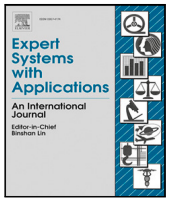
Expert Systems With Applications 225 (2023) 120130

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



# XAI-MethylMarker: Explainable AI approach for biomarker discovery for breast cancer subtype classification using methylation data

Sheetal Rajpal [a], Ankit Rajpal [a,*], Arpita Saggar [a], Ashok K. Vaid [b], Virendra Kumar [c], Manoj Agarwal [d], Naveen Kumar [a]

[a] Department of Computer Science, University of Delhi, Delhi, India
[b] Oncologist at Cancer Institute, Medanta, Gurugram, India
[c] Department of Nuclear Magnetic Resonance Imaging, All India Institute of Medical Sciences, New Delhi, India
[d] Department of Computer Science, Hans Raj College, University of Delhi, Delhi, India

ARTICLE INFO

*Keywords:*
Explainable AI
Methylation
Biomarker
Deep learning
Breast cancer subtypes

ABSTRACT

Breast cancer—a heterogeneous disease marked with a high mortality rate, necessitates early diagnosis and treatment. The availability of multi-omic data has revolutionized our understanding of how molecular changes mark the variations in different breast cancer subtypes. Epigenomic changes in the form of DNA methylation differentially impact the expression level of the genes that play a vital role in the onset and spread of these subtypes. So, in this paper, we study the role of these variations in distinguishing between the various breast cancer subtypes. The cardinality of the existing biomarker sets is often too large to be interpreted clinically, and their relevance in classification remains unclear. In this paper, we propose a two-stage XAI-MethylMarker—an explainable AI-based biomarker discovery framework applied to DNA methylation data to arrive at a small set of biomarkers for breast cancer classification. In the first stage, we build a deep-learning network $MethylNet$ that employs an autoencoder for dimensionality reduction and a feed-forward neural network to classify breast cancer subtypes. In the second stage, we propose a biomarker discovery algorithm, $MethylBDA$, which employs different explainable techniques for analyzing $MethylNet$ model and discovers a small set of 52 biomarkers. Using 5-fold cross-validation, we achieved a classification accuracy of $0.8145 \pm 0.07$ at a 95% confidence interval. To establish the clinical relevance of the discovered biomarkers, we performed a gene set analysis that revealed 14 druggable genes, nine genes linked to prognostic outcomes, and several enriched pathways are known to be significantly associated with distinct breast cancer subtypes.

## 1. Introduction

Breast cancer is the primary cause of mortality among women, with the highest incidence rate observed in 2020 (Sung et al., 2021). Thus, a timely and correct diagnosis of the disease is crucial. Intrinsic heterogeneity of breast cancer results in its categorization into various clinically and prognostically important subtypes. As patients with the same subtype are likely to respond to treatment therapies in a similar manner, identification of the correct subtype and discovery of subtype-specific biomarkers that may be used for prognostic evaluation and as therapeutic targets is important. Several classifications in terms of invasiveness, histological grading, TNM staging, and molecular subtyping has been established (Taherian-Fard, Srihari, & Ragan, 2015). The development of breast cancer is a multi-step process caused by the aggregation of molecular changes at genomic, epigenomic, and transcriptomic levels, which varies from individual to individual. Thus, molecular classification into different breast cancer subtypes, namely, Basal, Her2, Luminal A, Luminal B, and Normal-like best captures this heterogeneity. The availability of multi-omic data such as genomic, transcriptomic, and epigenomic data have revolutionized how these molecular changes mark the variations across subtypes. However, the high dimensionality of data, as well as the availability of only small-size imbalanced datasets, present a significant challenge.

Considering different omic data, aberrant epigenomic alterations in the form of DNA methylation adversely impact the expression level of the genes that play a vital role in the onset and spread of cancer (Holm et al., 2010; Jaenisch & Bird, 2003). The environmental factors such as exercises we perform, the diet we follow, and the life pressures we undergo, have a significant influence on DNA resulting in epigenomic

* Corresponding author.
*E-mail addresses:* sheetal.rajpal.09@gmail.com (S. Rajpal), arajpal@cs.du.ac.in (A. Rajpal), arpitasaggar.mca19.du@gmail.com (A. Saggar), akvaid@yahoo.com (A.K. Vaid), virendrakumar@aiims.edu (V. Kumar), manoj.agarwal@hrc.du.ac.in (M. Agarwal), nkumar@cs.du.ac.in (N. Kumar).

https://doi.org/10.1016/j.eswa.2023.120130
Received 13 September 2022; Received in revised form 9 March 2023; Accepted 9 April 2023
Available online 18 April 2023

changes such as methylation of chromosome segments. DNA methylation happens due to linking the methyl (CH3) group to cytosines in CpG dinucleotides resulting in a closed chromatin structure. These methylation changes may vary across CpG islands (CpGIs) and are measured using beta values denoting the extent of methylation. Higher beta values signify a higher level of DNA methylation (also called hypermethylation), and lower beta values signify a lower level of DNA methylation (also called hypomethylation). While hypomethylation is reported in the literature as linked with chromosome instability and affects oncogenes, hypermethylation is correlated with gene silencing affecting tumor suppressor genes (Daura-Oller, Cabre, Montero, Paternain, & Romeu, 2009). Thus, in this paper, we have focused on DNA methylation for dissecting molecular variations in breast cancer subtypes in this work.

As the hypermethylation and hypomethylation are known to be linked with progression of cancer, both supervised and unsupervised machine learning approaches (Liu, Chen, & Wong, 2021; Peng et al., 2019; Withnell, Zhang, Sun, & Guo, 2021) have been extensively employed to analyze this and other forms of data. While several researchers have leveraged methylation data for exploring new subtypes associated with improved prognosis and clinical outcomes using unsupervised techniques (Amor, Colomer, Monteagudo, & Naranjo, 2022; Stefansson et al., 2015; Wu, Tang, & Zhou, 2021), a large body of literature has studied it for breast cancer classification in supervised setting (Chen et al., 2019; Holm et al., 2010; Kuang et al., 2020; List et al., 2014; Liu, Peng, & Wang, 2020; Tao et al., 2019). Withnell et al. (2021) introduced XOmiVAE, a variational autoencoder (VAE)-based interpretable deep learning model for cancer classification utilizing high-dimensional omics data. They employed 33 TCGA tumor gene expression profiles. XOmiVAE can depict the contribution of each gene and latent dimension to each classification prediction and their association. XOmiVAE can also explain outcomes of deep learning network for unsupervised clustering as well as supervised classification. Tao et al. (2019) used normalized methylation data from the TCGA repository and applied multiple kernels with SVM to categorize breast cancer subtypes based on immunohistochemistry (IHC) markers, namely, ER, PR, and HER2 levels (Vallejos et al., 2010). They applied the Wilcoxon rank-sum test with FDR correction to select the genes with p-values less than 0.05 and obtained a classification accuracy of 0.65 using the selected biomarkers. However, compared to IHC-defined subtypes, PAM50 subtypes introduced by Parker et al. (2009) are more closely linked with prognostic and clinical outcomes. As the PAM50 defined subtypes are considered the gold standard (Eccles et al., 2013; Kim et al., 2019), they have been exploited by several researchers for breast cancer classification (Chen et al., 2019; List et al., 2014; Zhang, Wang, et al., 2018). List et al. (2014) have studied methylation data for PAM50-based classification of breast cancer using the random forest-based model. Using the TCGA dataset, they applied the Gini index measure to select a set of 38 DNA methylation genes and reported a classification accuracy of 0.753. Similar work is also reported by Chen et al. (2019), who investigated the associations between the methylation patterns of genes for four breast cancer subtypes using the datasets provided by Gene Expression Omnibus (GEO). To handle the class imbalance problem, they applied SMOTE filter. They shortlisted 9777 genes having a high MR (Maximum Relevance) score. Subsequently, using incremental feature selection with SVM, they arrived at a small set of 40 genes, resulting in an overall accuracy of 73.70% for classifying the four breast cancer subtypes.

Above-mentioned methodologies proposed in literature (Chen et al., 2019; List et al., 2014; Tao et al., 2019) have been successful in discovering DNA methylation biomarkers for different breast cancer subtypes to varying degrees, however, the relative relevance of the specific biomarkers remains unknown to the end-user. As these biomarkers are to be used in life-critical diagnosis, it is important to understand their role in breast cancer classification. In this paper, we have focused on using DNA methylation data to discover a small set of biomarkers that would enable us to dissect the heterogeneity of breast cancer. Towards this end, we have proposed a two-stage explainable AI-based biomarker discovery framework that we call XAI-MethylMarker (see Fig. 1). Motivated by the success of deep learning techniques in various application areas, including classification of cancer subtypes based on the selected biomarkers (Cristovao et al., 2020; Fakoor, Ladhak, Nazi, & Huber, 2013; Hosni, Abnane, Idri, de Gea, & Alemán, 2019; Karabulut & Ibrikci, 2017; Lin, Zhang, Cao, Li, & Du, 2020; Spanhol, Oliveira, Petitjean, & Heutte, 2016; Xiao, Wu, Lin, & Zhao, 2018; Xu et al., 2017), the proposed framework exploits the power of deep learning. In the first stage, we build a deep learning-based classifier network $MethylNet$, which comprises an autoencoder for dimensionality reduction, followed by constructing a neural network-based model for classification into five subtypes. In the second stage, the aforementioned neural network classifier model $MethylNet$ is analyzed to discover the relevant genes. For this purpose, we have developed a Biomarker Discovery Algorithm ($MethylBDA$) that employs several explainable AI techniques to unveil the entire working of the proposed MethylNet network to the end-user, thus enhancing their trust in the proposed methodology. The algorithm discovers a set of 52 differential DNA methylation biomarkers for breast cancer classification. Finally, we use the SHapley Additive exPlanations (SHAP) method to mark the contribution of individual biomarkers in breast cancer classification. Though several of these biomarkers are known to be linked with breast cancer, a few others identified may provide new insight to clinicians for further study.

In summary, although the existing literature on biomarker discovery extensively exploits various machine learning and statistical approaches for biomarker discovery (Liu, Xie, & Udell, 2021), to the best of our knowledge, the proposed XAI-driven deep learning framework for biomarker discovery is a novel attempt at biomarker discovery, that would not only aid breast cancer subtyping but may also be exploited by medical practitioners in devising therapeutic interventions. The contributions of this paper are enumerated below:

1. A generic explainable AI-based framework, $XAI - Methyl\ Marker$, for discovering biomarkers for breast cancer classification.
2. Discovery of a concise set of 52 DNA methylation biomarkers. Using 5-fold cross-validation, discovered biomarkers enabled us to achieve a classification accuracy of 0.8145 (±0.07) at a 95% confidence interval, which is comparable to state-of-the-art approaches.
3. The biomarkers discovered by XAI-MethylMarker outperformed the biomarkers obtained via other competitive feature selection methods in terms of classification performance.
4. Contribution of the identified biomarkers in breast cancer classification marked through SHapley Additive exPlanations (SHAP) method.
5. Results of gene set analysis:
   (a) Several enriched Reactome pathways (with FDR corrected p-values less than 0.05), all of which are known to be significantly associated with distinct breast cancer subtypes.
   (b) Presence of fourteen druggable genes and nine genes associated with prognosis in the discovered biomarkers.

This paper is structured as follows: in the second section, we present the datasets and propose a framework for biomarker discovery; in the third section, we give experimental details, outcomes, and interpretation of the finding and finally, give the conclusions in the fourth section.

## 2. Materials and methods

This section describes the dataset used for the experimentation. Next, we present a two-stage XAI-MethylMarker—an explainable AI-based biomarker discovery framework that is applied to DNA methylation data to arrive at a small set of biomarkers for breast cancer

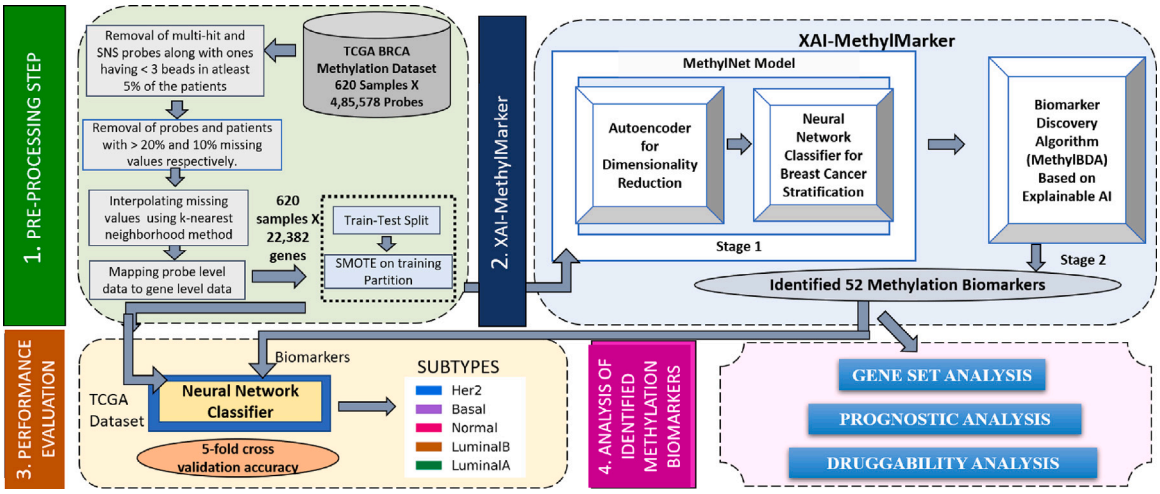



**Fig. 1.** Workflow of the proposed approach comprises four main components: (1) Pre-processing to clean and map probe-level data to gene-level data followed by normalization, train-test partitioning, and handling class imbalance problem using SMOTE filter (2) XAI-MethylMarker, two-stage explainable AI-based biomarker discovery framework to discover biomarkers, (3) Performance Evaluation of discovered biomarkers, and (4) Gene set analysis, prognostic analysis, and druggability analysis to establish the clinical relevance of the discovered biomarkers.

classification. To access the code of the proposed framework, please clickhere.[1]

### 2.1. Dataset

The dataset used in this paper was compiled as part of The Cancer Genome Atlas (TCGA) (UCSC, 2016) project contributed by seven National Cancer Institute (NCI) funded genome characterization centers. It includes multi-omic data for a variety of cancer types. Specifically, we use CpG level methylation 450k data having probe values in the range 0 and 1 indicating the level of methylation (beta value) for breast cancer patients. Higher beta values indicate higher levels of DNA methylation (hypermethylation), and lower beta values signify lower levels of DNA methylation (hypomethylation). The beta values in the methylation dataset follow a bimodal distribution. Though the DNA methylation data is available at 27k and 450k platforms, we have chosen the methylation450k platform for further analysis because the bimodal distribution in the methylation450k platform is significantly more apparent and evenly distributed than in the methylation27k platform. The dataset shows two peaks near 0.1 and 0.9 values and a gently sloped valley near 0.2–0.8. The data set has been retrieved from the University of California's Xena repository, which holds Methylation450K data for 4,85,578 probes for 888 patients. However, the current study considers 620 patients for whom the PAM50 class categories (Basal: 87, Her2: 31, LumA: 288, LumB: 127, and Normal-like: 87) are available. The demographic and clinical summary of the resultant dataset is given in Fig. 2.

*Mapping probe level data to gene-level data*

To arrive at the average gene-level methylation value, we first removed outliers by eliminating multi-hit and single-nucleotide polymorphism (SNP) related probes (CpG islands) and the probes comprising less than three beads in at least 5% of the patients. To deal with missing values in the dataset, probes with more than 20% NA's and patients with more than 10% NA's were removed. Subsequently, the remaining missing values are computed using the K-Nearest Neighborhood method ($K = 5$). The ChAMP library of R (Morris et al., 2014) is used to perform the above-mentioned operations. We also employed the boxplot method to find outliers, but only a few samples (10%–20%) showed sudden changes in DNA methylation beta-values. We kept these values for the final study since they corresponded to hypermethylation and hypomethylation in samples. The obtained data is finally mapped to gene-level corresponding to 22,382 genes using the function $CalculateSingleValueMethylationData$ of TCGA-Assembler2 Bioconductor package (Wei et al., 2018). These average gene-level beta values are computed for promoter regions TSS1500 and TSS2000.

### 2.2. $XAI - MethylMarker$ Framework

In this paper, we have carried out a retrospective study based on the TCGA breast cancer dataset (described in Section 2.1) and propose a two-stage explainable AI-based biomarker discovery framework that we call $XAI - MethylMarker$.

#### 2.2.1. Stage 1- $MethylNet$: Methylation based deep learning model for breast cancer classification

In this work, we have proposed a deep learning-based classifier model $MethylNet$ for breast cancer classification in the first stage. It is divided into two subnetworks (See Fig. 3). Since the huge dimensionality of the omic data is difficult to handle by any classifier, we compress the methylation data for 22,382 genes using an autoencoder in the first subnetwork. An autoencoder consists of an encoder and a decoder. The encoder network reduces a large number of input features to a compressed representation that is passed to the decoder network, which attempts to replicate input data. The weights of the autoencoder network are improved using the available training data by reducing information loss while the network's inputs are compressed and decompressed. The encoder consists of three layers with 5000, 2000, and 500 nodes, whereas the decoder consists of three levels with 2000, 5000, and 22,382 nodes (See Fig. 3). As a result, the autoencoder network transforms 22,382 methylation genes to a feature vector of length 500. The hidden layers utilize the ReLU activation function, whereas the output layer uses the linear activation function.

The second subnetwork of classifier model $MethylNet$ deploys a feed-forward neural network to perform multi-class breast cancer classification. The network comprises a hidden layer having 200 nodes, followed by an output layer comprising five nodes corresponding to the five breast cancer subtypes. The output of the encoder (compressed vector of size 500) serves as input to this network. The ReLU and softmax activations are used in the hidden and output layers, respectively, where the softmax activation function employed at each of the five neurons of the output layer yields the probability of belonging to each breast cancer subtype. Batch normalization is employed to cope with the internal covariate shift problem. To avoid overfitting, a dropout of 0.2 and 0.3 is used following the input and hidden layers, respectively.

[1] Project Page: https://github.com/SheetalRajpal/methyl-marker

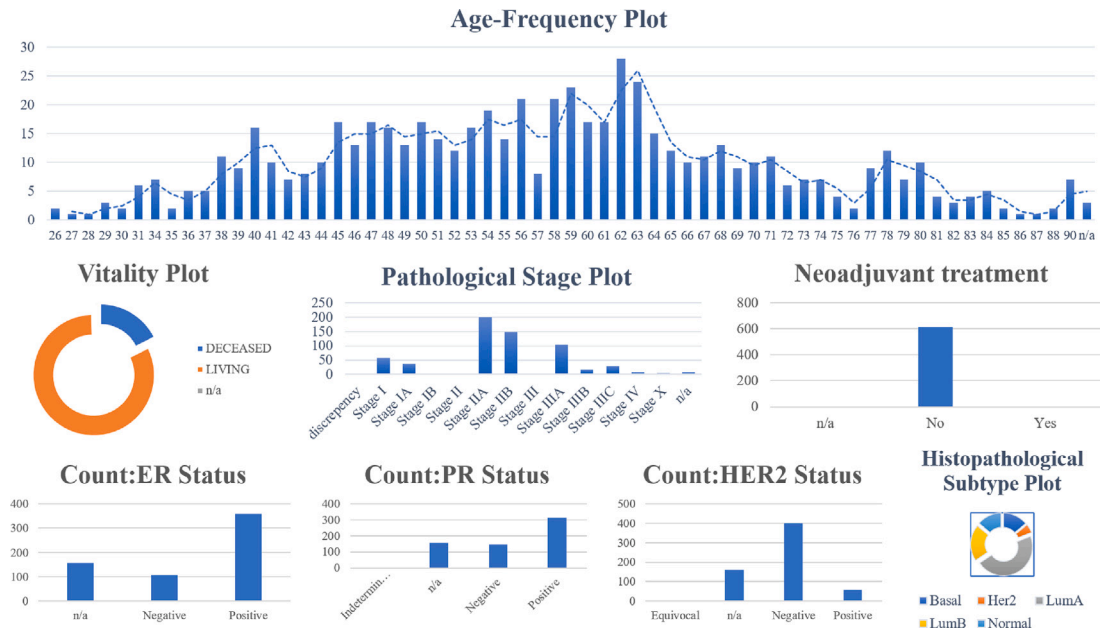



**Fig. 2.** Demographic and clinical details of TCGA BRCA for 620 patients, namely, plots corresponding to Age-Frequency, Vitality, Pathological Stage, Histopathological Subtype, ER/PR/Her2 Status, and Neoadjuvant Treatment.

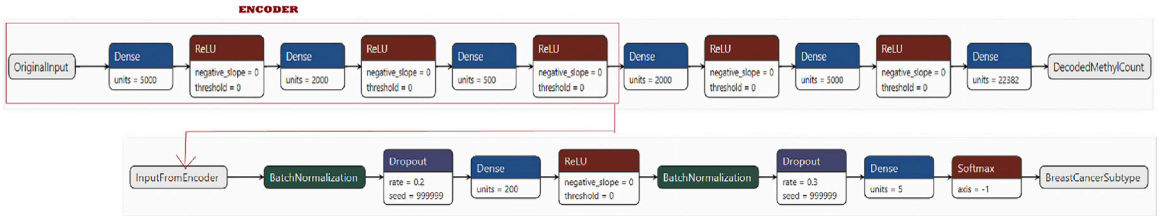



**Fig. 3.** Classifier Model *MethylNet*—the deep learning-based classifier network utilizing methylation dataset. It comprises an autoencoder for dimensionality reduction (22,382 genes to feature vector of size 500), followed by a feed-forward neural network classifier that accepts the output of the encoder as an input and performs classification into one of the five breast cancer subtypes.

#### 2.2.2. Stage 2- biomarker discovery algorithm: an explainable AI approach

In the second stage, we propose *BiomarkerDiscoveryAlgorithm* (*MethylBDA*) to identify DNA methylation biomarkers (see Algorithm 1) by interpreting the deep-neural network-based classifier model—*MethylNet*. To discover subtype-specific DNA methylation genes, *MethylBDA* leverages the following explainable AI methods: Gradient*input, DeepLIFT, Epsilon Layerwise Relevance Propagation, Integrated Gradient, Gradient SHAP (SHapley Additive exPlanations), and Deep SHAP (Lundberg & Lee, 2017).

The *MethylBDA* algorithm (see Algorithm 1 and its visual representation in Fig. 4) comprises two functions, namely, *discoverBiomarkers* and *subtypeCandidates*. Function *discoverBiomarkers* builds p (=10) distinct *MethylNet* models corresponding to $p$ distinct seeds. It then calls function *subtypeCandidates* that takes the classifier model of the first stage (*MethylNet*) along with the preprocessed dataset $X$ as an input and returns the subtype-specific candidate biomarker genes to function *discoverBiomarkers*. Finally, function *discoverBiomarkers* selects those subtype-specific candidate biomarker genes identified as the most relevant biomarkers across at least five seeds.

The working of *MethylBDA* algorithm initiates by calling the function *discoverBiomarkers*. First, an empty set of *SG* (Shortlisted Genes) is created. Next, it builds p (=10) *MethylNet* models (one for each seed) to classify breast cancer subtypes based on DNA methylation data. Each model is examined through the function *subtypeCandidates*, and the genes contributing to the model's prediction are shortlisted (*SG*). For patients of each subtype, the function *subtypeCandidates* chooses the most distinguishing genes (*geneSet*) using six explainable AI techniques. The function first defines an empty vector of genes *geneSet*, used to store genes associated with a subtype. *geneSet* is progressively updated by adding genes relevant to each subtype for all the explainable AI methods. The *XAI* method calculates the *scores* (relevance) of different genes by analyzing the model *MethylNet* using the whole dataset $X$. Then, given a *subtype*, we use the *elbow* method to choose the *topGenes* based on their contribution scores. Finally, for each subtype, we pick those sets of genes (*geneSet*) from *topGenes* which are relevant for subtype classification of at least one-fourth fraction of the patients (cutoff of 25%). Thereafter, among the subtype-specific genes (*geneSet*), those genes are selected as candidate biomarkers that are identified as important by at least five explainable AI techniques out of six. Following the interpretation of each trained model initialized with various seeds, these subtype-specific DNA methylation candidate biomarker genes are returned to function *discoverBiomarkers*. Finally, the *discoverBiomarkers* function chooses only those genes as biomarkers for each subtype that have been recognized as contributing the most to the predicted subtype for at least 50% of seeds.

## 3. Experimental results and discussion

In this section, we begin with the description of experimental details. After that, we describe the data preprocessing and hyperparameter tuning involved. Following that, we provide the DNA methylation biomarkers discovered using the *MethylBDA* algorithm and their efficacy in breast cancer subtype classification. We compare our work with List et al. (2014). Furthermore, we assess the reliability of the

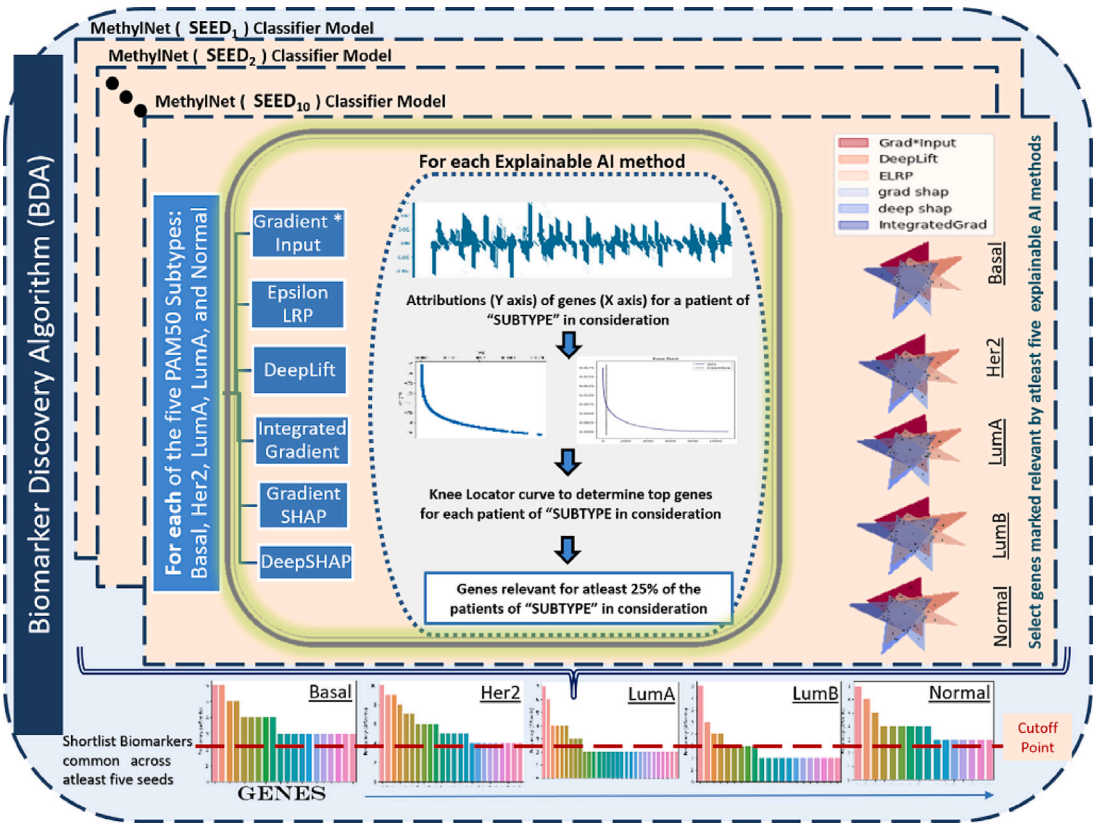



**Fig. 4.** Workflow of Biomarker Discovery Algorithm (*MethylBDA*). We first build 10 *MethylNet* models (one for each seed) to classify breast cancer subtypes based on DNA methylation data. For each of the five subtypes, each model is examined through the application of different explainable AI methods, and the most contributing genes to the model's prediction are shortlisted using the elbow technique. Subsequently, for each subtype, we pick those sets of genes that are relevant for subtype classification of at least one-fourth fraction of the patients (cutoff of 25%). Thereafter, among the subtype-specific genes, those genes are selected as candidate biomarkers that are identified as important by at least five explainable AI techniques out of six. Finally, those subtype-specific candidate biomarker genes were identified as the most relevant biomarkers across at least five seeds.

identified biomarkers in terms of biological significance. We provide the findings of enriched Reactome pathways, potentially druggable genes, and the genes linked with survival from the discovered DNA methylation biomarkers.

### 3.1. Experimental details

The proposed framework has been implemented in Python 3.7 in the Google Colaboratory environment (NVIDIA Tesla K80 GPU with 12 GB RAM). Python modules Pandas, Numpy, Imblearn, Keras, Matplotlib, Scikit-Learn, and Seaborn, have been used for the data pre-processing phase, model creation, and visual analysis. For the interpretation of deep neural network *MethylNet*, methods of the DeepExplain toolbox (https://github.com/marcoancona/DeepExplain) and SHAP libraries have been utilized.

#### 3.1.1. Data pre-processing

The present work uses gene-specific values for 620 patients derived by mapping probe-level data to gene-level data. The methylation levels computed for 22,382 genes are in the range [0,1], obviating the requirement for normalization. To facilitate classification by the neural network, each subtype is mapped to one hot encoding vector. To assess the classifier's effectiveness, 5-fold cross-validation is performed for each MethylNet model (built for each seed). In other words, for each MethylNet model, the classification experiment is repeated five times, with one fold retained as test data and the data from the remaining four folds retained as training data:

1. 80% of the data set aside for training is used to train the first sub-network of the MethylNet mode (autoencoder). This network takes 22,382 genes as its input and produces a compact representation of size 500 as its output.
2. The same 80% of the data utilized for training of the first sub-network of the MethylNet model (classifier) is employed to train the second sub-network of the MethylNet model
3. For the remaining 20% of the data set aside for testing, each instance is passed to the first sub-network (trained in step 1 before) to generate the encoded representation of size 500.
4. The encoded representation obtained in the previous step is passed to the second sub-network (trained in step 2 above) to be classified as one of the five subtypes of breast cancer.

Also, for each fold, 10% of the training set is reserved for validation to guide model building. To address the problem of class imbalance, we have used the Synthetic Minority Oversampling Technique (SMOTE) filter (Chawla, Bowyer, Hall, & Kegelmeyer, 2002) to ensure an equal distribution of data across all classes. SMOTE is a data augmentation strategy for creating balanced datasets that employ instances' nearest neighbors to generate synthetic, fairly similar samples.

#### 3.1.2. Hyperparameters

The hidden layers of the *MethylNet* model make use of L2 regularization ($\lambda = 10e-15$) to prevent overfitting and employs the ReLU activation function to address the vanishing gradient problem. The first sub-network of the *MethylNet* model, which consists of an autoencoder, employs a mean squared loss function and an Adam optimizer with a learning rate of 0.00001. We have used a batch size of 32 for model training. Early termination condition (reduction in validation loss ($\delta = 0.001$), patience level=50) has been employed to prevent the network from overfitting. The above hyperparameters

**Algorithm 1 Biomarker Discovery Algorithm ($MethylBDA$)**

**Input:**
MethylNet: Breast Cancer Classification Model
**X**: TCGA breast cancer methylation dataset of size $N \times M$, where $N$ indicates patient count and $M$ indicates gene count.
**XAI_MethodList**: List of XAI Methods:Gradient*Input, EpsilonLRP, DeepLIFT, IntegratedGradient, GradientSHAP, DeepSHAP

**Functions:**
**XAI**($MethylNet$, $X$): For each patient x of dataset X, XAI method yields the contribution of each gene in classifying the patient with a subtype using the network $MethylNet$
**elbow**($scoreMatrix$): For each patient in the given set, the method yields genes with maximum contribution determined using elbow method
**select**($list$, $cutoff$): Yields those genes from the given list, having $cutoff$ as its minimum occurrence frequency
**freqXAI**($list$, $cutoff$): Yields those frequent genes (obtained using different XAI methods) from the given list, having $cutoff$ as its minimum occurrence frequency
**freqSeeds**($list$, $cutoff$): Yields those frequent genes (across different seeds) from the given list, having $cutoff$ as its minimum occurrence frequency

**Notations:**
**SG**: Shortlisted Genes
**CG**: Candidate Genes

**Output:**Biomarkers

```
Procedure subtypeCandidates(MethylNeT, X)
    for subtype ∈ Basal, Her2, LumA, LumB, Normal do
        geneSet[subtype] ← {}
    end
    for XAI ∈ XAI_MethodList do
        scores ← XAI(MethylNet, X)
        for subtype ∈ Basal, Her2, LumA, LumB, Normal do
            topGenes ← elbow(scores[X[subtype]])
            cutoff ← len(X[subtype])/4
            genes[XAI, subtype] ← select(topGenes, cutoff)
            geneSet[subtype] ←
            geneSet[subtype] ∪ genes[XAI, subtype]
        end
    end
    for subtype ∈ Basal, Her2, LumA, LumB, Normal do
        CG[subtype] ← freqXAI(geneSet[subtype], cutoff)
    end
    return {CG[Basal], CG[Her2], CG[LumA], CG[LumB], CG[Normal]}
Procedure discoverBiomarkers()
    SG ← {}
    for seed ∈ p_Seeds do
        model ← MethylNet(seed, X)
        SG[Allsubtypes] ← SG[Allsubtypes] ∪ subtypeCandidates(model, X)
    end
    for subtype ∈ Basal, Her2, LumA, LumB, Normal do
        Biomarkers[subtype] ← freqSeeds(SG[subtype], cutoff)
    end
    return Biomarkers
```

were optimized using cross-validated grid-search over the following parameter grid values used in model building for autoencoder:

1. Regularization: L1, L2
2. Activation Functions: tanh, ReLU, Leaky ReLU
3. Learning Rate: 0.01, 0.001, 0.0001, 0.00001, 0.000001
4. Batch-size: 32, 64, 96, 128
5. Early stopping criterion
   (a) $\delta$=0.01,0.001,0.0001
   (b) patience level: 50,100,150

We have also experimented with alternative designs of the autoencoder, but the resulting models did not perform well. The second sub-network comprising the neural network to classify breast cancer patients into five subtypes employs an Adam optimizer learning algorithm ($learningrate = 0.0002$) and a categorical cross-entropy loss function. To avoid overfitting, a dropout of 0.2 and 0.3 is used following the input and hidden layers, respectively. Finally, the network has been trained using a batch size of eight for 1000 epochs.

#### 3.1.3. $MethylNet$ Results based on 22,382 input genes

Two sub-networks of $MethylNet$ model are trained separately. The encoded representation of methylation level data obtained from the autoencoder is passed to the second sub-network that classifies breast cancer patients. The trained model is used for evaluating the performance on the test set using 5-fold cross-validation. We determined the performance of the autoencoder under different hyperparameter settings (mentioned in Table 1). On experimenting with different values, we obtained the best set of hyperparameters for the autoencoder (mentioned in Section 3.1.2), which achieved a mean squared error (loss) of 0.0035. Using the entire set of 22,382 genes, we obtained a classification accuracy of 0.784 ± 0.042 at a 95% confidence interval using $MethylNetmodel$.

### 3.2. Using explainable AI for selection of biomarkers and their classification performance

For discovering the most differentiating DNA methylation subtype-specific genes, we used the proposed Biomarker Discovery Algorithm ($MethylBDA$) described in Section 2.2.2. Using different explainable AI methods, namely, Gradient*Input, DeepLIFT, $\epsilon$-Layerwise Relevance

**Table 1**
Autoencoder loss (Mean Squared Error) under different hyperparameter settings.

| Hyperparameters | | Loss |
|---|---|---|
| ***Best Hyperparameters:*** (***Regularization: L2, Activation: ReLU, Learning Rate:0.00003, Batch Size: 32, Early Stopping Delta: 0.001***) | | ***0.0035*** |
| RegularizationL (Lambda: 10e−15) | L1 | 0.0052 |
| | L1L2 | 0.3756 |
| Activation | tanh | 0.0052 |
| | LeakyReLU | 0.0050 |
| Learning Rate | 0.01 | 0.0088 |
| | 0.001 | 0.0062 |
| | 0.0001 | 0.0055 |
| | 0.00001 | 0.0056 |
| | 0.000001 | 0.0084 |
| Batch Size | 64 | 0.0061 |
| | 96 | 0.0064 |
| | 128 | 0.0065 |
| Early Stopping Delta | 0.01 | 0.0060 |
| | 0.0001 | 0.0044 |

**Table 2**
Accuracy of Neural Network classifier for 52 discovered biomarkers under different hyperparameter settings.

| Hyperparameters | | 5-fold Accuracy |
|---|---|---|
| ***Best Hyperparameters:*** (***Regularization: None, Activation: ReLU, Learning Rate:0.002, Batch Size: 16, Dropout: 0.3***) | | ***0.8145*** |
| Regularization(Lambda: 10e−15) | L1 | 0.7339 |
| | L2 | 0.7306 |
| | L1L2 | 0.7452 |
| Activation | tanh | 0.7822 |
| | LeakyReLU | 0.8048 |
| Learning Rate | 0.01 | 0.8103 |
| | 0.001 | 0.8107 |
| | 0.0001 | 0.7936 |
| | 0.00001 | 0.6274 |
| | 0.000001 | 0.5838 |
| Batch Size | 32 | 0.8090 |
| | 64 | 0.7968 |
| | 128 | 0.7903 |
| Dropout Rate | 0.1 | 0.7435 |
| | 0.2 | 0.7832 |
| | 0.4 | 0.7393 |
| | 0.5 | 0.7032 |

Propagation, Integrated Gradient, Gradient SHAP, and Deep SHAP, we selected a set of potential biomarker genes identified as the most relevant (for at least one-fourth of the patients of each subtype) by at least five methods. Finally, we discovered 52 biomarkers by shortlisting the genes designated as relevant in at least 50% of the seeds. These biomarkers include AC073508.1, AGR3, AOX1, ARHGAP40, ARHGDIB, CARD6, CFH, CHAD, CTD-2298J14.2, CTD-2370N5.3, CYP2F1, DEF6, DGKB, EMBP1, FBXO47, FJX1, FLACC1, FMOD, GBP4, IFI27, KLK3, KRTAP19-1, MESTIT11, MIR592, MNDA, MT1DP, OR10G9, OR10S1, OR2T4, OR51B2, OR51B6, OR5B17, OR8D4, OR8J3, PPP2R3 A, PRELP, PTPRQ, RHBDL1, RP11-12M5.1, RP11-159D12.2, RP11-344P13.6, RP4-761J14.9, RPL13AP, SERPINA3, SLC2A5, SNORD32B, SYNGR2, SYNPR-AS1, TUSC7, TXLNB, VTRNA1-2, and ZNF671.

To evaluate the efficacy of the 52 biomarkers discovered using the XAI-MethylMarker framework, we trained another neural network classifier comprising three hidden layers with 40, 10, and 5 neurons, respectively. To avoid overfitting, a dropout of 0.3 is used. For classification, hidden layers employ the ReLU activation function, and the output layer employs the softmax function. Further, we used a batch size of 16 for model training with the Adam optimizer (learning rate 0.002) and categorical cross-entropy loss function. The above hyperparameter values were obtained using cross-validated grid-search over the following parameter grid values used in model building (see Table 2):

1. Regularization: L1, L2
2. Activation Functions: tanh, ReLU, Leaky ReLU
3. Learning Rate: 0.01, 0.001, 0.0001, 0.00001, 0.000001
4. Batch-size: 32, 64, 96
5. Dropout: 0.1,0.2,0.3,0.4,0.5

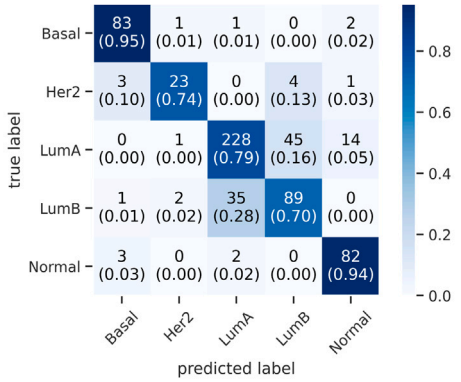
Using the five-fold cross-validation, we achieved a classification accuracy of 0.8145 ± 0.07 at a 95% confidence interval for the discovered 52 biomarkers. It is to be noted that, the classification accuracy (0.8145 ± 0.07 at a 95% confidence interval) using 52 biomarkers is found to be higher as compared to the classification accuracy (0.784 ± 0.042 at a 95% confidence interval) using the entire set of 22,382 genes. A probable cause may be the High-dimensional, low-sample-size (HDLSS) nature of the dataset under study. HDLSS data might induce severe overfitting and excessive training gradient variation while estimating parameters of deep neural networks (DNN), which could degrade performance. HDLSS data poses two key obstacles to using DNN for cancer prediction. High-dimensional features need DNNs to estimate a large number of parameters, but a limited number of samples is not enough to allow adequate training to identify patterns from redundant features, which might lead to over-fitting. Conversely, training with little sample size data increases training gradient variance, which leads to erroneous DNN gradient estimates.

The classification performance of 0.8145 ± 0.07 at a 95% confidence interval achieved using discovered 52 biomarkers is depicted in Fig. 5 through confusion matrix, heatmap of performance metrics and boxplot. The diagonal entries in the confusion matrix (see Fig. 5(a)) indicate the number of patients of each subtype that have been correctly classified. It is evident from the confusion matrix that using discovered 52 biomarkers, the model is able to predict most of the Basal (95%) and Normal-like patients (94%) correctly (Fig. 5(a)). However, the model achieves moderate classification accuracy above 70% for the remaining three subtypes. The worst results are observed for the Luminal B subtype (70% accuracy), where several of the luminal B patients are classified as luminal A and vice-versa. This may be attributed to the overlapping characteristics of luminal A and luminal B subtypes. The classification performance measures precision, recall, and F1-score for five-fold cross-validation are depicted using heatmap in Fig. 5(b). The heatmap shows that the model performs best for the Basal subtype, as evident from high precision, recall, and F1-score scores ($\geq 0.92$). However, for the Luminal B subtype, precision, recall, and F1-score metrics are 0.64, 0.70, and 0.67, respectively. The variations in classification performance across five-folds have been captured using the boxplots (see Fig. 5(c)). Note that the results are most stable for the Basal and Luminal A subtypes. However, we witness maximum variation for the Her2 subtype (Fig. 5(a)). Given the limited number of instances (32) for the Her2 subtype, such differences are only anticipated.

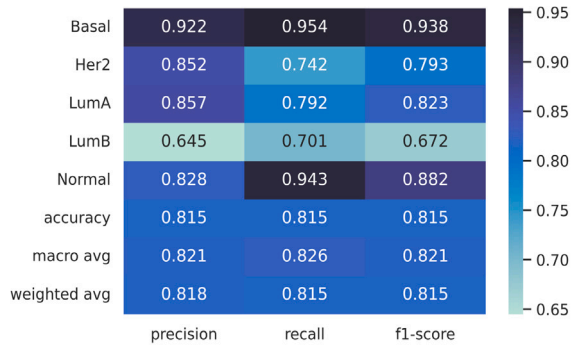
We also compared the performance of discovered biomarkers evaluated using neural network classifier against different classification methods, namely, SVM with radial basis function (SVM-rbf), SVM with the linear kernel (SVM-linear), and Random Forest (RF) in Table 3. It may be noted that the mean 5-fold cross-validation accuracy of 0.8145 resulting from the proposed neural network classifier is better than the other three classifiers.

### *3.3. Comparison of XAI-based feature selection with state-of-the-art feature selection methods*

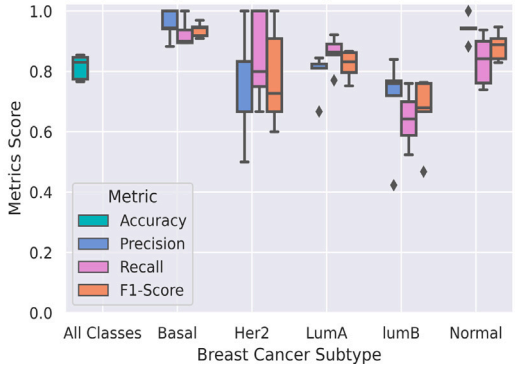
The XAI-based feature selection methodology is compared with various competitive feature selection methods, namely Mutual Information (MI), Recursive Feature Elimination (RFE), Random Forest (RF),




(a) Confusion Matrix signifying number of samples correctly classified (diagonal entries) for different breast cancer subtypes




(b) Heatmap-Summarization of precision, recall, and F1-score for different breast cancer subtypes with deeper hues denoting greater precision, recall, and F1-score values.




(c) BoxPlot indicating the variation in overall accuracy and variation in precision, recall, and f1-score across different breast cancer subtypes.

**Fig. 5.** Classification performance using 5-fold cross-validation for TCGA breast cancer methylation data using identified 52 biomarkers.

**Table 3**
Comparison of the performance of discovered biomarkers evaluated using neural network classifier against SVM-rbf, SVM-linear, and RF. The 5-fold cross-validation accuracy yielded by the proposed neural network classifier is better as compared to the other three classifiers that yielded comparable accuracy values.

| Fold | Classifier | | | |
|---|---|---|---|---|
| | SVM-rbf | SVM-linear | RF | Neural network |
| Fold 1 | 0.8301 | 0.7765 | 0.8201 | 0.7903 |
| Fold 2 | 0.7800 | 0.8100 | 0.8114 | 0.7500 |
| Fold 3 | 0.8096 | 0.8000 | 0.8124 | 0.8306 |
| Fold 4 | 0.8096 | 0.8200 | 0.7961 | 0.8548 |
| Fold 5 | 0.7866 | 0.7992 | 0.8222 | 0.8468 |
| **Mean** | 0.8032 | 0.8011 | 0.8124 | 0.8145 |

Extreme Gradient Boosting (XGB), and Least Absolute Shrinkage and Selection Operator (LASSO). A set of 52 highest-ranking biomarkers was obtained from each of the aforementioned methods. Table 4 shows the classification accuracy obtained using the proposed framework alongwith the other feature selection methods mentioned above. It is evident that the XAI-based feature selection method outperforms the other feature selection methods.

### 3.4. Relevance computation of discovered biomarkers using SHAP

To mark the relative attribution score of the selected 52 genes for breast cancer classification, we analyzed the classifier yielding 0.8145 $\pm$ 0.07 accuracy using the SHAP method. Given the outcome (prediction) of a machine learning model on a data instance, it operates by quantifying the overall contribution (marginal contribution quantified by SHAPley/SHAP values) of different features towards the prediction in a post-hoc manner. To evaluate the contribution $\phi_i$ of a specific feature $i$, several subsets of the feature set are considered, and the model output is evaluated using the true value of the feature in the instance at hand as well as using the baseline value of the feature. The difference between the two outcomes gives a measure of the contribution of feature $i$ for the specific subset. Finally, the contributions of feature $i$ computed for several subsets are averaged to yield the SHAP value of feature $i$, which yielded marginal contribution in terms of SHAP values. Thus, in this manner, using the SHAP library, the relevance /contribution/ Shap value of each gene is computed considering prediction towards a specific breast cancer subtype (output neuron).

Fig. 6 depicts 52 genes ordered by their overall influence on the model's output as measured by their SHAP Values. The color encoding marks the specific subtype, and the width of the bar indicates the extent of contribution of the gene (measured using SHAP value) in the subtype prediction. Also, Figs. 7(a), 7(b), 7(c), 7(d), and 7(e) depict the ten most significant genes for classification of Basal, Her2, Luminal A, Luminal B, and Normal like subtypes respectively. In each of these figures, the horizontal bar against a gene listed on the $y$-axis indicates the influence of the methylation level of the gene on the model's prediction. Red and blue colors indicate the genes' hypermethylation and hypomethylation. For example, for the Her2 subtype (see Fig. 7(b)), PTPRQ is the most relevant gene. Its hypomethylation contributes to the prediction of the Her2 subtype (positive impact on the model's output), and hypermethylation contributes to the prediction against the Her2 subtype, i.e., subtypes other than Her2 (negative impact on the model's output).

**Table 4**
Comparison between various feature selection methods and XAI-based feature selection. It is observed that XAI-based feature selection outperforms the other competitive methods, yielding maximum 5-fold cross-validation accuracy.

| Classifier | Feature Selection Method | | | | | |
|---|---|---|---|---|---|---|
| | Fold | MI | RFE | RF | XGB | LASSO | XAI |
| SVM-linear | Fold 1 | 0.8001 | 0.7901 | 0.7860 | 0.8011 | 0.7500 | 0.7765 |
| | Fold 2 | 0.7847 | 0.8182 | 0.8311 | 0.7900 | 0.7500 | 0.8100 |
| | Fold 3 | 0.7700 | 0.7966 | 0.7635 | 0.7765 | 0.7756 | 0.8000 |
| | Fold 4 | 0.7870 | 0.8012 | 0.8011 | 0.8000 | 0.7420 | 0.8200 |
| | Fold 5 | 0.8000 | 0.7801 | 0.7821 | 0.8000 | 0.7860 | 0.7992 |
| | **Mean** | **0.7884** | **0.7972** | **0.7928** | **0.7935** | **0.7607** | **0.8011** |
| SVM-rbf | Fold 1 | 0.8153 | 0.8100 | 0.8122 | 0.8151 | 0.7900 | 0.8301 |
| | Fold 2 | 0.8000 | 0.8182 | 0.7800 | 0.8198 | 0.7500 | 0.7800 |
| | Fold 3 | 0.8096 | 0.8100 | 0.7900 | 0.7996 | 0.7756 | 0.8096 |
| | Fold 4 | 0.7905 | 0.8102 | 0.8301 | 0.8101 | 0.7400 | 0.8096 |
| | Fold 5 | 0.7866 | 0.8000 | 0.8100 | 0.8043 | 0.7801 | 0.7866 |
| | **Mean** | **0.8004** | **0.8097** | **0.8045** | **0.8098** | **0.7671** | **0.8032** |
| Neural Network | Fold 1 | 0.8212 | 0.8222 | 0.8067 | 0.8300 | 0.7955 | 0.7903 |
| | Fold 2 | 0.8101 | 0.7900 | 0.8100 | 0.8016 | 0.7642 | 0.7500 |
| | Fold 3 | 0.8112 | 0.8021 | 0.7996 | 0.7990 | 0.7605 | 0.8306 |
| | Fold 4 | 0.7867 | 0.8137 | 0.8210 | 0.8141 | 0.7802 | 0.8548 |
| | Fold 5 | 0.8001 | 0.8001 | 0.7905 | 0.8230 | 0.7796 | 0.8468 |
| | **Mean** | **0.8059** | **0.8056** | **0.8055** | **0.8135** | **0.7760** | **0.8145** |
| KNN | Fold 1 | 0.7745 | 0.7996 | 0.7989 | 0.8000 | 0.7451 | 0.8203 |
| | Fold 2 | 0.7965 | 0.8015 | 0.8063 | 0.7845 | 0.7620 | 0.7900 |
| | Fold 3 | 0.7601 | 0.8002 | 0.7812 | 0.7900 | 0.7430 | 0.7996 |
| | Fold 4 | 0.7800 | 0.7987 | 0.7886 | 0.7845 | 0.7765 | 0.7996 |
| | Fold 5 | 0.7899 | 0.7760 | 0.7965 | 0.7952 | 0.7614 | 0.7866 |
| | **Mean** | **0.7802** | **0.7952** | **0.7943** | **0.7908** | **0.7576** | **0.7992** |
| Adaboost | Fold 1 | 0.7965 | 0.7901 | 0.7860 | 0.8001 | 0.7510 | 0.8096 |
| | Fold 2 | 0.7810 | 0.8003 | 0.8311 | 0.7900 | 0.7321 | 0.8067 |
| | Fold 3 | 0.7995 | 0.7964 | 0.7635 | 0.7886 | 0.7600 | 0.8002 |
| | Fold 4 | 0.7870 | 0.8012 | 0.8011 | 0.7945 | 0.7750 | 0.7945 |
| | Fold 5 | 0.7833 | 0.7801 | 0.7821 | 0.8000 | 0.7802 | 0.7960 |
| | **Mean** | **0.7894** | **0.7936** | **0.7928** | **0.7946** | **0.7597** | **0.8014** |

**Table 5**
Comparison of classification performance for breast cancer classification on TCGA breast cancer dataset.

| Research Group | Features | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|---|
| XAI-MethylMarker | 52 | 0.8145 | 0.818 | 0.815 | 0.815 |
| List et al. (2014) | 38 | 0.753 | 0.751 | 0.753 | 0.746 |

### 3.5. Comparison with state-of-the-art frameworks

To compare our findings with state-of-the-art studies, we searched for similar works but could only find List et al. (2014) for direct comparison as they experimented on the same dataset (TCGA DNA Methylation). The authors have studied methylation data for a five-class breast cancer classification problem using random forest-based categorization models. They employed the varSelfRF method for recursive feature selection using random forest. Subsequently, they used the Gini index to arrive at a set of 38 DNA methylation features. The authors reported average accuracy of 0.753 over 10 runs of the .632 bootstrap method. In contrast, the proposed framework discovers a competitive set of 52 biomarkers with significantly better accuracy, precision, recall, and f1-score (see Table 5) of 0.8145 ± 0.07, 0.818, 0.815, and 0.815, respectively (compared to 0.753, 0.751, 0.753, and 0.746).

### 3.6. Gene set analysis and trustworthiness of the biomarkers

In this section, we examine the role of discovered 52 biomarkers in differentiating five breast cancer subtypes.

*Visual analysis of DNA methylation biomarkers*

The heatmap in Fig. 8(a) (plotted using PROMO tool (Netanely, Stern, Laufer, & Shamir, 2019) depicts the variation in methylation

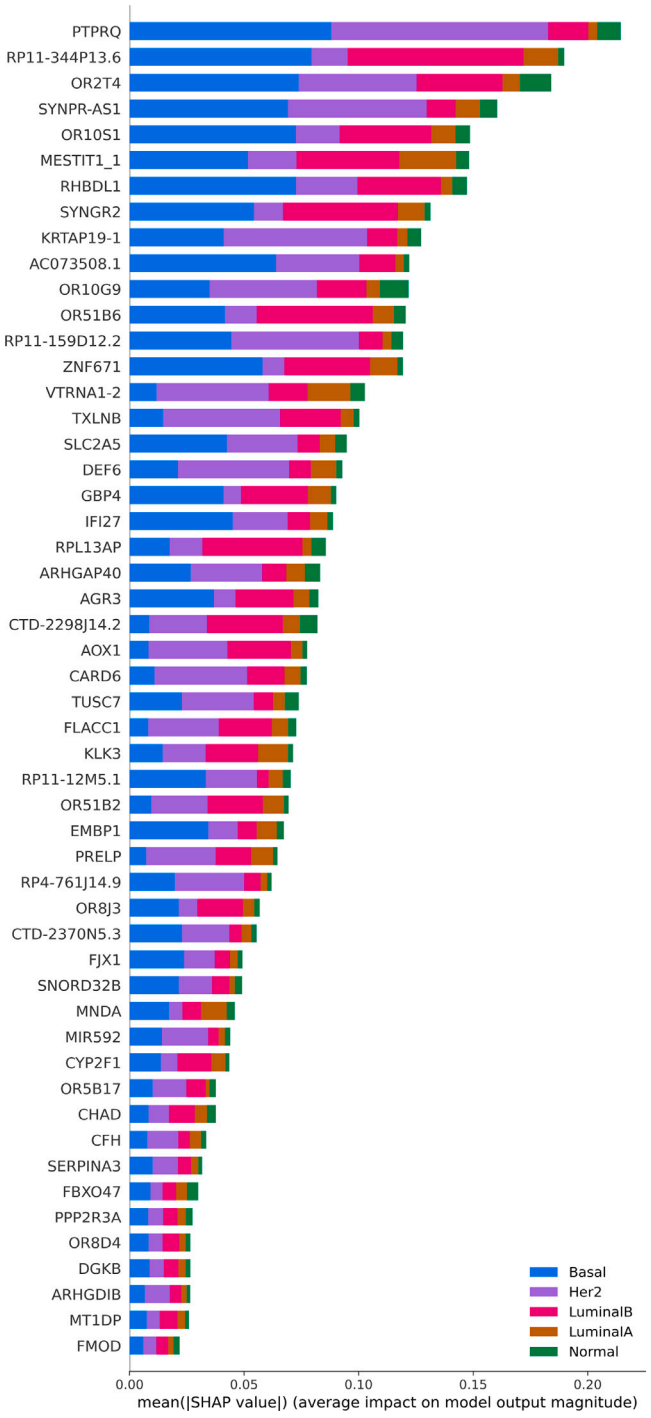
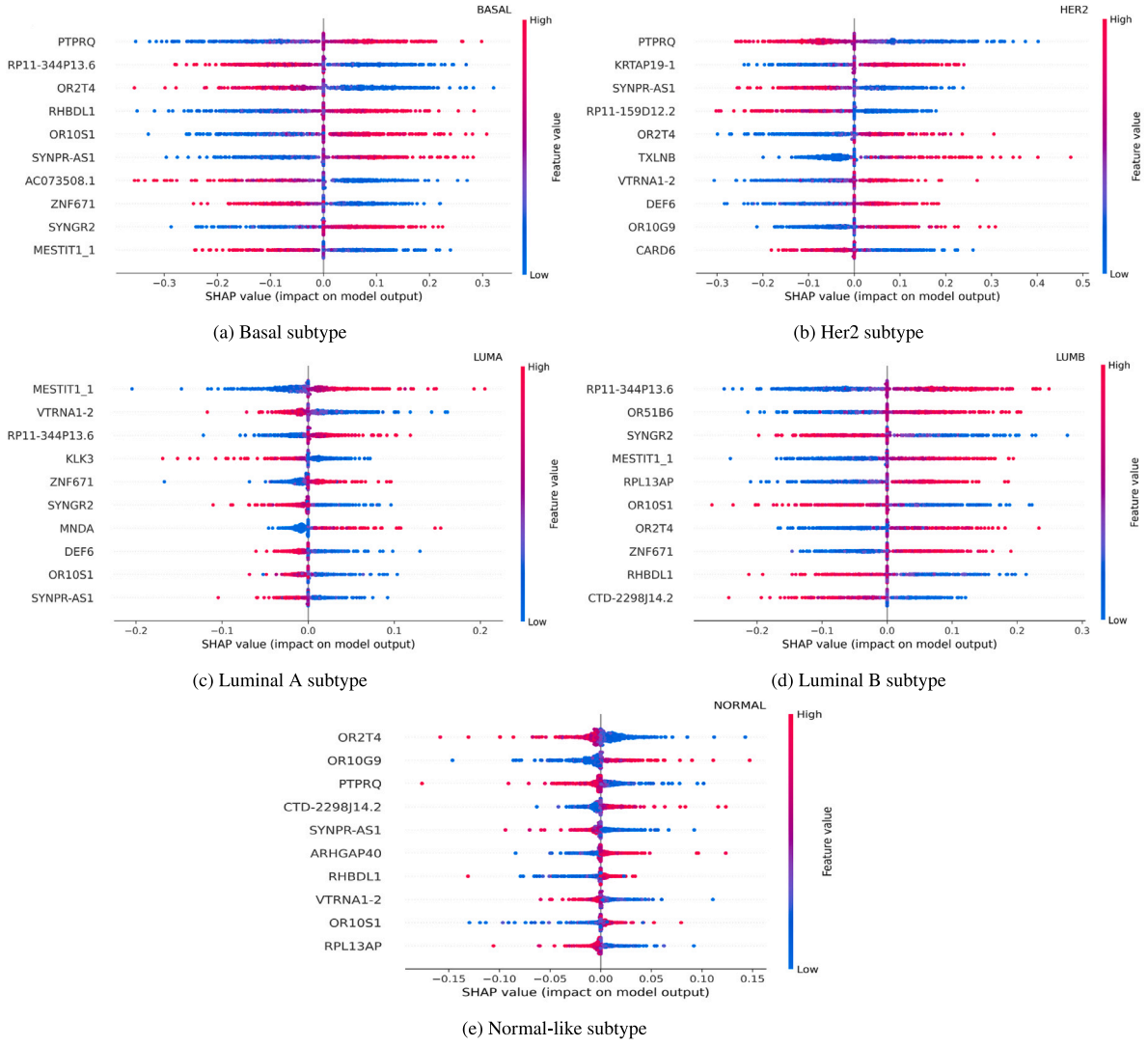



**Fig. 6.** 52 genes ordered by their average impact on the model's output as measured by their SHAP Values. The color encoding marks the specific subtype, and the width of the bar indicates the extent of contribution of the gene (measured using SHAP value) in the subtype prediction.

of the discovered biomarkers across different subtypes. Further, t-Distributed Stochastic Neighbor Embedding (t-SNE) visualization in Fig. 8(b) shows segregation of the five breast cancer types based on the selected biomarkers. The data distribution shows that these genes have an overall aggregated capacity in segregating Basal, Her2, and Normal subtypes, even though samples of LumB and LumA overlap to some extent.




(a) Basal subtype

(b) Her2 subtype

(c) Luminal A subtype

(d) Luminal B subtype

(e) Normal-like subtype

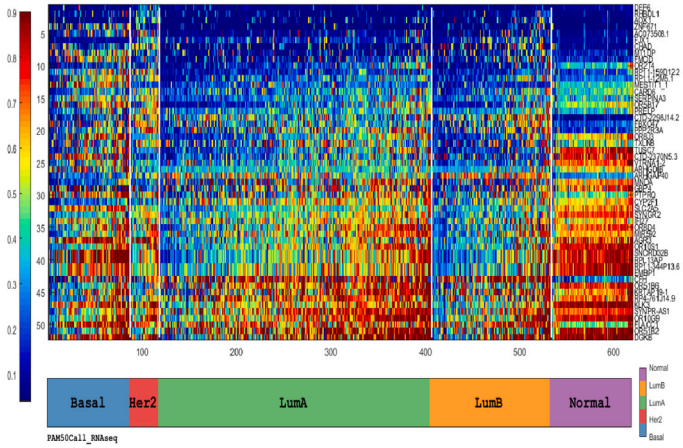
**Fig. 7.** Top 10 significant genes for classification of Basal, Her2, Luminal A, Luminal B, and Normal-like subtypes, respectively. In each of these figures (a–e), the horizontal bar against a gene listed on the *y*-axis indicates the influence of the methylation level of the gene on the model's prediction, and dots depict the instances (patients) in the dataset exhibiting different methylation levels for these genes. Red and blue colors indicate the genes' hypermethylation and hypomethylation, respectively. For example, for the Her2 subtype (see Fig. 7(b)), PTPRQ is the most relevant gene. Its hypomethylation contributes to the prediction of the Her2 subtype (positive impact on the model's output), and hypermethylation contributes to prediction against the Her2 subtype, i.e., subtypes other than Her2 (negative impact on the model's output).

*Reactome pathways enriched by DNA methylation biomarkers*

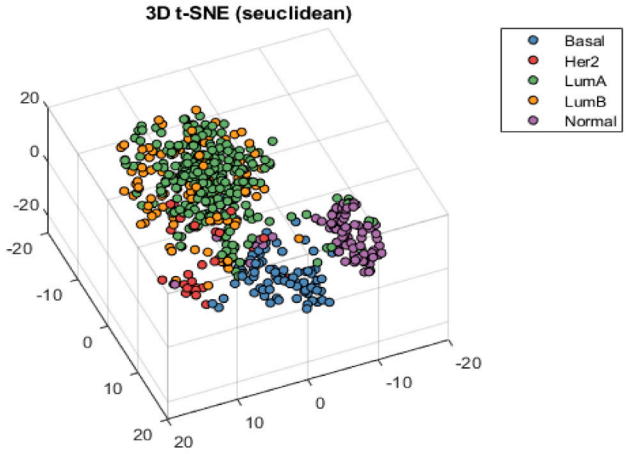
To identify the biological pathways driven by the biomarkers (selected by the *MethylBDA*), we carried out the overrepresentation test (using WebGestalt). It was found that eleven of these biomarker genes, namely FMOD, PRELP, OR10G9, OR10S1, OR2T4, OR51B2, OR51B6, OR5B17, OR8D4, OR8J3, and DGKB lead to seven enriched Reactome pathways with FDR corrected (False Discovery Rate) p-values less than 0.05. The enriched pathways, namely, Defective CHST6 causes MCDC1, Defective ST3GAL3 causes MCT12 and EIEE15, Defective B4GALT1 causes B4GALT1-CDG, Olfactory Signaling Pathway, G alpha(s) signaling events, GPCR downstream signaling, and Signaling by GPCR, have been shown in dark blue in Fig. 9. These pathways are known to be linked with different breast cancer subtypes. For example, G protein-coupled receptors (GPCRs) are the biggest family of cell-surface receptors involved in the onset and progression of breast cancer, more so with the Basal and Her2 subtypes (Lappano, Jacquot, & Maggiolini, 2018). Several studies suggest that the Olfactory signaling pathway enriched by different olfactory receptors such as OR10G9, OR2T4, OR51B2, and OR51B6 is linked with different breast cancer subtypes (Masjedi, Zwiebel, & Giorgio, 2019; Weber et al., 2018).

*Druggability of DNA methylation biomarkers*

To examine the potential druggability of the discovered set of biomarkers, Drug Gene Interaction Database (DGIdb) is utilized. This online database (https://www.dgidb.org/) takes a list of genes as inputs and outputs the potentially druggable genes. It was found that the proposed framework could discover 14 potentially druggable genes, namely, CFH, RHBDL1, OR51B2, OR10G9, SERPINA3, DGKB, AOX1, PTPRQ, CYP2F1, PPP2R3 A, CARD6, SLC2A5, KLK3, and IFI27. For example, targeting CFH (Complement Factor H) gene, a protein-coding gene, may assist in controlling the progression of breast cancer in luminal A patients (Alshabi, Vastrad, Shaikh, & Vastrad, 2019; Tishchenko, Milioli, Riveros, & Moscato, 2016). Similarly, RHBDL1 is a Ras gene over-expressed in Her2 subtype patients in later stages of cancer. So, a drug that suppresses it may assist in controlling the disease (Canzoneri, Lacunza, Larrain, Croce, & Abba, 2014; Claeys, 2018). Similarly, Diacylglycerol (DG) kinase (DGK) gene is also known to be over-expressed in the case of resistant Her2 breast tumors. As suppression of this gene assists in inducing programmed cell death, it may be targeted for resistant tumors (Lapin et al., 2014; Sakane, Hoshino, Ebina, Sakai, & Takahashi, 2021). Also, as non-methylation of AOX1 is known to have




(a) Heatmap highlighting segregation of the patients of different breast cancer types based on variation in discovered 52 DNA methylation biomarkers




(b) t-SNE Visualization depicts that the discovered 52 DNA methylation biomarkers have an overall aggregated capacity in segregating Basal, Her2, and Normal subtypes, even though samples of LumB and LumA overlap to some extent

**Fig. 8.** Heatmap and T-SNE Visualization using the proposed 52 DNA Methylation Biomarkers.

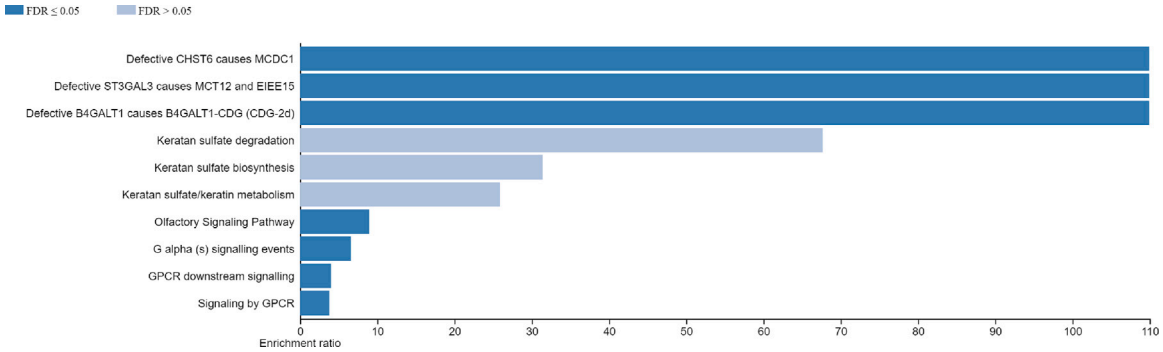



**Fig. 9.** Reactome Pathways with *p*-value less than 0.05 hit by discovered 52 Biomarkers. These enriched pathways include Defective CHST6 causes MCDC1, Defective ST3GAL3 causes MCT12 and EIEE15, Defective B4GALT1 causes B4GALT1-CDG, Olfactory Signaling Pathway, G alpha(s) signaling events, GPCR downstream signaling, and Signaling by GPCR.

an inhibiting effect on the progression of Her2 breast cancer, it may serve as a druggable gene (Zhang, Liu, et al., 2018). PPP2R3 A gene contributes towards regulating PP2 A activity and tumor-related signal proteins. So, it may also be used for the tailored treatment of breast cancer patients (He et al., 2021). Also, KLK3 (Kallikrein-related peptidase 3) appears to be substantially linked with Androgen Response elements for all Breast cancer subtypes (Hanamura et al., 2021). Further, the SERPINA3 gene is known to promote tumor invasion in basal patients. Targeted medicines have been devised for the SERPINA3 gene, which can be used with first-line chemotherapeutic treatments to achieve an enhanced survival rate of basal breast cancer patients (Zhang et al., 2021). Similarly, the IFI27 gene is upregulated in basal patients and

**Table 6**
Proposed framework applied to the multi-omics (RNA+CNV+Methyl) data resulted in the discovery of a set of 127 biomarkers yielding a mean accuracy of 83.80%.

| Random Seed | RNA+CNV+Methyl (127 genes: 37 RNA, 38 CNV, 52 Methylation) |
|---|---|
| 1 | 0.8386 |
| 2 | 0.8479 |
| 3 | 0.8218 |
| 4 | 0.8518 |
| 5 | 0.8481 |
| 6 | 0.8329 |
| 7 | 0.8349 |
| 8 | 0.8310 |
| 9 | 0.8366 |
| 10 | 0.8365 |
| Mean Accuracy | 0.8380 |

can serve as a promising target for therapeutic intervention (Cervantes-Badillo et al., 2020; Li et al., 2015). Thus, these genes can not only guide subtype diagnosis but may also potentially serve as candidates for devising drug therapy.

*Prognostic analysis of DNA methylation biomarkers*

Kaplan–Meier (KM) plotter (Nagy, Lánczky, Menyhárt, & Győrffy, 2018) is a popular tool for investigating the role of a set of genes in prognosis. It segregates the patients into two groups based on the methylation level of genes. To achieve this, it uses the optimal cutoff criterion to identify the split point and generates Kaplan–Meier curves that show the overall survival probability for each group. We found that nine of the discovered biomarkers (having p-values less than 0.05), namely, OR10S1, KLK3, EMPB1, RP11-344P13.6, FLACC1, PPP2R3 A, SYNGR2, DEF6, and CTD-2298JI4.2 are associated with survival. Fig. 10 shows the Kaplan Meier survival curves for these nine genes. These curves clearly distinguish the prognosis of the hypomethylation and hypermethylation groups. The subplots also depict the hazard ratio at 95% confidence interval and logrank *p*-value obtained using the Cox regression model. A hazard ratio of more than two in Figs. 10(a), 10(b), 10(c), 10(d), 10(g), and 10(h) indicates that the patients in the hypomethylation group who are alive at any stage in time have at least twice the chance of dying as patients in the other hypermethylation groups. Again, a hazard ratio of less than 0.5 in Figs. 10(e), 10(f), and 10(i) indicates the patients who are alive in the hypermethylation group at any stage in time have at most half the likelihood of dying as patients in the hypomethylation group. Further, low logrank values indicate that the two groups are clearly distinguishable.

### 3.7. Experimentation on multi-omics data

We applied the proposed framework to the multi-omics (RNA+CNV+Methyl) data, which resulted in the discovery of a set of 127 biomarkers (comprising 37 RNA-Seq genes, 38 copy number variation (CNV) genes, and 52 Methylation genes). Using these 127 biomarkers yielded a mean accuracy of 83.80% for the patients whose all three omic variants were available in the TCGA dataset, outperforming the state-of-the-art method proposed by Lin et al. (2020) (please see Table 6).

## 4. Conclusion and future work

Initial omic data profiling, such as Sanger sequencing and microarray techniques result in single-omic data collection such as single nucleotide polymorphism (which yields copy number variation) and gene expression. The advent of high throughput next-generation sequencing techniques, capable of faster capturing of omic/multi-omic data aids in comprehending the disease and its heterogeneity at the genomic, transcriptomic, and epigenomic levels. This led to the development of a very promising approach—targeted treatment. Epigenomic alterations in the form of DNA methylation occur at an early stage in cancer progression and influence the expression level of genes involved in the further progression of different breast cancer subtypes. Thus, in this paper, we have investigated the impact of these alterations in discriminating between distinct breast cancer subtypes. Towards this end, we have proposed *XAI-MethylMarker*—an explainable AI-based biomarker discovery framework. The framework, when applied to DNA methylation data, led to the discovery of a set of 52 DNA methylation biomarkers capable of differentiating different breast cancer subtypes. Using 5-fold cross-validation, a classification accuracy of $0.8145 \pm 0.07$ at a 95% confidence interval was obtained. The results generated using the proposed framework are competitive with respect to the state-of-the-art. This study is a first of its kind in which the behavior of a deep learning model has been analyzed using methods of explainable AI to uncover the relevance of a small set of DNA methylation biomarkers in discriminating between distinct breast cancer subtypes.

Further, to establish the clinical relevance of the discovered biomarkers, we have performed gene set analysis (GSA). Using GSA, we found that eleven of these biomarker genes, namely FMOD, PRELP, OR10G9, OR10S1, OR2T4, OR51B2, OR51B6, OR5B17, OR8D4, OR8J3, and DGKB lead to seven enriched Reactome pathways (FDR corrected p-values less than 0.05), all of which are known to be significantly associated with distinct Breast Cancer subtypes. Also, close analysis reveals the presence of fourteen druggable genes (namely, CFH, RHBDL1, OR51B2, OR10G9, SERPINA3, DGKB, AOX1, PTPRQ, CYP2F1, PPP2R3 A, CARD6, SLC2A5, KLK3, and IFI27) and nine biomarkers (namely, OR10S1, KLK3, EMPB1, RP11-344P13.6, FLACC1, PPP2R3 A, SYNGR2, DEF6, and CTD-2298JI4.2) linked with the prognostic outcome in discovered set suggesting their potential use in early diagnosis and targeted therapy. This paves the path for applying Explainable AI approaches for various oncological studies. In future research, we intend to apply explainable AI approaches to discover relevant biomarkers for other forms of cancer.

## CRediT authorship contribution statement

**Sheetal Rajpal:** Conceived the experiments and led the development of the framework, Writing – original draft, Led the analysis and interpretation of the data. **Ankit Rajpal:** Writing – original draft, Led the analysis and interpretation of the data. **Arpita Saggar:** Supported in the development of the framework. **Ashok K. Vaid:** Led the analysis and interpretation of the data. **Virendra Kumar:** Writing – review & editing. **Manoj Agarwal:** Writing – review & editing. **Naveen Kumar:** Writing – review & editing, Led the analysis and interpretation of the data.

## Declaration of competing interest

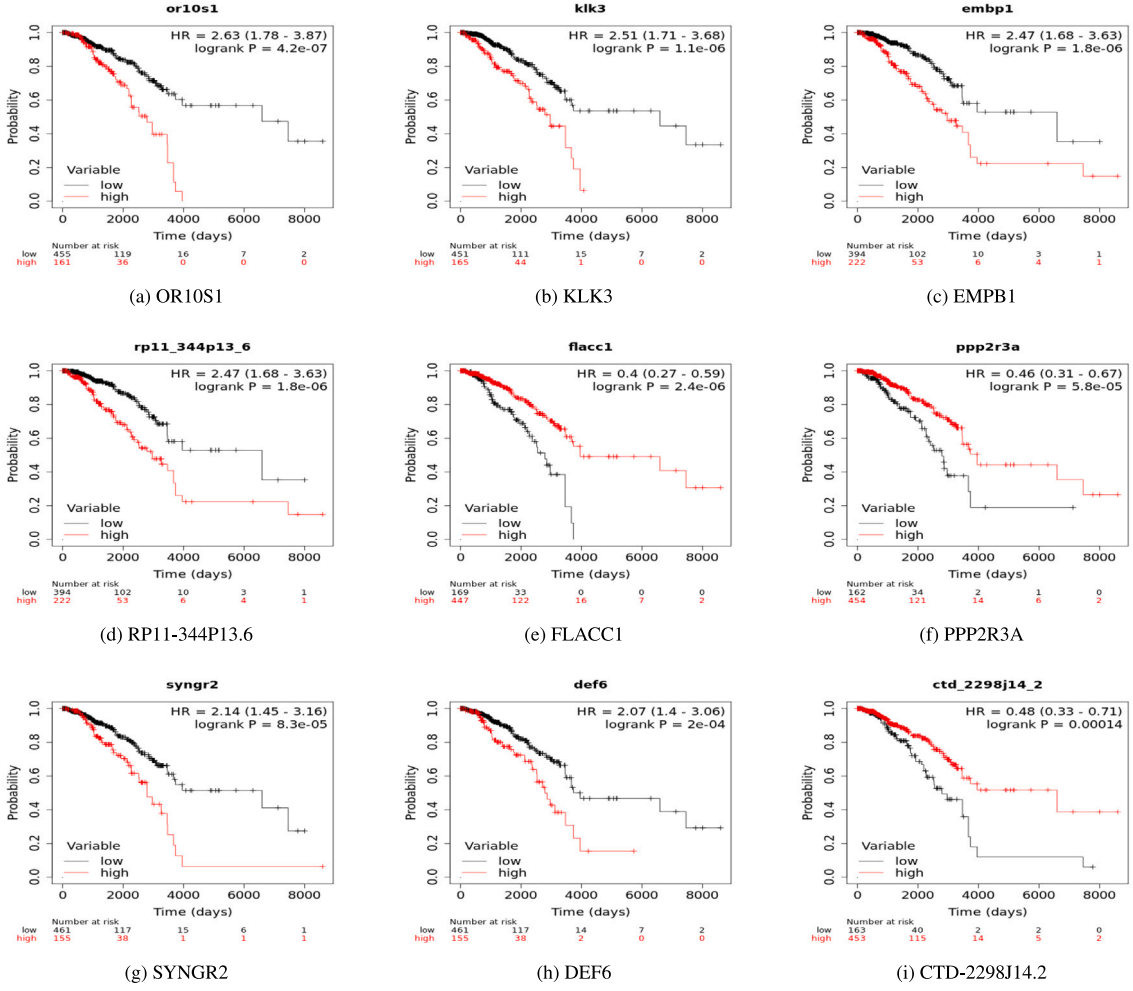
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

Data will be made available on request.

## Acknowledgments

The authors are extremely grateful to Tavpritesh Sethi, Associate Professor, IIIT Delhi for his guidance and insightful comments on the results of the Ph.D. Clinic – ACM India Council initiative. All authors read and approved the final draft of the manuscript.




(a) OR10S1 (b) KLK3 (c) EMPB1

(d) RP11-344P13.6 (e) FLACC1 (f) PPP2R3A

(g) SYNGR2 (h) DEF6 (i) CTD-2298J14.2

**Fig. 10.** Kaplan–Meier curves showing survival probabilities of the two contrasting groups (hypomethylation and hypermethylation groups based on methylation expression). The nine genes out of 52 genes (OR10S1, KLK3, EMPB1, RP11-344P13.6, FLACC1, PPP2R3 A, SYNGR2, DEF6, and CTD-2298JI4.2) are associated with survival having *p*-value less than 0.05.

## References


Alshabi, A. M., Vastrad, B., Shaikh, I. A., & Vastrad, C. (2019). Exploring the molecular mechanism of the drug-treated breast cancer based on gene expression microarray. *Biomolecules*, *9*(7), 282.

Amor, R. d., Colomer, A., Monteagudo, C., & Naranjo, V. (2022). A deep embedded refined clustering approach for breast cancer distinction based on DNA methylation. *Neural Computing and Applications*, *34*(13), 10243–10255.

Canzoneri, R., Lacunza, E., Larrain, M. I., Croce, M. V., & Abba, M. C. (2014). Rhomboid family gene expression profiling in breast normal tissue and tumor samples. *Tumor Biology*, *35*(2), 1451–1458.

Cervantes-Badillo, M. G., Paredes-Villa, A., Gómez-Romero, V., Cervantes-Roldán, R., Arias-Romero, L. E., Villamar-Cruz, O., . . ., et al. (2020). IFI27/ISG12 down-regulates estrogen receptor $\alpha$ transactivation by facilitating its interaction with CRM1/XPO1 in breast cancer cells. *Frontiers in Endocrinology*, *11*, 792.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chen, L., Zeng, T., Pan, X., Zhang, Y.-H., Huang, T., & Cai, Y.-D. (2019). Identifying methylation pattern and genes associated with breast cancer subtypes. *International Journal of Molecular Sciences*, *20*(17), 4269.

Claeys, A. (2018). *An optimized pipeline for matching patients and drug-based therapies in precision oncology* (Ph.D. thesis), Ghent University.

Cristovao, F., Cascianelli, S., Canakoglu, A., Carman, M., Nanni, L., Pinoli, P., et al. (2020). Investigating deep learning based breast cancer subtyping using pan-cancer and multi-omic data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

Daura-Oller, E., Cabre, M., Montero, M. A., Paternain, J. L., & Romeu, A. (2009). Specific gene hypomethylation and cancer: new insights into coding region feature trends. *Bioinformation*, *3*(8), 340.

Eccles, S. A., Aboagye, E. O., Ali, S., Anderson, A. S., Armes, J., Berditchevski, F., . . ., & Thompson, A. M. (2013). Critical research gaps and translational priorities for the successful prevention and treatment of breast cancer. *Breast Cancer Research*, *15*(5), 1–37.

Fakoor, R., Ladhak, F., Nazi, A., & Huber, M. (2013). Using deep learning to enhance cancer diagnosis and classification. In *Proceedings of the international conference on machine learning, vol. 28*. USA: ACM New York.

Hanamura, T., Christenson, J. L., O'Neill, K. I., Rosas, E., Spoelstra, N. S., Williams, M. M., et al. (2021). Secreted indicators of androgen receptor activity in breast cancer pre-clinical models. *Breast Cancer Research*, *23*(1), 1–15.

He, J.-J., Shang, L., Yu, Q.-W., Jiao, N., Qiu, S., Zhu, W.-X., . . ., & Zhang, Q. (2021). High expression of protein phosphatase 2 regulatory subunit B"alpha predicts poor outcome in hepatocellular carcinoma patients after liver transplantation. *World Journal of Gastrointestinal Oncology*, *13*(7), 716.

Holm, K., Hegardt, C., Staaf, J., Vallon-Christersson, J., Jönsson, G., Olsson, H., . . ., & Ringnér, M. (2010). Molecular subtypes of breast cancer are associated with characteristic DNA methylation patterns. *Breast Cancer Research*, *12*(3), 1–16.

Hosni, M., Abnane, I., Idri, A., de Gea, J. M. C., & Alemán, J. L. F. (2019). Reviewing ensemble classification methods in breast cancer. *Computer Methods and Programs in Biomedicine*, *177*, 89–112.

Jaenisch, R., & Bird, A. (2003). Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature Genetics*, *33*(3), 245–254.

Karabulut, E. M., & Ibrikci, T. (2017). Discriminative deep belief networks for microarray based cancer classification. *Biomedical Research-Tokyo*, *28*(3), 1016–1024.

Kim, H. K., Park, K. H., Kim, Y., Park, S. E., Lee, H. S., Lim, S. W., . . ., & Park, Y. H. (2019). Discordance of the PAM50 intrinsic subtypes compared with immunohistochemistry-based surrogate in breast cancer patients: potential implication of genomic alterations of discordance. *Cancer Research and Treatment: Official Journal of Korean Cancer Association*, *51*(2), 737–747.

Kuang, Y., Wang, Y., Zhai, W., Wang, X., Zhang, B., Xu, M., . . ., & Liu, H. (2020). Genome-wide analysis of methylation-driven genes and identification of an eight-gene panel for prognosis prediction in breast cancer. *Frontiers in Genetics*, *11*, 301.

Lapin, V., Shirdel, E., Wei, X., Mason, J., Jurisica, I., & Mak, T. (2014). Kinome-wide screening of HER2+ breast cancer cells for molecules that mediate cell proliferation or sensitize cells to trastuzumab therapy. *Oncogenesis*, *3*(12), e133.

Lappano, R., Jacquot, Y., & Maggiolini, M. (2018). GPCR modulation in breast cancer. *International Journal of Molecular Sciences*, *19*(12), 3840.

Li, S., Xie, Y., Zhang, W., Gao, J., Wang, M., Zheng, G., . . ., & Tao, X. (2015). Interferon alpha-inducible protein 27 promotes epithelial–mesenchymal transition and induces ovarian tumorigenicity and stemness. *Journal of Surgical Research*, *193*(1), 255–264.

Lin, Y., Zhang, W., Cao, H., Li, G., & Du, W. (2020). Classifying breast cancer subtypes using deep neural networks based on multi-omics data. *Genes*, *11*(8), 888.

List, M., Hauschild, A.-C., Tan, Q., Kruse, T. A., Baumbach, J., & Batra, R. (2014). Classification of breast cancer subtypes by combining gene expression and DNA methylation data. *Journal of Integrative Bioinformatics*, *11*(2), 1–14.

Liu, L., Chen, X., & Wong, K.-C. (2021). Early cancer detection from genome-wide cell-free DNA fragmentation via shuffled frog leaping algorithm and support vector machine. *Bioinformatics*, *37*(19), 3099–3105.

Liu, X., Peng, Y., & Wang, J. (2020). Integrative analysis of DNA methylation and gene expression profiles identified potential breast cancer-specific diagnostic markers. *Bioscience Reports*, *40*(5), BSR20201053.

Liu, B., Xie, M., & Udell, M. (2021). Controlburn: Feature selection by sparse forests. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining* (pp. 1045–1054).

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems* (pp. 4768–4777).

Masjedi, S., Zwiebel, L. J., & Giorgio, T. D. (2019). Olfactory receptor gene abundance in invasive breast carcinoma. *Scientific Reports*, *9*(1), 1–12.

Morris, T. J., Butcher, L. M., Feber, A., Teschendorff, A. E., Chakravarthy, A. R., Wojdacz, T. K., et al. (2014). ChAMP: 450k chip analysis methylation pipeline. *Bioinformatics*, *30*(3), 428–430.

Nagy, Á., Lánczky, A., Menyhárt, O., & Győrffy, B. (2018). Validation of miRNA prognostic power in hepatocellular carcinoma using expression data of independent datasets. *Scientific Reports*, *8*(1), 1–9.

Netanely, D., Stern, N., Laufer, I., & Shamir, R. (2019). PROMO: an interactive tool for analyzing clinically-labeled multi-omic cancer datasets. *BMC Bioinformatics*, *20*(1), 1–10.

Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., . . ., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology*, *27*(8), 1160.

Peng, C., Wu, X., Yuan, W., Zhang, X., Zhang, Y., & Li, Y. (2019). MGRFE: multilayer recursive feature elimination based on an embedded genetic algorithm for cancer classification. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *18*(2), 621–632.

Sakane, F., Hoshino, F., Ebina, M., Sakai, H., & Takahashi, D. (2021). The roles of diacylglycerol kinase $\alpha$ in cancer cell proliferation and apoptosis. *Cancers*, *13*(20), 5190.

Spanhol, F. A., Oliveira, L. S., Petitjean, C., & Heutte, L. (2016). Breast cancer histopathological image classification using convolutional neural networks. In *2016 International joint conference on neural networks* (pp. 2560–2567). IEEE.

Stefansson, O. A., Moran, S., Gomez, A., Sayols, S., Arribas-Jorba, C., Sandoval, J., . . ., et al. (2015). A DNA methylation-based definition of biologically distinct breast cancer subtypes. *Molecular Oncology*, *9*(3), 555–568.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, *71*(3), 209–249.

Taherian-Fard, A., Srihari, S., & Ragan, M. A. (2015). Breast cancer classification: linking molecular mechanisms to disease prognosis. *Briefings in Bioinformatics*, *16*(3), 461–474.

Tao, M., Song, T., Du, W., Han, S., Zuo, C., Li, Y., . . ., & Yang, Z. (2019). Classifying breast cancer subtypes using multiple kernel learning based on omics data. *Genes*, *10*(3), 200.

Tishchenko, I., Milioli, H. H., Riveros, C., & Moscato, P. (2016). Extensive transcriptomic and genomic analysis provides new insights about luminal breast cancers. *PLoS One*, *11*(6), Article e0158259.

UCSC (2016). UCSC xena. https://xenabrowser.net/datapages/?hub=https://tcga.xenahubs.net:443. (Accessed 06 February 2020).

Vallejos, C. S., Gómez, H. L., Cruz, W. R., Pinto, J. A., Dyer, R. R., Velarde, R., . . ., et al. (2010). Breast cancer classification according to immunohistochemistry markers: subtypes and association with clinicopathologic variables in a peruvian hospital database. *Clinical Breast Cancer*, *10*(4), 294–300.

Weber, L., Maßberg, D., Becker, C., Altmüller, J., Ubrig, B., Bonatz, G., . . ., et al. (2018). Olfactory receptors as biomarkers in human breast carcinoma tissues. *Frontiers in Oncology*, *8*, 33.

Wei, L., Jin, Z., Yang, S., Xu, Y., Zhu, Y., & Ji, Y. (2018). TCGA-assembler 2: software pipeline for retrieval and processing of TCGA/CPTAC data. *Bioinformatics*, *34*(9), 1615–1617.

Withnell, E., Zhang, X., Sun, K., & Guo, Y. (2021). XOmiVAE: an interpretable deep learning model for cancer classification using high-dimensional omics data. *Briefings in Bioinformatics*, *22*(6), bbab315.

Wu, Z.-H., Tang, Y., & Zhou, Y. (2021). DNA methylation based molecular subtypes predict prognosis in breast cancer patients. *Cancer Control*, *28*, Article 1073274820988519.

Xiao, Y., Wu, J., Lin, Z., & Zhao, X. (2018). A deep learning-based multi-model ensemble method for cancer prediction. *Computer Methods and Programs in Biomedicine*, *153*, 1–9.

Xu, Y., Jia, Z., Wang, L.-B., Ai, Y., Zhang, F., Lai, M., . . ., & Chang, C. (2017). Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. *BMC Bioinformatics*, *18*(1), 281.

Zhang, H., Liu, Y., Ge, A., Wang, X., Sun, H., Bi, H., . . ., & Zhao, Y. (2018). Association between AOX1, IRF4 methylation in peripheral blood leukocyte DNA and the risks of breast cancer: a case-control study. *Zhonghua Liu Xing Bing Xue Za Zhi=Zhonghua Liuxingbingxue Zazhi*, *39*(9), 1265–1269.

Zhang, Y., Tian, J., Qu, C., Peng, Y., Lei, J., Li, K., . . ., & Liu, S. (2021). Overexpression of SERPINA3 promotes tumor invasion and migration, epithelial-mesenchymal-transition in triple-negative breast cancer cells. *Breast Cancer*, *28*(4), 859–873.

Zhang, S., Wang, J., Ghoshal, T., Wilkins, D., Mo, Y.-Y., Chen, Y., et al. (2018). lncRNA gene signatures for prediction of breast cancer intrinsic subtypes and prognosis. *Genes*, *9*(2), 65.