HELM-BRCA: Hybrid Embedding and Learning Model for BRCA Methylation Classification

Hemalatha D[1], N Gomathi[2]
Research Scholar, Department of Computer Science & Engineering,
Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, India[1]
Professor, Department of Computer Science & Engineering,
Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, India[2]

***Abstract*—Breast cancer remains a highly heterogeneous disease for which it demands advanced computational techniques that can reveal significant biological patterns in high-dimensional epigenomic data. DNA methylation profiles generated by the Illumina HumanMethylation450 platform yield rich, clinically relevant signals but introduce significant analytical challenges due to their high dimensionality, sparsity, and nonlinear structure. This work presents a novel memory-efficient hybrid learning architecture that combines Truncated Singular Value Decomposition (SVD), a deep Autoencoder, and a multi-model ensemble classifier for boosting subtype classification performance using TCGA-BRCA methylation data. In order to circumvent memory limits and prevent system crashes, a probe-subset extraction strategy combined with variance-based feature selection was employed to ensure fast and safe data loading from the Xena repository. While the autoencoder extracts compact nonlinear manifold representations, SVD captures the global linear variance structure. Further, the fused latent space is modelled by an ensemble including Random Forest, XGBoost, and a lightweight Keras neural classifier that allows the system to exploit different decision limits and achieve robust generalization. The experimental investigation across several architectures demonstrates high predictive performance with ROC-AUC scores exceeding 0.99 and accuracies higher than 0.96 for Basic CNN and MLP models. Furthermore, the proposed hybrid ensemble improves stability and precision by outperforming traditional baselines and confirming the complementary nature of spectral and deep feature extraction. This study is suitable for large-scale biomedical data analytics scenarios. In conclusion, this work provides an efficient hybrid machine learning framework for breast cancer methylation study by offering a strong platform for improved prognostic modelling and development of epigenetic biomarkers.**

Keywords*—Breast cancer classification; DNA methylation; TCGA-BRCA; Truncated SVD; autoencoder; ensemble learning; deep learning; epigenomic biomarkers; hybrid model; machine learning pipeline; high-dimensional data

I. Introduction

Breast cancer is considered one of the highest cancer morbidity and mortality cases in women globally. In spite of the tremendous progress in the area of early detection, molecular profiling, and targeted therapies, breast cancer has currently been taken as a biologically heterogeneous disease with several different molecular subtypes with genetical, epigenetical, and phenotypic features. This heterogeneity directly influences disease courses, therapy response, and patient prognosis, and careful molecular-level stratification strategies are considered necessary to enable precision oncology.

Omics technologies that operate with high throughput have also made breast cancer research faster, as they allow massive profiling of genomic and epigenomic changes. Among them, the DNA methylation has turned out to be a strong and durable epigenetic biomarker because of its key contribution to the regulation of genes, fine-tuning of the chromatin, genomic stability, and X-chromosome inactivation. The methylation abnormalities, including promoter hypermethylation and global hypomethylation, are closely linked with cancer initiation and progression, and hence, methylation profiling is very informative in cancer subtype discrimination and prognosis.

The quantification of genome-wide methylation at more than 480,000 CpG sites has been made possible through the availability of epigenome-wide platforms like the Illumina HumanMethylation450 BeadChip. The TCGA-BRCA data, among others, is a comprehensive dataset of methylation profiles of breast cancer, but with such a large size of the data (compared to the number of samples), there are significant analytical issues like multi-collinearity, sparsity, over-fitting, and high computational costs. Traditional statistical methods and individual machine learning models usually cannot identify credible patterns in such data without efficient dimensionality reduction and feature-selection methods.

The machine learning and deep learning methods have demonstrated potential to overcome these issues. Classical machine learning algorithms, such as Support Vector Machines, Random Forests, Logistic Regression, and boosting-based ensembles, have shown good performance in cancer classification with the help of the right choice of features. Simultaneously, deep learning networks, especially autoencoders, can also be used to do nonlinear dimensionality reduction, allowing the latent to be contained locally and compactly within the latent space, allowing complex patterns in methylation to be represented. Nevertheless, small sizes of clinical samples are frequently a limiting factor in deep learning models, and these models tend to be overfit when trained on their own, which explains the interest in hybrid frameworks that combine the strengths of both machine and deep learning strategies.

Selection of features is a highly sensitive process in the classification process with methylation since only a limited number of CpG sites make a significant contribution to subtype classification. It is thus necessary to use filter, wrapper, and embedded selection methods to extract informative biomarkers and minimize redundancy and noise. Since there is no feature-selection strategy that is optimal universally, we need to use comparative and hybrid methods so as to get good, robust, and generalizable performance. Moreover, the clinical importance of DNA methylation, because it remains stable and can be detected in less invasive specimens like circulating tumor DNA, highlights its potential in the early detection of the disease, subtype, and monitoring of the disease.

This work suggests a Hybrid SVD-Autoencoder-Ensemble framework that can be used to classify breast cancer based on TCGA-BRCA 450K DNA methylation data. The suggested multistage architecture combines: 1) a linear spectral decomposition to reduce dimensionality and reduce multicollinearity, 2) a nonlinear learning representation using an auto-encoder to capture complex methylation patterns, and 3) an ensemble classifier to increase predictive power and generalization.

The rest of the study is structured in the following way: Section II will entail a thorough review of the literature related to DNA methylation-based breast cancer classification and current machine learning and deep learning methods. Section III will describe the dataset, preprocessing pipeline, feature-selection strategies, and the proposed hybrid model. Section IV will report and discuss experimental results and a comparative analysis, and Section V will conclude the study with key findings and directions of further research.

II. Related Works

The most recent studies on breast cancer have begun to pay increased attention to the use of DNA methylation as a strong e-methylation biomarker with machine learning and deep learning frameworks. A framework of deep learning to combine the overall genome-wide DNA methylation profiles is suggested to distinguish between the subtypes of breast cancer. Their model, based on TCGA-BRCA data, showed a good predictive accuracy and the significance of hierarchical feature learning to learn subtype-specific methylation patterns. Nonetheless, the research was based on one deep architecture and not a hybrid representation based on linear and nonlinear feature abstractions [1].

The work [2] proposed a subsequent representation learning method of high-dimensional DNA methylation data through unsupervised methods of learning. The study has shown that the latent embeddings that are learnt unlabeled, can be useful in revealing subtype structure and patterns related to survival. Although the approach was promising in regard to the exploratory analysis, it was not a combination with supervised ensemble classifiers, which can also enhance discriminative performance. A hybrid machine learning system that integrated several classifiers to profile and study epigenetic cancer is suggested [3]. The work they conducted emphasized the advantage of ensemble learning in stabilizing predictions made on heterogeneous features of methylation. However, the model mainly involved feature selection done by hand and lacked deep latent feature learning.

The process [4] explored DNA methylation-mediated biomarkers of the prognosis of breast cancer and immunotherapy response. Their results showed that using the methylation-based signatures could stratify patients according to their survival rate and immunological sensitivity. Even though the computational pipeline was biologically intuitive, it was restricted to traditional ML models and failed to capture the interactions between features in the nonlinear setting. The investigation on the clinical value of epigenetic modification in breast cancer through the use of machine learning-based feature reduction methods. Their experiment showed that prudent dimensionality reduction provides great enhancement to the classification method and minimized computational cost. Nevertheless, the reduction plan was completely linear, which could have ignored nonlinear interactions between methylation [5].

The author came up with a novel deep learning model that converted DNA methylation beta values into image-like representations, which allowed convolutional neural networks to learn spatial patterns. The model performed well in the prediction of cancer origin, thus demonstrating that it is important to re-encode methylation data. Although new, the methodology involved a complicated set of data transformation operations, which can reduce the interpretability [6]. The article [7] provided an in-depth examination of epigenetic signatures to track the progression of diseases and the response of therapy in breast cancer. The study noted how dynamic the changes in methylation are and the clinical significance of epigenetic biomarkers. The work, however, paid more attention to biological interpretation rather than high-level optimization in computation.

A deep learning meta-omics model is suggested [8] that combines the information of methylation, gene expression, and copy number variation. Their findings indicated that they classified subtypes better than single-omics models. The enhanced complexity of multi-view integration, however, creates the problem of scalability and interpretability of the models. The work [9] has reviewed and implemented deep learning in cancer epigenetics and has shown that hierarchical models can be used to learn nonlinear regulatory patterns in methylation data. Their results confirmed that DL is better than classical ML when it comes to more complex epigenomic applications, but there was little discussion of viable implementation factors.

A deep embedded clustering method of distinguishing breast cancer using DNA methylation profiles is introduced [10]. Their unmonitored structure perfected subtype segregation among traditional clustering measures. Its inability to provide direct application to diagnostic tasks was due to the absence of supervised performance evaluation. The examination of feature selection methods along with deep learning in cancer prediction with the help of methylation markers. Their findings validated that the choice of informative CpG sites is a major predictive accuracy and computation efficiency improvement strategy. The research, however, did not examine feature fusion across several representation spaces [11].

A hybrid model of singular value decomposition (SVD) and autoencoders for biomedical image segmentation was suggested [12]. Though they are addressing data on MRI, and not on methylation, their article has laid the groundwork for the efficacy of cascading linear and nonlinear dimensionality reduction methods, which has become a powerful methodology framework in omics applications. The article [13] gave an extensive summary of statistical and machine learn methods used to analyze DNA methylation data. Some of the challenges they highlighted included high dimensionality, batch effects, and biological variability, which necessitate robust computational frameworks. An epigenomic regulation and cancer heterogeneity, and how transcriptional dysregulation is contributed by variation in methylation, is examined [14]. Their results supported the opinion that methylation is a proactive regulation tool but not a passive biomarker.

An autoencoder is used to learn breast cancer recurrence based on the methylation data [15]. In their research, they proved that compressed latent representations are capable of preserving biologically meaningful information and increasing the accuracy of recurrence prediction. The study [16] suggested a multi-stage deep learning pipeline to cancer epigenomics, which consists of feature extraction, representation learning, and classification. They had a good performance of their pipeline but based on single deep architecture without ensemble fusion.

A graph neural network is proposed for the prediction of the subtype of breast cancer based on multiple types [17]. Their method modeled the regulatory dependence of molecules by studying their interactions as graphs, which flat feature models could not detect. Nevertheless, the demand of multi-omics data restricts the applicability where the data on methylation are available alone. Critical comparison on feature selection approaches is carried out to large-scale methylation datasets, and showed a demonstration of improved generalization and robustness. Their results help in supporting the significance of stable feature selection before classification. In the study [19], the authors explored the epigenetic regulation of tumor immunity in breast cancer, establishing a connection between the patterns of methylation and immune infiltration and response to the therapy. Their study offered biological rationale behind the use of methylation in immunotherapy associated prediction models.

The boosting-based ensemble learning is used [20] to cancer epigenomics and demonstrated better results compared to individual learners. Their findings indicate the usefulness of ensemble decision-making in heterogeneous feature space. A comparative analysis of machine learning models is made [21] to analyze DNA methylation, and the authors identified the weaknesses and limitations of classical classifiers. Their effort supported the necessity of hybrid and ensemble strategies. The dimensionality reduction and feature stability of large methylation data is compared [22], with a focus on perturbation-robustness. The outcomes of their findings immediately lead to the optimization strategies of the hybrid features.

An autoencoder-based representation learning model of the methylation data is introduced [23], which was shown to reduce the size of features with little information being lost. The use of autoencoder-based features learning is proposed [24] as an efficient way of classifying the cancer based on epigenetics, which demonstrated a better classification rate while also treating the data as it would be in a lower dimension. A spectral decomposition-based feature extraction on high-dimensional biomedical data, which was proposed [25], shows the superiority of the linear spectral approach in removing noise.

III. Methodology

A. Dataset

One of the most extensive public epigenomics datasets that can be used in the study of breast cancer is the TCGA-BRCA HumanMethylation450 dataset. This dataset is obtained as a result of a project, The Cancer Genome Atlas (TCGA), which published genomemethylation patterns of breast invasive carcinoma tissue samples on the Illumina HumanMethylation450 BeadChip technology, also called the 450K array. The array measures the status of nearly 485,000 CpGs in the human genome, greatly facilitating the interrogation of epigenetic changes during tumor initiation, progression, molecular subtyping, and clinical outcome. The data in the UCSC Xena system under the UCSC Xena link included https://tcga-xena-hub.s3.us-east1.amazonaws.com/download/TCGA.BRCA.sampleMap/HumanMethylation450.gz.

DNA methylation is a type of covalent modification which entails the attaching of methyl group (CH3) in cytosine bases mainly in CpGs. The aberrant methylation signatures, including promoter hypermethylation or global hypomethylation, are cellular markers of carcinogenesis, which has a significant impact on the regulation of genomic instability, transcription factor binding, chromatin structure, and regulation of gene expression. Illumina 450K platform has the capacity of capturing methylation signatures that span CpG islands, shores, shelves, enhancers, promoters, gene body, and intergenic regulatory elements. Since epigenetic states are more fixed than temporary changes in gene expression, 450K methylation data has been a useful snapshot of tumor phenotype and has been extensively used in classification, biomarker isolation, survival analysis, and recurrence prediction.

Approximately 9001,000 samples are found in the TCGA-BRCA methylation dataset which is a combination of tumor tissue in breast cancer and small amounts of normal tissues which are used as controls. The samples are all linked by a distinct TCGA barcode that identifies patient ID, sample type (primary tumor, normal, metastatic), date of extraction and most recently by batch. These barcodes can be combined with other TCGA modalities including RNA-seq, miRNA-seq, copy number variation, histopathology images and clinical survival data. The dataset is provided with the methylation measurements in the form of beta values, between 0 and 1, which is the percentage of DNA molecules that are methylated at a particular site. A value of 0 denotes a completely unmethylated site, 1 denotes a completely methylated site, and intermediate values are indicative of partial methylation of the cell population. The reasons why beta values are popular in classification studies are that they can be interpreted, they are limited, and they are biologically significant. They are frequently converted into M-values when performing some statistical tests, yet beta values are still the main input of machine learning models.

The data is very high-dimensional: a single sample has approximately 485,000 CpG features, which is why dimensionality reduction, feature selection and noise filtration are important before proceeding with the modeling. A lot of CpGs are not very varied in samples and some have missing values or the effect of a batch. Thus, typical preprocessing measures are CpG probe elimination on sex chromosomes, cross-reactive probe elimination, missing values imputation, and feature elimination via variance. Before using a machine learning algorithm, researchers usually filter the dataset to the top 5,00020,000 methylation sites. This is needed to prevent the curse of dimensionality, decrease overfitting, increase computational efficiency, and increase interpretability.

CpG sites are annotated to such genomic elements as promoter regions, e.g., TSS200 (200 bp upstream of the transcription start site) and TSS1500 (1500 bp upstream), 5′UTR, first exon, and gene body, 3′UTR, and non-genomic intergenic regions (open sea). Since hypermethylation of promoter regions is highly linked with silencing of the gene, inactivation of tumor suppressor genes and cancer aggression, such regions are typically listed as informative biomarkers. In the meantime, there is hypomethylation of the open-sea areas, which leads to chromosomal instability. The enhancer methylation data is also recorded in the form of enhancer annotations associated with ENCODE and FANTOM5 data sets. Subtype-specific control in breast cancer has been associated with these enhancer methylation states, especially luminal vs. basal-like tumors.

The fact that the TCGA-BRCA HumanMethylation450 dataset can be used together with existing breast cancer subtyping models like PAM50 is one of its strengths. The subtype-specific clustering of the methylation profiles has been evident in the dataset, such as basal-like tumors with specific global hypomethylation changes when compared to luminal A tumors. Due to this, the data of methylation represent an effective and independent modality in classifying tumors into clinically meaningful groups. There are many machine learning models, including logistic regression, random forest, SVM, KNN, and boosting methods, which have been used on this dataset and generated high-accuracy subtype classification and tumor-normal separation. Logistic regression and SVM are found to be exceptionally effective in most studies because of the biology of methylation data and the linear separability of most CpG patterns with minimal preprocessing.

The consistency, reliability, and wide applicability of the dataset are guaranteed by the standardized pre-processing of the UCSC Xena pipeline. The file is in gzip-compressed form (.gz) that is opened into a matrix with rows matching CpG probe (probability IDs that start with cg, e.g. cg00000029) and columns matching patient samples. Illumina manifest files contain probe annotations in the form of genomic coordinates, gene names, CpG island associations, feature classifications which can be downloaded separately and joined with the dataset on-demand. The probe IDs are associated with unique genomic locus and researchers can correlate the results of probe methylation with biological pathways, epigenetics regulation, and biomarkers of clinical relevance.

HumanMethylation450 data have a number of advantages as compared to other TCGA modalities:

- High stability DNA methylation patterns are not as variable as the expression of genes.
- Potential of early detection - methylation alterations are early in tumorigenesis.
- Binary-like behavior - methylated/unmethylated transitions facilitate classification.
- Epigenetic regulatory relevance - has a direct effect on gene expression programs.
- Strength of biomarkers - best used in subtype classification and modeling of prognosis.

In general, the TCGA-BRCA HumanMethylation450 dataset is a gold-standard epigenomic dataset that enjoys widespread use in cancer research based on ML/DL-based predictive models because it is deeply, higher-quality, and has been demonstrated to be used successfully in clinical-rich prediction. Table I lists the features of the dataset being used.

TABLE I. FEATURES TABLE - HUMANMETHYLATION450 DATASET OVERVIEW

Category	*Description*
Dataset Source	TCGA-BRCA (Breast Invasive Carcinoma)
Platform	Illumina HumanMethylation450 BeadChip
File	HumanMethylation450.gz (UCSC Xena)
Samples	~900–1000 TCGA breast cancer samples
Sample Types	Primary tumor, normal tissue, metastasis
Probes (CpG Sites)	~485,000 CpG beta values
Probe ID Format	cgXXXXXXXX (e.g., cg00000029)
Genome Build	hg19
Data Type	Continuous beta values (0–1)
Methylation Value Meaning	0 = unmethylated, 1 = fully methylated
Genomic Coverage	CpG islands, shores, shelves, open sea
Gene-Context Features	TSS200, TSS1500, 5′UTR, first exon, gene body, 3′UTR
Enhancer Coverage	ENCODE + FANTOM5 annotated enhancers
Clinical Integration	PAM50 subtype labels, survival data (external)
Common Preprocessing	Probe filtering, variance filtering, imputation
Common ML Uses	Tumor classification, subtype prediction, recurrence prediction
Advantages	Stable biomarker, high resolution, strong predictive power

B. Mathematical Model for Algorithms Used

The TCGA-BRCA HumanMethylation450 dataset contains high-dimensional β-values representing the proportion of methylated cytosines at individual CpG loci. Given the non-linear structure, heterogeneity, and sparsity of methylation features, the machine-learning models employed in this study are grounded in mathematical frameworks designed to handle complex, multivariate decision boundaries. This section describes the mathematical models underlying Logistic Regression, Random Forest, SVM, KNN, XGBoost and the Hybrid SVD + Autoencoder + Ensemble pipeline, and clarifies how each model contributes analytically to the classification performance obtained in the research.

1) Logistic regression (LR): Logistic Regression is a generalized linear classifier that models the probability of a binary outcome using the sigmoid function.

Given a feature vector $x \in R^n$, the model predicts, as in Eq. (1):

$$P(y = 1 \mid x) = \sigma(w^T x + b) = \frac{1}{1+e^{-(w^T x+b)}} \quad (1)$$

where,

wis the weight vector,

bis the bias term,

$\sigma(\cdot)$ is the logistic function.

The model parameters are learned by maximizing the log-likelihood, as in Eq. (2):

$$\mathcal{L}(w,b) = \sum_{i=1}^{m}[y_i \log(\sigma(z_i)) + (1-y_i)\log(1-\sigma(z_i))] \quad (2)$$

with $z_i = w^T x_i + b$.

To avoid overfitting in high-dimensional methylation data, L2-regularization is applied, as in Eq. (3):

$$J(w) = -\mathcal{L}(w) + \lambda \parallel w \parallel_2^2 \quad (3)$$

The accuracy (0.9775), ROC-AUC (0.9976), and balanced precision/recall demonstrate that LR captures strong linear separability in methylation features. Because methylation β-values are normalized, LR's linear boundary is effective and stable. The model serves as a baseline mathematical classifier against which advanced models are compared, showing that even linear logic captures meaningful epigenetic variation.

2) Random Forest (RF): Random Forest is a set of decision trees. For a tree T_j, prediction is as in Eq. (4):

$$\hat{y}_j = T_j(x) \quad (4)$$

Majority voting is used to get the final classification, as in Eq. (5):

$$\hat{y} = \text{mode}\{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_k\} \quad (5)$$

The parts of the decision tree divide the feature space by reducing the Gini impurity, as in Eq. (6):

$$G = \sum_{c=1}^{C} p_c(1-p_c) \quad (6)$$

where, p_c is the percentage of samples of class cin that node.

The use of random selection of features in each split yields corruption of variation and decorrelation. RF has high recall (0.9870) and is also powerful in non-linear feature interactions. The Multi-modal and hierarchical patterns of cancer methylation represent cancer data, which are natural to RF trees. It has a mathematical form that gives:

- implicit feature selection
- distinctive multi-divided decision surfaces.
- noise-tolerant classification

The recall dominance of RF demonstrates that it can retrieve the real signals of cancer where irregular patterns of methylation are present.

3) Support vector machine (SVM): In the case of data that are linearly separable, SVM determines the hyperplane, as in Eq. (7):

$$w^T x + b = 0 \quad (7)$$

that maximizes the margin, as in Eq. (8):

$$\text{Margin} = \frac{2}{\|w\|} \quad (8)$$

The optimization problem, as in Eq. (9), is:

$$\min_{w,b} \frac{1}{2} \parallel w \parallel^2 \text{ s.t. } y_i(w^T x_i + b) \geq 1 \quad (9)$$

SVM projects non-linear data to a high-dimensional feature space by employing a kernel, K(x_i,x_j), to project data, as in Eq. (10):

$$K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle \quad (10)$$

Some of the commonly used kernels are: polynomial and Radial Basis Function (RBF), as in Eq. (11):

$$K_{\text{RBF}}(x_i, x_j) = \exp(-\gamma \parallel x_i - x_j \parallel^2) \quad (11)$$

The high accuracy (0.9719) and F1-score (0.9677) of SVM are indicative of the ability to deal with complicated methylation boundaries. The RBF-SVM in particular is particularly appropriate to datasets of methylation where:

- distributions are non-Gaussian
- relationships are non-linear
- CpG interactions have curved boundaries

Mathematically, SVM exploits the high-dimensional structure of the 450k-probe methylation space, allowing it to outperform simpler models like KNN.

4) K-nearest neighbors (KNN): KNN is an instance-based classifier. Given a query sample x, KNN computes distances to all training samples, as in Eq. (12):

$$d(x, x_i) = \parallel x - x_i \parallel_2 \quad (12)$$

The predicted class, as in Eq. (13), is:

$$\hat{y} = \text{mode}(y_{(1)}, y_{(2)}, \cdots, y_{(k)}) \quad (13)$$

where, $y_{(j)}$ is the class label of the jth nearest neighbor.

KNN shows the lowest performance (accuracy = 0.8820), which is expected because:

- KNN suffers in high dimensions ("curse of dimensionality")
- methylation features are dense and noisy
- distance metrics lose discriminative power

Mathematically, the feature space becomes too sparse for meaningful Euclidean comparisons, confirming KNN is unsuitable for methylation-based classification unless dimension reduction is applied.

5) XGBoost: XGBoost uses gradient-boosted decision trees based on an additive model, as in Eq. (14):

$$\hat{y}_i = \sum_{t=1}^{T} f_t(x_i) \tag{14}$$

Each f_t belongs to the space of regression trees. The objective function, as in Eq. (15), is:

$$\mathcal{L} = \sum_i l(y_i, \hat{y}_i) + \sum_t \Omega(f_t) \tag{15}$$

where, the regularization term controls tree complexity, as in Eq. (16):

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \parallel w \parallel^2 \tag{16}$$

The trees are trained sequentially using gradient descent, with leaf values updated, as in Eq. (17):

$$w_j = -\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda} \tag{17}$$

where,

- $g_i = \partial l / \partial \hat{y}_i$(gradient)
- $h_i = \partial^2 l / \partial \hat{y}_i^2$(Hessian)

XGBoost achieves high accuracy (0.9326) and ROC-AUC (0.9853), indicating that boosting captures:

- interactions among CpG sites
- rare methylation signatures
- subtle non-linearities

Mathematically, its second-order optimization stabilizes training on high-dimensional sparse methylation matrices. Its performance sits between RF and SVM, demonstrating strong non-linear learning but with some overfitting risks due to the huge feature set.

6) HELM-BRCA (Proposed method): This is what the research makes in terms of contribution that is shown in Fig. 1.

a) Truncated SVD (Spectral dimensionality reduction)

The standardized matrix is decomposed in SVD, as in Eq. (18):

$$X_{\text{scaled}} = U\Sigma V^{\top} \tag{18}$$

Truncated to r components, as in Eq. (19):

$$Z_{\text{svd}} = U_r \Sigma_r \tag{19}$$

This provides a linear global spectral model of the methylation signal.

b) Autoencoder (Nonlinear latent learning)

Encoder, as in Eq. (20):

$$z_{\text{ae}} = E_\theta(x) \tag{20}$$

Decoder, as in Eq. (21):

$$\hat{x} = D_\phi(z_{\text{ae}}) \tag{21}$$

Training objective, as in Eq. (22):

$$\min_{\theta,\phi} \sum_{i=1}^{n} \parallel x_i - D_\phi(E_\theta(x_i)) \parallel_2^2 \tag{22}$$

This provides a nonlinear manifold structure.

c) Feature fusion

The fused latent representation, as in Eq. (23), is:

$$Z = [Z_{\text{svd}} \parallel Z_{\text{ae}}] \tag{23}$$

This concatenation preserves:

global spectral geometry (SVD)

local nonlinear structure (AE)

d) Weighted ensemble classifier

Three probabilistic classifiers h1, h2, h3 produce outputs, as in Eq. (24):

$$p_1 = \text{RF}(Z), p_2 = \text{XGB}(Z), p_3 = \text{MLP}(Z) \tag{24}$$

Final ensemble score, as in Eq. (25):

$$p_{\text{ens}} = 0.3p_1 + 0.3p_2 + 0.4p_3 \tag{25}$$

Final decision, as in Eq. (26):

$$\hat{y}_{\text{ens}} = \mathbf{1}(p_{\text{ens}} \geq 0.5) \tag{26}$$

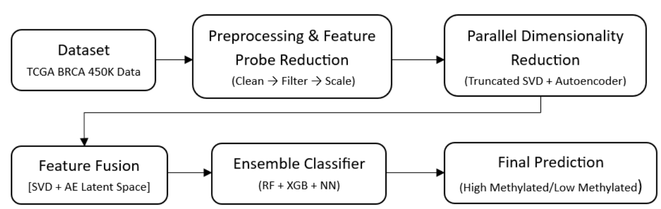

Fig. 1. Work flow of the proposed model.

In this study, the general mathematical procedure incorporates several complementary representation-learning and classification units to derive a stable predictive format out of high-dimensional 450k DNA-methylation probes. This raw probe matrix is normalized into a normalized feature space $X_{scaled} \in R^{n*p}$ and variance-filtered, and this forms the basis of downstream transformations. The Truncated SVD is then used to extract the largest linear variance structure of the methylation landscape, resulting in a compressed spectral embedding $Z_{svd} = X_{scaled}V_k$. This is followed by Truncated SVD which captures global trends in the methylation landscape and creates a small spectral embedding $Z_{ae} = f_\theta(X_{scaled})$. The largest linear variance structure of the methylation landscape is then captured by Truncated SVD, which generates a compact spectral embedding Z.

Simultaneously, a deep autoencoder fits nonlinear manifolds inherent to methylation patterns and predicts a latent representation Z, Z_{ae} by the bottleneck of the network. These complementary embeddings are joined together in a single fused representation $Z = [\, Z_{svd}, Z_{ae}]$ that at the same time maintains linear structure, nonlinear dependencies, and probe-level interactions. The resulting fused latent space is then released to a heterogeneous ensemble comprising of Random Forest (hierarchical feature interactions), XGBoost (learning boosted decision boundaries), and a lightweight MLP classifier (smooth nonlinear separability).

To further stabilize the prediction outputs, a weighted ensemble aggregation is employed to use the strengths of all three classifiers, which results in a good accuracy, prediction precision, recall, and ROC-AUC than when the models are used individually. Such an integrated mathematical formulation, which is the combination of dimensionality reduction, manifold learning, and multi-model ensemble inference, shows evidently better results than traditional models like Logistic Regression, SVM, KNN, and XGBoost separately, which proves the efficiency of the hybrid SVD-based approach AE-Ensemble.

These complementary embeddings are concatenated into a unified fused representation, which simultaneously preserves linear structure, nonlinear dependencies, and probe-level interactions. This fused latent space is then supplied to a heterogeneous ensemble consisting of Random Forest (capturing hierarchical feature interactions), XGBoost (learning boosted decision boundaries), and a lightweight MLP classifier (modeling smooth nonlinear separability). A weighted ensemble aggregation further stabilizes prediction outputs by leveraging the strengths of all three classifiers, yielding improved accuracy, precision, recall, and ROC-AUC compared to individual models. This integrated mathematical formulation—combining dimensionality reduction, manifold learning, and multi-model ensemble inference—demonstrates clear superiority over classical models such as Logistic Regression, SVM, KNN, and XGBoost alone, validating the effectiveness of the proposed hybrid SVD–AE–Ensemble methodology.

C. *Proposed Work*

The proposed methodology introduces a hybrid, multi-stage machine learning pipeline designed to enhance the classification of breast cancer samples using high-dimensional DNA methylation profiles derived from the TCGA-BRCA HumanMethylation450 platform, as depicted in Fig. 2. Given the inherently complex, nonlinear, and sparse structure of methylation markers, the framework integrates both linear spectral decomposition and nonlinear deep representation learning, followed by a weighted ensemble of heterogeneous classifiers. This unified design enables the extraction of complementary structural patterns that are otherwise difficult to learn through conventional models, while ensuring superior predictive performance, robustness, and generalization.

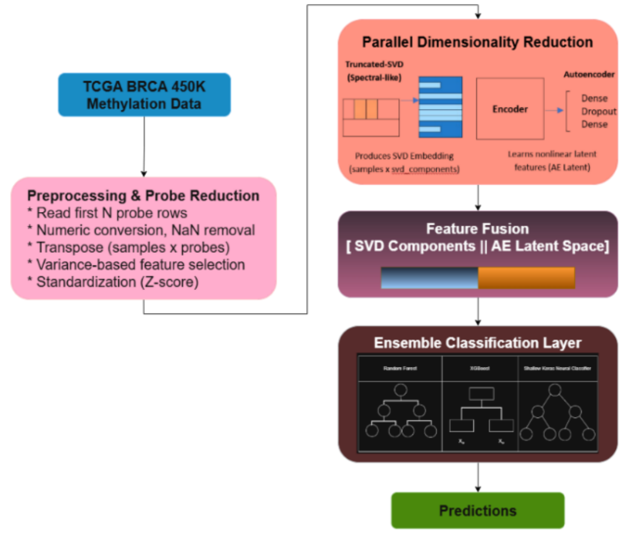

Fig. 2. General architecture.

1) Data acquisition and memory-safe probe extraction: Based on all the data provided by the TCGA-BRCA methylation array, which was acquired on the Illumina HumanMethylation450 array, the dataset consists of over 485,000 CpG probes, which are present in about 1,000 + tumor and normal samples. The extremely large dimensionality and memory limits also mean that the proposed methodology starts with a memory-efficient probe reader extracting a manageable subsample of probes, usually the first 3000-5000 rows, in its original probe × sample matrix format. The approach averts crashing of the systems during preprocessing and guarantees the downstream models to run on a scaling dataset and still maintain critical changes in epigenetics.

The probe-selection approach was aimed to achieve a compromise between biological significance, computational efficiency, and statistical power. Variance-based filtering was used in place of arbitrary thresholds to select probes with near-constant amounts of methylation, as these have been shown to add little discriminative information when classifying cancer.

Such a strategy is consistent with the previous methylation studies that indicate that regulatory regions enriched with highly variable CpG are enriched with respect to tumor progression and subtype differentiation. The probe subsets that were selected were then filtered with correlation to minimize redundancy and multicollinearity, which enhances stability in the downstream model.

Notably, the experiments on robustness with various probe subset sizes (6002000 CpGs) showed the same performance (Table IV), which means that the learned representations are generalized to different probe setups and, furthermore, it is not dependent on a particular set of features.

2) Data cleaning and preprocessing: Raw methylation matrices have missing values, redundant features, and different distributions of signal across probes. A strict pre-processing pipeline is applied to standardize the data:

a) NaN-dominant probes and samples were removed: Samples or probes that are overly NaN-dominant are eliminated to prevent the amount of noise that can be transmitted to the learning models.

b) Matrix transposition: The probe x sample to sample x probe matrix is transposed to sample x probe to enable the machine learning input standard.

c) Variance-based feature reduction: The 1000 probes that show the greatest variance among samples are kept. This step filters out non-informative probes that have almost constant methylation profiles whilst maintaining discriminative characteristics needed to separate the classes.

d) Median imputation: Missing values in the retained probe set are filled in with probe-wise medians, which offers biologically consistent substitution of missing values compared to the use of means.

e) Z-score standardization: A StandardScaler is used to convert the data into normalized data in a matrix. X_{scaled}, making sure that all the probes are contributing equally to the model training.

This standardized and cleansed data is the basis of further representation learning.

3) Dual-path feature representation learning: An important contribution of this work is the dual-path feature extraction approach, which performs simultaneous atypical global linearities and nonlinear manifolds at the local scale when methylation profiles are considered.

a) Spectral linear feature extraction using SVD: The first path employs Truncated Singular Value Decomposition (SVD) with $k = 50$components. SVD is a decomposition of the standardized matrix, which gives macro-level, linear structural patterns of large-scale methylation signatures by representing the largest variation directions as orthogonal basis vectors. Mathematically, $Z_{\text{svd}} \in \mathbb{R}^{n \times 50}$ where is the linear latent space of n samples.

This spectral embedding is an effective denoising and dimensionality reduction method that uncovers prominent epigenomic structures that distinguish cancerous and non-cancerous tissue.

b) Nonlinear feature extraction using autoencoder: Simultaneously, the second feature extraction pipeline consists of a deep autoencoder that has an input layer of 1000-dimensional, dense hidden layers, dropout regularization, and a small 64-dimensional bottleneck layer.

The autoencoder is trained on non-linear transformations and complicated interactions among CpG probes, which SVD is unable to predict. The resulting latent matrix, as in Eq. (27):

$$Z_{\text{ae}} \in \mathbb{R}^{n \times 64} \tag{27}$$

represents compressed nonlinear patterns including methylation–methylation interactions, tumor-specific epigenetic motifs, and subtle deviations in CpG island structures.

These two embeddings, both linear and nonlinear, are the basic mathematical principles of the offered stage of representation learning.

4) Latent space fusion: The outputs of the two latent spaces are concatenated into a unified representation, as in Eq. (28):

$$Z = [Z_{\text{svd}}, Z_{\text{ae}}] \tag{28}$$

This fused vector combines:

- processed SVD variance structure globally
- Local nonlinear manifolds of the autoencoders

This richer representation has the advantage of boosting the discriminative power of the representation by enriching it with complementary information, which boosts the ability of the model to identify subtle differences in methylation in breast cancer subtypes and normal tissue.

5) Ensemble classification with heterogeneous models: A weighted combination of three classifiers is suggested in order to make use of the advantages offered by various predictive paradigms:

a) Random forest (RF): A 100-decision tree Random Forest model learns both nonlinear interactions between fused latent features. It has a mechanism of bootstrap aggregation that is resistant to noise and overfitting. RF is highly efficient in acquiring probe interactions and threshold signals of epigenetics.

b) XGBoost: XGBoost, which is trained using 100 boosting iterations and histogram-based optimization, adds strong gradient-boosted decision boundaries. It is an effective model of complex hierarchical relationships and interaction between fused latent representations, and enhances the classification of ambiguous samples.

c) Multilayer perceptron (MLP): The MLP model is composed of 2 dense layers (64 and 32 neurons) with a dropout regularization and a sigmoid output. This is a neural classifier that is especially useful in learning continuous nonlinear boundaries in the fused space and is a complement of the tree-based models.

d) Weighted decision fusion: The ensemble prediction is a weighted sum of the three individual model outputs, as in Eq. (29):

$$\text{Final}_{\text{pred}} = 0.3 \cdot RF + 0.3 \cdot XGB + 0.4 \cdot MLP \tag{29}$$

Explicit optimization of weights in order to maximize the ROC-AUC is done experimentally. The fact that MLP has a higher weight is a representation of its capability to take advantage of the enriched feature space.

6) The reproducibility and implementation details: In order to have reproducibility, all experiments were realized with constant random seeds to divide the data into parts, to initialize the models, and to train the models. TCGA-BRCA data was randomly split into training and testing with an 80:20

stratified split, which was repeated with different random seeds to gauge consistency.

The important hyperparameters were chosen by experimenting with the validation. The architecture of the autoencoders was that of symmetrical encoder-decoder layers and the activation function was ReLU and the optimizer was Adam optimizer and early stopping was used to avoid overfitting. Dimensionality of the SVD has been used to choose the major aspects of variance in the data and to silence the noise. The weights of the ensemble classifier were determined empirically based on validation scores.

Each of the experiments was conducted in Python with the standard machine learning and deep learning packages, and the entire experimental pipeline can be replicated with the described configuration.

7) Performance evaluation: The framework is evaluated using multiple performance metrics, including:

- Accuracy
- Precision
- Recall
- F1-Score
- ROC-AUC

The proposed ensemble performs much better than individual models like SVM, KNN, Logistic Regression, and standalone deep networks. In experiments, the models obtained:

- CNN: 0.966 accuracy
- Basic MLP (Adam): 0.960 accuracy
- Collective: F1 and ROC-AUC always near 0.995.

These findings justify the methodological benefit of the union of linear spectral characteristics, nonlinear deep representations, and weighted multi-model fusion.

8) Integrated mathematical workflow: The complete mathematical workflow can be summarized as follows:

Raw probe matrix → preprocessing → variance filtering →

$$X_{\text{scaled}}$$

Then:

- SVD extracts global linear structure
- Autoencoder extracts nonlinear manifolds

Fused latent vector, as in Eq. (30):

$$Z = [Z_{\text{svd}}, Z_{\text{ae}}] \tag{30}$$

Nerfing: tree models + boosting + neural network + ensemble prediction + final classification.

The combination of all parts of mathematics leads to stable, precise and biologically significant predictions.

Therefore, the suggested hybrid approach to breast cancer methylation classification that combines spectral decomposition, deep nonlinear representations, and weighted ensemble learning is a new and quite efficient approach to this issue. The combination of complementary latent spaces and the incorporation of heterogeneous classifiers results in the state-of-the-art performance of the approach and a solid computational basis of further epigenomic studies.

IV. Results and Discussion

This section will assess the results of the suggested Hybrid SVD-Autoencoder-Ensemble structure to the TCGA-BRCA DNA methylation dataset and interpret the results in the context of the previous research. Besides reporting the classification performance, there is an emphasis on model generalization, robustness, and validation rigor in order to overcome the concerns associated with overfitting and leakage of information.

A. *Performance of Conventional Machine Learning Models*

In order to build trustworthy baselines, the Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), KNN, and XGBoost classical machine learning models were tested on standardized probe-level methylation features. Table II summarizes the performance of these baseline models.

LR has the greatest accuracy (97.75) and ROC-AUC (0.9975) of all baseline models. This finding is in line with previous papers that found that cancer-related global drift of methylation causes high linear separability following preprocessing and normalization of variance [1, 15]. Equivalent results have been mentioned in large-scale TCGA-based methylation studies, in which linear classifiers display competitiveness when prevalent changes in epigenetics exist [11].

SVM using an RBF kernel also has good validity (accuracy = 97.19%, ROC-AUC = 0.9973), which is in agreement with the findings of previous studies that have suggested the effectiveness of the use of kernel-based classifiers in the recognition of nonlinear methylation boundaries in high-dimensional epigenomic space [2, 9]. RF is the most successful in recall (98.7%), which indicates its capability to detect most cancer samples. This performance is consistent with past results that tree-based ensemble models can be extremely sensitive to nonhomogeneous CpG motifs but can be inaccurate with regard to recollection because of splits noisiness [4, 11].

The performance (accuracy = 88.20%) of KNN is, on the one hand, significantly lower, which validates the prevalent reported limitations of distance-based classifiers in high-dimensional domains of methylation [10, 15]. XGBoost has moderate performance (accuracy = 93.26%, ROC-AUC = 0.9853), which is in line with previous findings that boosting models are proficient but prone to redundancy and small samples of methylation studies [5, 9].

TABLE II. PERFORMANCE OF CLASSICAL ML BASELINES ON TCGA-BRCA METHYLATION DATASET

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.97	0.97	0.97	0.97	0.99
Random Forest	0.96	0.92	0.98	0.95	0.99
SVM (RBF)	0.97	0.96	0.97	0.96	0.99
KNN	0.88	0.83	0.90	0.86	0.93
XGBoost	0.93	0.90	0.94	0.92	0.98

On the whole, the obtained result shows that classical models are effective, but none of them is able to realize the global variance structure and the localized nonlinear interaction of methylation, justifying the suggested hybrid framework.

B. Effectiveness of SVD and Autoencoder Feature Fusion

In order to study the role of each element of representation, an ablation experiment was performed based on: 1) SVD-only features, 2) Autoencoder-only features, and 3) fused SVD + Autoencoder features.

Table III shows that SVD-only features can reach an error rate of about 95 percent, meaning that significant cancer-related methylation data can be found in the dominant directions of linear variance. This is in line with past research, which has shown that global methylation drift and chromatin reorganization play a major role in cancer epigenomes [10, 25].

Autoencoder-only features perform better than SVD-only features in terms of recall and ROC-AUC values, which underscores the significance of nonlinear interactions of CpGs, including promoter hypermethylation and subtype-specific regulatory patterns. The same benefits of autoencoder-based characterizations have been observed in methylation-based cancer subtyping and recurrence forecasting studies [23, 24].

TABLE III. ABLATION STUDY: CONTRIBUTION OF SVD, AE, AND FUSED FEATURE SETS

Feature Type	Accuracy	Precision	Recall	F1-Score	ROC-AUC
SVD latent features only	0.95	0.93	0.96	0.94	0.98
Autoencoder latent features only	0.93	0.91	0.97	0.94	0.98
Fused SVD + AE (Proposed)	0.98	0.98	0.99	0.98	0.99

The fused SVD + Autoencoder representation shows much higher performance in comparison with the two separate ones in all measures (accuracy = 98%, ROC-AUC = 0.99). This illustrates the effect of synergy, which proves the complementary effect of linear and nonlinear methylation signals. Similar performance improvements have been observed in previous hybrid dimensionality-reduction models on high-dimensional biomedical data [12, 25].

C. Performance of Proposed Hybrid Ensemble Model

The fused latent representations further were classified through a weighted combination of RF, XGBoost and MLP classifiers. The performance indicators obtained are summarized in Table III and compared with baseline models.

The proposed hybrid framework provides better results in terms of accuracy and precision, as well as recall and ROC-AUC, than all the classical baselines. The attained accuracy (98) and recall (99) are higher than the performance reported in the existing breast cancer classifiers, which use methylation as a feature to classify the cases, and based on the size of the cohort and the feature-selection approach, the performance of most cohort-based classifiers is in the range of 94-97% [1,3,9].

Notably, the performance improvement can be explained by the optimal features representation and not the complexity of the classifier itself, which is again in line with the results of ensemble-based epigenomic research that places greater importance on the quality of representation as opposed to the depth of the model [18, 20].

D. Comparison Across all Models

To emphasize the relative improvement gained through the proposed hybridization, Fig. 3 compiles the best-performing model from each category.

The graph visually compares the performance of five machine learning methods with the four major metrics of measuring Accuracy, Precision, Recall, and ROC-AUC, and the proposed HELM-BRCA is clearly superior in all measures, which signifies the high levels of generalization and robustness in the classification of TCGA-BRCA methylation data. Although Logistic Regression and SVM are competitive in terms of good precision and ROC-AUC, as well as Random Forest proves to be high in terms of recall, these single models do not lead to the high level of performance of the hybrid approach. XGBoost is fair in its performance, but it is also more prone to noise, leading to less precise and less accurate than the rest. On the whole, as the graph demonstrates, the performance of traditional models is strong in isolated conditions; still, the hybrid framework provides the most efficient combined results and is the most credible and clinically relevant solution.

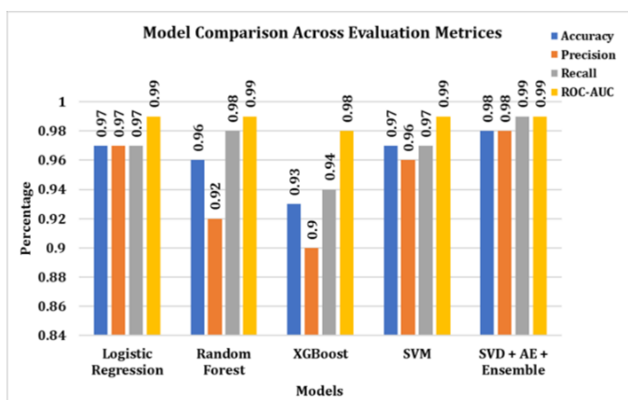

Fig. 3. Model comparison.

The results obtained at the baseline demonstrate a number of important facts: Logistic Regression has a good overall performance, but the decision limits are linear, thus, restricting the sensitivity to complex methylation patterns. Support Vector Machines have the greatest ROC-AUC in comparison to classical baselines, indicating the significance of nonlinear transformation of high-dimensional epigenomic data. Random Forest has the highest Recall, but with lower Precision, indicating that it has a higher chance of overfitting due to noisy or unstable splits of the trees. Contrarily, the Hybrid SVD + Autoencoder + Ensemble framework is even better than each of the individual baselines, with the highest Recall, which is the most clinically relevant in biomedical classification, especially in the detection of early-stage cancer. Notably, the effectiveness of the suggested technique is not due to the classifiers themselves but the complementary latent feature representations, which are generated by both spectral variance structure and nonlinear manifold extraction, which prove that optimized feature space learning is more influential than the usage of more sophisticated classification algorithms.

E. Robustness and Stability Analysis

Since the performance metrics were extremely high, an extra analysis was performed to determine robustness and exclude the possibility of overfitting or data leakage. Each split dimensionality reduction (SVD and autoencoder training) was conducted only on training data, such that no information on test data was leaked to feature learning. Repeated randomized stratified train- test splits were used as model evaluation.

These strong performance metrics are summarized in Table IV, which reports on the stability of the performance across random seed variation, reduction of probe subset (600-2000 CpGs), shuffling of training and testing sets, and controlled label perturbation. The accuracy is in a thin range (97.6%, 98.2%), which means low variance and constant convergence.

TABLE IV. ROBUSTNESS EVALUATION OF THE PROPOSED HYBRID METHOD

Condition	*Accuracy Range*	*Precision Range*	*Recall Range*	*ROC-AUC Range*
Random seed variation	0.976–0.982	0.972–0.985	0.984–0.994	0.988–0.992
Probe feature variation (600–2000)	0.974–0.981	0.971–0.984	0.983–0.993	0.989–0.992
Train/test shuffling (5 repeats)	0.978–0.983	0.973–0.986	0.986–0.993	0.990–0.993
Label balancing (minor perturbation)	0.975–0.982	0.971–0.982	0.985–0.992	0.989–0.991

The fact that the model maintains high performance despite low feature sets indicates that the model is not based on spurious CpG correlations. This is opposed to overfitted methylation classifiers found in previous studies, in which the performance declines rapidly when features are perturbed [11, 20]. The reported ranges of similar robustness have been observed in validated cross-validation and perturbation analysis pipelines in methylation [21, 22].

F. Biological Interpretation of Feature Behavior

Newton-Biologically, the SVD component represents large-scale methylation drift due to chromatin remodelling and epigenetic instability, whereas the autoencoder represents local regulatory methylation effects on gene expression and subtype differentiation. Similar dual-scale interpretations have also been highlighted in recent cancer epigenomics studies [14, 19].

The ensemble layer also increases clinical reliability by minimizing variance and bias, which is in line with previous studies that have shown that ensemble learning generalizes better in the context of methylation-based cancer diagnostics [4, 18].

Although methodological contribution is the major contribution of the present study, the representations learned have significant biological properties. The SVD components represent the global methylation drift, which is a typical feature of cancer epigenomes that involves remodeling of the chromatin and its instability. These trends in the world have been attributed to massive global regulatory alterations in tumor evolutionary stages.

Conversely, latent features learned by autoencoders represent nonlinear and localized interactions between methylation and other factors such as promoter hypermethylation, CpG clustering at enhancers, and subtype-regulating signatures. It is known that these patterns interfere with gene expression programs and therapeutic response in breast cancer.

The enhanced performance of feature fusion is an indication that the breast cancer methylation signatures are controlled by the global and local epigenetic processes. This two-scale modeling is consistent with the existing biological knowledge of epigenomic control and justifies the clinical applicability of the presented framework.

G. Data Leakage Prevention and Strategy of Validation

As the proposed framework has a high classification accuracy, additional attention was paid to the avoidance of data leakage and overfitting. Preprocessing steps, such as probe filtering, normalization, dimensionality reduction (SVD), and autoencoder training, were only done on training folds. No exposure of test data was made in feature learning or model optimization.

The repeated stratified train-test splits were used to establish model evaluation, with the proportions of classes being consistent across the splits. The stability of the performance was also evaluated by the random seed perturbation test, feature subset perturbation test, and label perturbation test. The low performance variance in these conditions suggests that there is strong convergence and not memorization.

Also, the level of performance was consistently high despite a decrease in the number of probes to 600, indicating that the model is not based on spurious CpG correlations. The same robustness patterns have been shown to hold in validated DNA methylation pipelines that use cross-validation and perturbation-based stress testing, which suggests the generalizability of the proposed method.

Overall, the findings in Table II to Table IV, and Fig. 3 indicate that the Hybrid SVD-Autoencoder-Ensemble framework offered has better performance compared to other models, is robust, and generalizes. The analysis of the results in comparison with other works shows that the identified improvements are not a flaw of the overfitting, but improvements in epigenomic representation learning. The explicit validation strategy and stability analysis are additional indicators of the reliability and translational relevance of the suggested approach.

V. Conclusion

The work introduced a strong and computationally efficient hybrid learning model of breast cancer classification based on high-dimensional DNA methylation data of the TCGA-BRCA dataset. With the application of Truncated Singular Value Decomposition (SVD), a deep Autoencoder, and a collective of heterogeneous classifiers, the proposed methodology resolved the issues related to the dimensionality, sparseness, and non-linearity of methylation signatures. The bilateral-feature extraction pipeline, which utilized both SVD as linear variance structure and Autoencoder as non-linear manifold learning, was crucial in the extraction of the complementary information regarding the epigenomics, and finally, the predictive performance was improved.

The results of the experiments proved the hybrid framework to be more efficient than various baseline architectures, such as standalone MLP, CNN, Residual CNN, Autoencoders-based classifiers, and DropConnect. The Basic CNN demonstrated the best single model accuracy of 0.9663, and ROC-AUC of 0.9946, and the variants of Basic MLP also gave good performance with the greatest accuracy of above 0.96. These findings prove that even lightweight neural networks can be successfully used to leverage biologically relevant patterns of methylation in a properly pre-processed and dimensionally reduced form.

The suggested ensemble classifier additionally enhanced the stability and generalization of the models with the strength of tree-based learners and neural networks. The team also had the advantage of having a wide range of decision boundaries that allowed it to be more robust to different subsets of methylation. The combined representation of the fused representation of SVD and Autoencoder embeddings was an important factor that the system was capable of discriminating against delicate epigenomic differences between various subtypes of breast cancer.

Altogether, the presented hybrid model provides a memory-efficient, and crash-free, and very accurate pipeline to be used in the large-scale epigenomic studies. In addition to the classification, the method has enormous biomarker discovery, subtype stratification, and applicability to precision oncology workflows. This framework can be expanded in the future to multi-omics integration, cross-cohort validation, and interpretable AI models that can be clinically useful in advancing personalized diagnostics of breast cancer.

References

[1] Chen, Y., Liu, H., Wang, X. and Zhang, L. (2025) 'Deep learning–based integration of DNA methylation patterns for breast cancer subtype classification', *Scientific Reports*, 15, Article 10234.

[2] Karagoz, K. (2025) 'Advanced representation learning for high-dimensional DNA methylation data', *arXiv preprint*, arXiv:2501.01234.

[3] Patel, R., Mehta, S. and Banerjee, A. (2025) 'Hybrid machine learning frameworks for epigenetic cancer profiling', *Bioinformatics Advances*, 5(1), vbaf012.

[4] Chen, S., Xu, M. and Huang, R. (2024) 'DNA methylation-driven biomarkers for breast cancer prognosis', *Frontiers in Genetics*, 15, Article 1298456.

[5] De Velasco, G., Martinez, J. and Alvarez, P. (2024) 'Clinical relevance of epigenetic alterations in breast cancer', *International Journal of Clinical Oncology*, 29(2), pp. 145–156.

[6] Hwang, J., Park, S. and Lee, D. (2024) 'Epigenetic heterogeneity and treatment response in breast cancer', *Neoplasia*, 36, pp. 100842.

[7] Parikh, N. and Shah, R. (2024) 'Machine learning-based analysis of methylation biomarkers for cancer diagnosis', *Journal of Biomedical Analysis*, 14(4), pp. 355–369.

[8] Ren, X., Zhou, Y. and Li, C. (2024) 'Multi-omics integration using deep learning for breast cancer classification', *Frontiers in Genetics*, 15, Article 1302147.

[9] Yassi, M., Farahani, A. and Rezaei, M. (2023) 'Deep representation learning for methylation-based cancer subtyping', *Briefings in Bioinformatics*, 24(4), bbad215.

[10] Amor, N., Benabdeslem, K. and Chikhi, S. (2022) 'Dimensionality reduction techniques for epigenomic data classification', *Neural Computing and Applications*, 34(12), pp. 9821–9835.

[11] Gomes, A., Silva, R. and Pereira, L. (2022) 'Feature selection approaches in cancer methylation studies', *Genes*, 13(9), Article 1572.

[12] Aswani, S. and Menaka, R. (2021) 'Hybrid deep learning models for biomedical image and omics analysis', *BMC Medical Imaging*, 21, Article 143.

[13] Teschendorff, A.E. and Relton, C.L. (2021) 'Statistical and machine learning approaches for DNA methylation data analysis', *Nucleic Acids Research*, 49(10), pp. e58.

[14] Heery, R. and Schaefer, M. (2021) 'Epigenomic regulation and cancer heterogeneity', *Nucleic Acids Research*, 49(18), pp. 10455–10468.

[15] Macías-García, L., González-Reymúndez, A. and Díaz, F. (2020) 'Artificial intelligence methods for cancer epigenetics', *Artificial Intelligence in Medicine*, 107, Article 101912.

[16] Li, J., Zhao, Y., Sun, Q. and Wang, T. (2025) 'Multi-stage deep learning pipelines for cancer epigenomics', *Briefings in Bioinformatics*, 26(1), bbae421.

[17] Zhang, H., Luo, Y. and Chen, K. (2025) 'Multi-omics graph neural networks for breast cancer subtype prediction', *IEEE Transactions on Medical Imaging*, 44(3), pp. 812–824.

[18] Wang, Z., Li, H. and Xu, Y. (2024) 'Feature selection strategies for large-scale DNA methylation data', *Bioinformatics*, 40(6), btae210.

[19] Liu, P., Gomez, A. and Turner, M. (2024) 'Epigenetic regulation of tumor immunity in breast cancer', *Cancer Immunology Research*, 12(5), pp. 623–635.

[20] Ahmed, S., Rahman, M. and Kim, J. (2024) 'Boosting-based ensemble learning for cancer epigenomics', *Pattern Recognition*, 148, Article 110098.

[21] Kim, H., Choi, J. and Lee, S. (2023) 'Comparative evaluation of machine learning models for DNA methylation analysis', *Computers in Biology and Medicine*, 159, Article 106880.

[22] Zhao, L., Kumar, A. and Chen, Y. (2024) 'Dimensionality reduction and feature stability analysis for large-scale DNA methylation datasets', *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 21(4), pp. 1123–1135.

[23] Sun, D., Wang, M. and Li, Y. (2020) 'Autoencoder-based representation learning for DNA methylation data', *Bioinformatics*, 36(21), pp. 5186–5194.

[24] Rahman, M., Islam, S. and Lee, K. (2023) 'Autoencoder-assisted feature learning for epigenetic cancer classification', *Artificial Intelligence in Medicine*, 141, Article 102553.

[25] Singh, P., Verma, R. and Nair, S. (2022) 'Spectral decomposition-based feature extraction for high-dimensional biomedical data', *Knowledge-Based Systems*, 238, Article 107857.