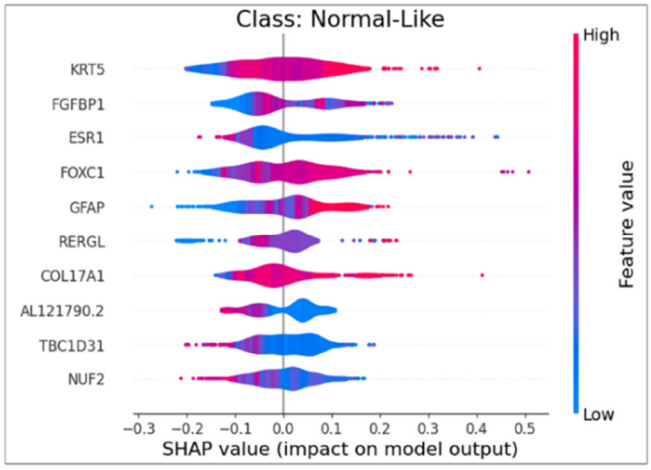

Fig. 7. SHAP violin plot of the Normal-Like BRCA class.

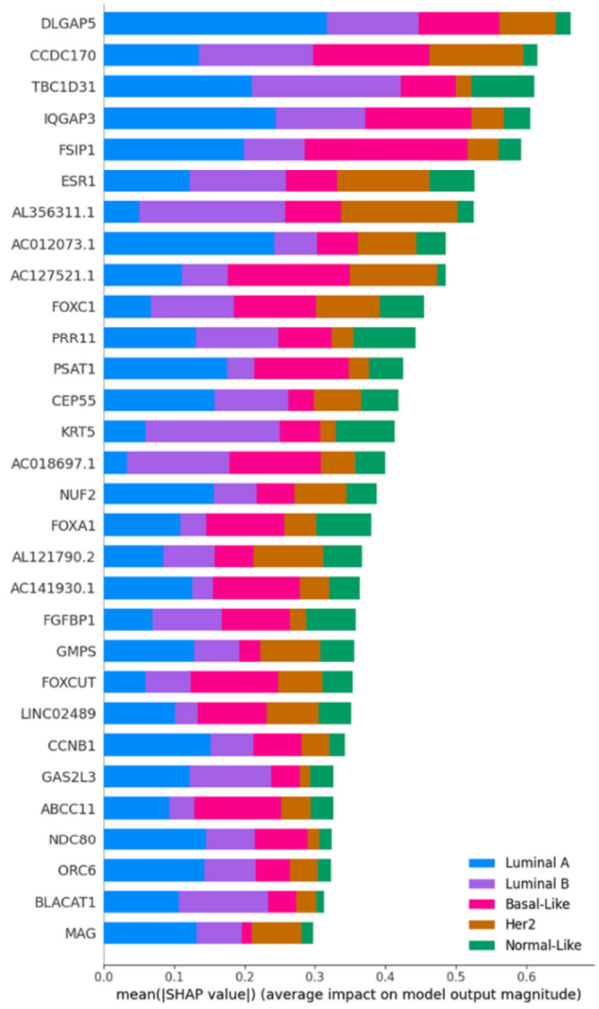

Fig. 8. SHAP summary plot representing the top 30 genes that contribute to BRCA subtype classification globally.

IV. CONCLUSION

Early detection of Breast Cancer (BRCA) and the identification of the molecular subtype of BRCA are crucial to the proper treatment of a patient. Molecular subtypes are defined by gene expression patterns and cellular functions, which traditional histopathology cannot fully capture. However, by utilizing high-throughput RNA-sequencing gene expression data, Machine Learning (ML) models offer a valuable means to analyze the molecular state of cancer cells. Furthermore, Explainable AI (XAI) technology has the potential to facilitate the identification of influential genes responsible for the molecular state of a cancerous cell through the implementation of a robust ML model. This study explores the implementation of an interpretable ML model for the classification of BRCA subtypes utilizing RNA-sequencing multi-omics data. The model demonstrated an accuracy of 92.54%, surpassing the performance of previous approaches in identifying the five BRCA subtypes. The model exhibits robust performance, achieving a precision of 92.55% and an F1-score of 92.33%, signifying substantial advancements over existing studies. These findings contribute to the early and precise detection of BRCA subtypes while also identifying critical genes associated with the disease, advancing precision oncology. Ultimately, this may positively influence patient survival rates and improve overall quality of their life.

DATASET AVAILABILITY

The prepared dataset utilized in this study and the necessary codes to reproduce it were published and made publicly available by the authors in the Harvard Dataverse [42].

REFERENCES

[1] R. Gonzales Martinez and D.-M. van Dongen, "Deep learning algorithms for the early detection of breast cancer: A comparative study with traditional machine learning," *Informatics in Medicine Unlocked*, vol. 41, Jan. 2023, Art. no. 101317, https://doi.org/10.1016/j.imu.2023.101317.

[2] A. Bekkouche, M. Merzoug, M. Hadjila, and W. Ferhi, "Towards Early Breast Cancer Detection: A Deep Learning Approach," *Engineering, Technology & Applied Science Research*, vol. 14, no. 5, pp. 17517–17523, Oct. 2024, https://doi.org/10.48084/etasr.8634.

[3] J. S. Parker *et al.*, "Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes," *Journal of Clinical Oncology*, vol. 27, no. 8, pp. 1160–1167, Mar. 2009, https://doi.org/10.1200/JCO.2008.18.1370.

[4] A. Fernandez-Martinez *et al.*, "Limitations in predicting PAM50 intrinsic subtype and risk of relapse score with Ki67 in estrogen receptor-positive HER2-negative breast cancer," *Oncotarget*, vol. 8, no. 13, pp. 21930–21937, Feb. 2017, https://doi.org/10.18632/oncotarget.15748.

[5] P. Turova *et al.*, "The Breast Cancer Classifier refines molecular breast cancer classification to delineate the HER2-low subtype," *npj Breast Cancer*, vol. 11, no. 1, Feb. 2025, Art. no. 19, https://doi.org/10.1038/s41523-025-00723-0.

[6] J. M. Choi, C. Park, and H. Chae, "meth-SemiCancer: a cancer subtype classification framework via semi-supervised learning utilizing DNA methylation profiles," *BMC Bioinformatics*, vol. 24, no. 1, Apr. 2023, Art. no. 168, https://doi.org/10.1186/s12859-023-05272-6.

[7] M. Hamaneh and Y.-K. Yu, "A Simple Method for Robust and Accurate Intrinsic Subtyping of Breast Cancer," *Cancer Informatics*, vol. 22, Jan. 2023, Art. no. 11769351231159893, https://doi.org/10.1177/11769351231159893.

[8] L. Zhong, Q. Meng, and Y. Chen, "A Cascade Flexible Neural Forest Model for Cancer Subtypes Classification on Gene Expression Data," *Computational Intelligence and Neuroscience*, vol. 2021, no. 1, Oct. 2021, Art. no. 6480456, https://doi.org/10.1155/2021/6480456.

[9] S. Cascianelli, I. Molineris, C. Isella, M. Masseroli, and E. Medico, "Machine learning for RNA sequencing-based intrinsic subtyping of breast cancer," *Scientific Reports*, vol. 10, no. 1, Aug. 2020, Art. no. 14071, https://doi.org/10.1038/s41598-020-70832-2.

[10] Y. Huang, P. Zeng, and C. Zhong, "Classifying breast cancer subtypes on multi-omics data via sparse canonical correlation analysis and deep learning," *BMC Bioinformatics*, vol. 25, no. 1, Mar. 2024, Art. no. 132, https://doi.org/10.1186/s12859-024-05749-y.

[11] M. M. Islam, S. Huang, R. Ajwad, C. Chi, Y. Wang, and P. Hu, "An integrative deep learning framework for classifying molecular subtypes of breast cancer," *Computational and Structural Biotechnology Journal*, vol. 18, pp. 2185–2199, Aug. 2020, https://doi.org/10.1016/j.csbj.2020.08.005.

[12] S. Zhang *et al.*, "lncRNA Gene Signatures for Prediction of Breast Cancer Intrinsic Subtypes and Prognosis," *Genes*, vol. 9, no. 2, Feb. 2018, Art. no. 65, https://doi.org/10.3390/genes9020065.

[13] S. Meshoul, A. Batouche, H. Shaiba, and S. AlBinali, "Explainable Multi-Class Classification Based on Integrative Feature Selection for Breast Cancer Subtyping," *Mathematics*, vol. 10, no. 22, Nov. 2022, Art. no. 4271, https://doi.org/10.3390/math10224271.

[14] J. M. Choi and H. Chae, "moBRCA-net: a breast cancer subtype classification framework based on multi-omics attention neural networks," *BMC Bioinformatics*, vol. 24, no. 1, Apr. 2023, Art. no. 169, https://doi.org/10.1186/s12859-023-05273-5.

[15] T. Wang *et al.*, "MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification," *Nature Communications*, vol. 12, no. 1, Jun. 2021, Art. no. 3445, https://doi.org/10.1038/s41467-021-23774-w.

[16] J. Xu, P. Wu, Y. Chen, Q. Meng, H. Dawood, and H. Dawood, "A hierarchical integration deep flexible neural forest framework for cancer subtype classification by integrating multi-omics data," *BMC Bioinformatics*, vol. 20, no. 1, Oct. 2019, Art. no. 527, https://doi.org/10.1186/s12859-019-3116-7.

[17] Y. Lin, W. Zhang, H. Cao, G. Li, and W. Du, "Classifying Breast Cancer Subtypes Using Deep Neural Networks Based on Multi-Omics Data," *Genes*, vol. 11, no. 8, Aug. 2020, Art. no. 888, https://doi.org/10.3390/genes11080888.

[18] J. N. Weinstein *et al.*, "The Cancer Genome Atlas Pan-Cancer analysis project," *Nature Genetics*, vol. 45, no. 10, pp. 1113–1120, Oct. 2013, https://doi.org/10.1038/ng.2764.

[19] A. Colaprico *et al.*, "TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data," *Nucleic Acids Research*, vol. 44, no. 8, May 2016, Art. no. e71, https://doi.org/10.1093/nar/gkv1507.

[20] "R: The R Project for Statistical Computing." R-project. https://www.r-project.org/.

[21] T. M. Chowdhury, F. Tabassum, S. Islam, and A. R. M. Kamal, "A Pan-cancer Classification Model using Multi-view Feature Selection Method and Ensemble Classifier." arXiv, Jan. 12, 2025, https://doi.org/10.48550/arXiv.2501.06805.

[22] J. S. Cramer, "The Origins of Logistic Regression." Social Science Research Network, Dec. 01, 2002, https://doi.org/10.2139/ssrn.360300.

[23] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep. 1995, https://doi.org/10.1007/BF00994018.

[24] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 785–794, https://doi.org/10.1145/2939672.2939785.

[25] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, Montreal, Canada, 2018, pp. 6639–6649.

[26] P. Cunningham and S. J. Delany, "k-Nearest Neighbour Classifiers - A Tutorial," *ACM Computing Surveys*, vol. 54, no. 6, Jul. 2021, Art. no. 128, https://doi.org/10.1145/3459665.

[27] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, https://doi.org/10.1023/A:1010933404324.

[28] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, CA, USA, 2017, pp. 4768–4777.

[29] X. Dai *et al.*, "Breast cancer intrinsic subtype classification, clinical use and future trends," *American Journal of Cancer Research*, vol. 5, no. 10, pp. 2929–2943, Sep. 2015.

[30] S. X. Ge, D. Jung, and R. Yao, "ShinyGO: a graphical gene-set enrichment tool for animals and plants," *Bioinformatics*, vol. 36, no. 8, pp. 2628–2629, Apr. 2020, https://doi.org/10.1093/bioinformatics/btz931.

[31] M. Kanehisa and S. Goto, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, Jan. 2000, https://doi.org/10.1093/nar/28.1.27.

[32] C. VanGenderen, T. A. A. Harkness, and T. G. Arnason, "The role of Anaphase Promoting Complex activation, inhibition and substrates in cancer development and progression," *Aging*, vol. 12, no. 15, pp. 15818–15855, Aug. 2020, https://doi.org/10.18632/aging.103792.

[33] O. Wattanathamsan and V. Pongrakhananon, "Emerging role of microtubule-associated proteins on cancer metastasis," *Frontiers in Pharmacology*, vol. 13, Sep. 2022, Art. no. 935493, https://doi.org/10.3389/fphar.2022.935493.

[34] M. K. Shanmugam *et al.*, "Role of novel histone modifications in cancer," *Oncotarget*, vol. 9, no. 13, pp. 11414–11426, 2018, https://doi.org/10.18632/oncotarget.23356.

[35] D. Vugic *et al.*, "Replication gap suppression depends on the double-strand DNA binding activity of BRCA2," *Nature Communications*, vol. 14, no. 1, Jan. 2023, Art. no. 446, https://doi.org/10.1038/s41467-023-36149-0.

[36] A. L. Parker, M. Kavallaris, and J. A. McCarroll, "Microtubules and Their Role in Cellular Stress in Cancer," *Frontiers in Oncology*, vol. 4, Jun. 2014, Art. no. 153, https://doi.org/10.3389/fonc.2014.00153.

[37] R. Roskoski, "Cyclin-dependent protein serine/threonine kinase inhibitors as anticancer drugs," *Pharmacological Research*, vol. 139, pp. 471–488, Jan. 2019, https://doi.org/10.1016/j.phrs.2018.11.035.

[38] S. Rodrigues-Ferreira, A. Molina, and C. Nahmias, "Microtubule-associated tumor suppressors as prognostic biomarkers in breast cancer," *Breast Cancer Research and Treatment*, vol. 179, no. 2, pp. 267–273, Jan. 2020, https://doi.org/10.1007/s10549-019-05463-x.

[39] M. S. Ong *et al.*, "Cytoskeletal Proteins in Cancer and Intracellular Stress: A Therapeutic Perspective," *Cancers*, vol. 12, no. 1, Jan. 2020, Art. no. 238, https://doi.org/10.3390/cancers12010238.

[40] B. Nami and Z. Wang, "Genetics and Expression Profile of the Tubulin Gene Superfamily in Breast Cancer Subtypes and Its Relation to Taxane Resistance," *Cancers*, vol. 10, no. 8, Aug. 2018, Art. no. 274, https://doi.org/10.3390/cancers10080274.

[41] K. A. L. Collins *et al.*, "Proteomic analysis defines kinase taxonomies specific for subtypes of breast cancer," *Oncotarget*, vol. 9, no. 21, pp. 15480–15497, Mar. 2018, https://doi.org/10.18632/oncotarget.24337.

[42] T. M. Chowdhury, "Replication Data for: An Efficient and Interpretable Machine Learning Model for Predicting Breast Cancer Subtypes using Gene Expression Profiles." Harvard Dataverse, Apr. 26, 2025, https://doi.org/10.7910/DVN/WLEHDV.