Journal of Theoretical and Applied Information Technology
31st March 2026. Vol.104. No.6

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

HCRHE-NET: HIGH-CONFIDENCE RESIDUAL HYBRID ENSEMBLE NETWORK FOR BREAST CANCER DETECTION FROM TCGA-BRCA DNA METHYLATION DATA

HEMALATHA D[1], N. GOMATHI[2]

[1]Research Scholar, Department of Computer Sciences & Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, India
[2] Professor, Department of Computer Sciences & Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, India
E-mail: [1]hema24294@gmail.com, [2]gomathin@veltech.edu.in

ABSTRACT

Due to noise, redundancy and uncertainty with respect to predictive confidence, proper classification of high-dimensional biological data remains a significant challenge. To overcome these shortcomings, this paper proposes a High-Confidence Residual Hybrid Ensemble (HCRHE) Network, which is a composite of residual learning, deep neural modelling, and confidence-conscious decision fusion to classify diseases with high confidence. The Cancer Genome Atlas (TCGA) contains large-scale data on DNA methylation that are susceptible to overfitting and unstable forecasts by using conventional deep learning models to evaluate the proposed methodology. To learn latent patterns due to reconstruction, HC-RHE architecture consists of a primary-prediction base multilayer perceptron (MLP), together with a residual learning path that is constructed using an autoencoder and residual MLP. A new machine called a confidence-based fusion technique allows the dynamical weighting of the base and residual prediction in terms of model certainty and makes adaptive decision-making. Also, forecasts with high confidence margins are retained and a high-confidence filtering process used, which maximizes reliability with minimal coverage loss. An accuracy-optimized threshold selection strategy is also provided in order to enhance the performance of classification further. Vast comparative experiments are conducted with the state-of-the-art deep learning baselines, including CNN, autoencoder-based classifiers, Dense DropConnect, residual CNN frameworks, and Basic MLP (Adam and SGD). The proposed HC-RHE has a much better accuracy at 98.7 compared to all other methods. These results prove success of confidence-aware residual fusion as they reflect continuous improvement over the best baseline CNN model. The proposed framework is, all in all, exceptionally promising to the field of clinical decision-support systems and gives a credible, intuitive, and high-confidence classification paradigm of high-dimensional biomedical data.

Keywords: *High-Confidence Learning, Residual Hybrid Ensemble, Deep Neural Networks, Autoencoder, DNA Methylation, Biomedical Classification, TCGA, Confidence-Aware Fusion, Cancer Prediction*

1. INTRODUCTION

Gene expression is the fundamental process of biology through which genetic information in the form of DNA is translated and processed into functional molecules that regulate cell shapes, behaviour and fate. Gene expression in normal growth, development, and tissue homeostasis requires an exact regulation of gene expression. These tightly controlled systems often become disturbed to cause pathological disorders, in the most obvious case, cancer. Dysregulation in apoptosis, cell-cycle control, and DNA repair pathways through aberrant activation of oncogenes or inhibition of tumour suppressor genes such as TP53 and PTEN may facilitate tumour start and progression. Thus, the accurate characterisation and prediction of gene expression patterns are fundamental to understanding the causes of disease, as well as enhancing the detection of disease in its initial stages, prognosis and designing of specific therapies.

Among the various levels of regulation of gene expression, it is the epigenetic processes that are especially relevant. The changes in gene operation that are hereditary and which do not involve alteration of the underlying DNA code are termed as epigenetics. The primary epigenetic processes that interact to regulate the activity of transcription are

DNA methylation, histone modifications, non-coding RNAs, and chromatin accessibility. The most studied change in the epigenetic activity is the DNA methylation or the addition of a methyl group to the cytosine nucleotide, which is usually at CpG dinucleotide. DNA methyltransferases (DNMTs) including as DNMT1, DNMT3A and DNMT3B catalyses this process, which is vital to X-chromosome inactivation, genomic imprinting, embryonic development and the maintenance of genomic stability.

Among other control levels that determine the expression of genes, the most important ones are the epigenetic processes. Epigenetics are the changes in the action of the genes which are not inheritable and are not linked to the modification of the coding DNA. DNA methylation, histone modifications, non-coding RNAs, and chromatin accessibility are the main epigenetic processes that communicate to control transcriptional activity. The most examined epigenetic change is DNA methylation which is the addition of a methyl group on the cytosine base that is usually in CpG dinucleotides. This process is catalysed by the DNA methyltransferases (DNMTs) such as DNMT1, DNMT3A and DNMT3B, and is needed to inactivate the X chromosomes for development as well as the ability to maintain genomic stability, genomic imprinting and the development of the embryonic stage.

The current advancements in high-throughput technologies such as Illumina HumanMethylation450 and EPIC arrays and the next-generation sequencing platforms have enabled genome-wide profiling of hundreds of thousands of CpG sites per sample. Large-scale projects such as The Cancer Genome Atlas (TCGA) have generated comprehensive DNA methylation data sets of many cancer types, such as breast cancer (TCGA-BRCA). These datasets provide tired-of-never-before opportunities to explore epigenetic regulation, yet, they also come with employing significant analytical challenges. The main peculiarities of DNA methylation data are high dimensions, sparsity, noises, and often a low number of samples relative to the number of measured attributes. In this case, the complex, nonlinearity of relationship between non-linear pattern of CpG methylation and biological phenotype cannot be understood by the conventional statistical technique.

Machine learning (ML) techniques enable the creation of data-driven modelling of high-dimensional biological data to become a promising tool in addressing those problems. In classical machine learning pipelines, a such one is often followed by handcrafted feature extraction, then regression or classification using machine learning methods such as random forests, logistic regression, or support vectors. These approaches have been found to be useful in cancer epigenetics, but they are often found to be unable to scale between datasets and often depend on manually-designed features. Whereas the feature selection methods can help alleviate the dimensionality issues, they can remove the biologically important data especially in cases where different CpG sites are so important in interaction.

Due to its capability to generate hierarchical feature representations without the input of humans in the case of raw data, the specialised branch of machine learning known as deep learning (DL) has attracted an increasing number of interested parties. Examples of effective deep neural networks in the context of bioinformatics include multilayer perceptrons (MLPs) and convolutional neural networks (CNNs), recurrent neural networks (RNNs), transformer-based networks, and have been successfully applied to protein tumour subtypling, tissue-of-origin classification, survival prognostication, and protein discovery as biomarkers. Deep learning methods are particularly well suited to discover the complex regulatory interactions underlying DNA methylation patterns since they are nonlinear and high-order interactions.

Recent studies have indicated the extent to which deep learning models methylation-based regulation of genes. DeepMethyGene enhanced prediction accuracy on TCGA breast cancer data by adding convolutional kernels of various sizes in addition to residual learning to forecast the gene expression using DNA methylation profiles [1]. This study highlighted the importance of gathering both local and long-range effects of methylation around locus genes. Similarly, DeepPGD integrated CNN and BiLSTM blocks with attention systems to model both the sequential and spatial relationships in the methylation data jointly, and the classification of methylated and unmethylated DNA sites was better [4]. These designs are characterized by the usefulness of hybrid deep learning models, which combine complementing feature extraction techniques.

Ensemble learning has also been explored to improve resilience and generalisation more than single-model architectures. The ensemble deep network framework of EDNTOM, including CNN, RNN, and BERT-based elements and a novel attention-based weighting scheme, was found to be more successful in nanopore methylation detection tasks compared to traditional ensemble voting ones [3]. Ensemble methods can reduce model variation and reduce the biases present in individual learners

through the use of multiple representations of features. Nevertheless, any investigation on ensemble techniques in the analysis of methylation is still underrepresented, particularly in large cancer samples.

Transformer-based and foundation models have also been used recently in methylation research. To obtain state-of-the-art accuracy in predicting methylation sites, techniques such as DeepMethylation were used which utilized GloVe embeddings and transformer encoders to learn both local and global features of the DNA sub-sequences [6]. Large-scale pretraining methods, including models such as MethylNet [10], further demonstrated the potential of automated and modular deep learning systems to construct embeddings, generate predictions, and discover latent heterogeneity in DNA methylation data with minimal human intervention. These developments imply that learners using methylation will become task-specific and more generalisable.

Alongside such developments, there are still several issues. The CNN-based models might also have problems with long-range relationships in one-dimensional genomic sequences, and even though RNNs are applicable to sequential modelling, distance interactions and computational efficiency often become problematic. Transformer models are used to address some of these problems, but tend to require large amounts of training data and processing power. In addition, many of the existing approaches do not focus much on trustworthy classification based on disease levels and rather focus on either estimating the location of methylation sites or estimating gene expression. Issues such as class imbalance, overfitting, and poor calibration of confidence make clinical translation a harder task.

The DNA methylation has proved to have great potential in advancing accuracy in diagnosis and prognosis of breast cancer. In studies involving the use of methylation marks, there is evidence that effective cancer prediction and classification of subtypes is possible including the analysis of paracancerous tissues that exhibit early molecular alterations before overt tumours develop [9]. Beside emphasizing the necessity of more reliable and scalable analytical frameworks, systematic reviews have also pointed to the growing applicability of machine learning to finding predictive methylation biomarkers in cancers [7]. Though it has been established that feature selection with deep learning can be used to improve the performance of prediction, and overcome the issue of computational constraints, the optimal integration methods remain a subject of discussion [8].

These results demonstrate that there is an urgent necessity of advanced modelling frameworks that integrate the strengths of residual structures, the use of ensembles, deep learning without compromising scalability, interpretability and resilience. Confidence-aware decision processes are particularly important in biological applications where the consequences of false positive and false negative outcomes are particularly severe to the therapeutic process. Models that have been successful in delivering confidence estimates that are reliable and strike an effective balance between sensitivity and specificity are likely to be used in real-life scenario applications.

To address the problems, a high-confidence residual hybrid ensemble framework of breast cancer detection is proposed in this work using DNA methylation data of the TCGA-BRCA cohort. The proposed approach aims to achieve higher classification accuracy and stability compared with the existing deep learning models through combining residual learning to extract stable deep features, heterogeneous base learners to identify complementary methylation patterns, and confidence-based ensemble aggregation. The efficiency of the proposed architecture on popular metrics of performance is evidenced by an extensive experimental analysis of the architecture against a variety of reference frameworks.

In general, our contribution to the growing body of literature at the interface of artificial intelligence and cancer epigenetics is to provide a robust, ensemble-based deep learning model that can detect methylation-induced breast cancer. The proposed system is built upon previous studies of deep methylation modelling [11], offering a high-performing, scalable system with tremendous potential of accuracy oncology applications and clinical decision support.

2. RELATED WORKS

Since DNA methylation is critical in the regulation of gene expression without altering the underlying DNA sequence, it has been one of the most studied epigenetics mechanisms. The diagnosis and prognosis of cancer would be well indicated by the aberrant patterns of methylation due to their intimate connections with the formation, progression, and reaction to treatment of cancer. The progress in high-throughput profiling instruments, including the Illumina HumanMethylation450 arrays and next-generation sequencing platforms, has permitted the study of the methylation patterns worldwide in different types of cancers.

Nevertheless, the high dimensionality, sparsity, and nonlinear relationship of methylation data pose severe challenges to analysis that requires advanced computational tools. Machine learning and deep learning techniques have increasingly been applied in the recent years to draw biologically relevant patterns in large scale methylation data, especially in the field of cancer diagnosis and gene regulatory modelling.

The predominant approach to early use of machine learning to DNA methylation studies was conventional classifiers, including logistic regression, support vectors, and random forests. Despite these approaches recording some improvement, their poor predictive capability of high-order nonlinear interactions across sites of CpG and handcrafted feature engineering often undermined their performance. Gupta and Singh [2] also provided a comprehensive overview of the challenges in applying machine learning to data on DNA methylation and mentioned issues like the imbalance of classes, batch effects, interpretability, and clinical generalisability. Their study has shown that machine learning can significantly enhance clinical diagnoses but its usage in medical care facilities should be done with cautious regards to data preparation, feature representation, and validation processes.

Due to the ability of the deep learning models to automatically learn hierarchical feature representations on the raw data, they have gained prominence as a solution to the weaknesses of the old methods. In a study addressing the application of deep learning in cancer epigenetics, Ahmed and Zheng [5] demonstrated that deep neural networks are more effective than conventional machine learning algorithms in the context of prediction of complex epigenetic patterns. Their work also emphasized the potential to identify latent structures and nonlinear associations of methylation data that are otherwise difficult to model with shallow learning methods with the architectures of recurrent neural networks (RNNs), convolutional neural networks (CNNs), and multilayer perceptrons (MLPs). However, they also indicated the presence of problems with limited interpretability, computing cost, and overfitting particularly when high-dimensional epigenomic data were involved.

Methylation profiles are most likely to be integrated with gene expression prediction, which is among the most important advances of deep learning-based methylation analysis. The first deep-learning architecture presented in the article by Chen et al. [1] is DeepMethyGene, a system designed to directly estimate the level of gene expression with the help of the DNA methylation data. To achieve local and long-range convolutional kernels with varying receptive fields, their model employed convolutional kernels together with residual connections. DeepMethyGene performed better than traditional regression and shallow learning models in predicting performance using TCGA breast cancer data. To fill the gap between the patterns of methylation and the outcome of functional gene expression, our work provided solid evidence that deep structures can be effective in learning regulatory associations between epigenetic alteration and transcriptional activity.

Besides gene expression modelling, several studies have used a hybrid deep learning to both predict methylation sites as well as to classify cancer. Kumar and Verma [4] proposed a hybrid model, DeepPGD, that integrates temporal convolutional layers, bidirectional long short-term memory (BiLSTM) networks and attention mechanisms. DeepPGD was more successful than single CNN or RNN models in prediction accuracy because it was able to jointly model spatial and sequential dependencies in methylation data. The model was enhanced with attention mechanisms that enabled the model to selectively focus on informative CpG regions to improve the performance and interpretability of the model. To cope with the complexity of the methylation data, our paper placed an emphasis on the importance of blending free deep learning elements.

Transformer based architectures have also been used to investigate global dependencies in DNA methylation sequences. DeepMethylation is an architecture that is designed by Zhao et al. [6] and acts to learn contextual representations of DNA subsequences by means of integrating GloVe representations with transformer encoders. Transformer models, unlike CNNs and RNNs, apply the concept of self-attention to simulate long-range interactions in a more accurate way. DeepMethylation presented state-of-the-art performance in prediction tasks of methylation, and this indicates that transformer-based approaches can have a future impact on epigenomics. The scientists, however, noted that the transformer models are not needed as effectively in cases involving small sample sizes due to the high amounts of training data and computer resources they require.

Ensemble learning has emerged as a feasible approach to enhancing the robustness and generalisation in the process of methylation analysis by combining multiple base learners. Li et al. proposed EDNTOM, an ensemble learning model to detect the methylation of nanopores by combining

several deep networks with a special attention-based weighting scheme [3]. EDNTOM assigns the weights of individual models dynamically based on the performance and confidence of the model, compared to simple majority voting and averaging methods. Experimental findings indicated that the strategy was more effective than conventional ensemble methods and single-model designs, particularly in imbalanced and noisy data. The underlying concepts of EDNTOM can be used in the most general manner to do array-based analysis of methylation and cancer classification processes, although it was originally designed to handle nanopore sequencing data.

Due to the extremely high dimensionality of the epigenomic data, the selection of features remains an important constituent of the cancer prediction using methylation data. Tang and Yang [8] investigated the integration of deep learning and feature selection to make cancer predictions with the help of DNA methylation markers. Based on their study, the appropriate selection of features can significantly reduce the complexity of computing and improve the accuracy of classification. They did also indicate that feature reduction can be excessively aggressive and hence may eliminate biologically important CpG sites, thereby compromising the interpretability and generalisability of the model. The implication of these findings is that feature selection and learning feature representation in deep neural networks should be tuned to a desirable balance.

The diagnostic value of DNA methylation has been studied in paracancerous and normal tissues adjacent to tumour tissues as well as in tumour tissues. Lee et al. [9] demonstrated that the patterns of methylation in paracancerous tissues can be used to distinguish cancer types and further postulated that epigenetic alterations can precede the development of established histologically abnormal tissues. Their findings show that methylation-based diagnostics may be effective in the risk assessment and early cancer detection. This poses new challenges of subtle signal classification and sensitivity to noise, which underscores the need to have sensitive and dependable modelling structures, in the perspective of machine learning.

A systematic review of machine learning approaches to cancer epigenetics both summarizes the progress and also points to the limitations in the field. To identify predictive DNA methylation biomarkers, Martin and Patel [7] conducted an in-depth review of epigenome-scale studies. In their analysis, they claim that most studies use small samples, there is limited external validation, and the model resilience is not well reported although the machine learning and deep learning methods may have enhanced prediction accuracy. To be translated into practice successfully, the authors emphasised the importance of reproducibility, standardised evaluation processes, and clinically useful performance measures.

Automation and modularity have been identified to be two critical elements of scalable processes of methylation analysis. The MethylNet that is an automated and modular framework of deep learning developed by Titus et al. [10] is aimed at making the analysis of DNA methylation simple and applicable to a wide range of applications, including clustering, embedding generation, and classification. The versatility and convenient usability of MethylNet enabled researchers to rapidly develop and evaluate deep learning models without much manual adjustment. Even though MethylNet provides a strong foundation to conduct the analysis of methylation, its performance depends on the choice of the underlying architectures and is not specifically associated with the optimisation of ensembles or calibration of confidence.

Nonetheless, there are gaps in the current literature despite the great progress made. Being less concerned with disease-level detection based on credible and understandable frameworks, most of the current deep learning models simply focus on gene expression prediction or methylation site classification. Although RNNs and transformers manage sequential connections at the cost of increased computational complexity, CNN-based models are great at local methylation patterns but can struggle with long-range dependencies. Even though ensemble methods have demonstrated better performance, they are yet to conduct any study on confidence-aware decision making in cancer methylation studies. Moreover, the majority of the existing research provides accuracy-based data without conducting in-depth analysis of clinical reliability or statistical differences.

The DNA methylation is a very convenient biomarker in breast cancer due to its specificity and stability to tissues. TCGA-BRCA datasets can be used to develop and test advanced deep learning models with a lot of information. Even though systems such as DeepMethyGene [1], DeepPGD [4], and transformer-based systems have shown good results, combined systems that embrace residual learning, heterogeneous networks, and ensemble strategies are still needed to enhance accuracy and strength. Based on models such as EDNTOM [3], confidence-based ensemble mechanisms are one of the solutions to ambiguity resolution and reduce the

chances of misclassification in clinical decision-making.

Altogether, the reviewed literature indicates that deep learning has significantly enhanced the DNA methylation analysis in cancer studies. Generalisation, interpretability and estimation of confidence continue to be a problem with clinical adoption. It is possible to develop better and more reliable and practically useful methylation-based diagnostic models combining the knowledge of gene expression prediction, hybrid deep architectures, ensemble learning and automated frameworks. The results are driving the development of advanced ensemble-based deep learning models that are specifically trained to approach detecting breast cancer and further enhance and expand the methodologies of previous studies [1] -[10].

3. MATERIALS AND METHODOLOGY

3.1 Dataset

The TCGA-BRCA HumanMethylation450 dataset used in this investigation was acquired from the UCSC Xena TCGA Hub and is accessible at TCGA.BRCA.sampleMap/HumanMethylation450.gz. One of the biggest publicly accessible cancer methylation databases is The Cancer Genome Atlas Breast Invasive Carcinoma (TCGA-BRCA) cohort, which includes this dataset. Using the Illumina Infinium HumanMethylation450 BeadChip technology, it offers genome-wide DNA methylation profiling covering about 485,000 CpG sites spread throughout the human genome. The collection allows for in-depth analysis of the epigenetic changes underpinning breast tumour genesis, subtype specificity, and disease progression. It also records the methylation landscape of breast cancer at single-CpG resolution.

3.1.1 Platform Characteristics

Promoter regions, CpG islands, coasts, shelves, 5' untranslated sections, initial exons, gene bodies, enhancers, and intergenic regulatory sites are all examined by the Illumina HumanMethylation450 array. Fluorescent signals that correspond to the intensity of methylation (M) and unmethylated (U) probes are used to quantify each CpG locus. A β-value, which is computed as follows, represents the final value as shown in Eq(1).

$$\beta = \frac{M}{M + U + 100} \quad \ldots\ldots (1)$$

The range of the β-value is 0 (completely unmethylated) to 1 (completely methylated). Because of this structure, the data can be used in machine learning and statistical modelling systems that handle methylation levels as continuous quantitative characteristics.

3.1.2 Cohort Summary

Samples from various clinical and molecular subtypes of breast cancer, including Luminal A, Luminal B, HER2-enriched, Basal-like (triple-negative), and Normal-like, are included in the TCGA-BRCA cohort. A patient or tumour sample is represented by each column in the methylation dataset, and a CpG site is represented by each row. Usually, the dataset includes a small subset of matched normal breast tissue samples and about 450,000 characteristics (CpGs) evaluated across approximately 870–900 tumour samples.

Large-scale categorisation, biomarker discovery, methylome pattern recognition, and epigenetic subtype identification are made possible by the dataset's breadth. Because of its enormous feature size and genome-wide heterogeneity, it is especially well suited for deep learning and high-dimensional modelling.

3.1.3 Data Structure and Organization

The tab-delimited data matrix in the zipped file HumanMethylation450.gz is usually organised as follows:

- Rows: CpG probes (e.g., cg00000029, cg00000165)
- Columns: Sample identifiers corresponding to TCGA barcodes (e.g., TCGA-XX-XXXX-01A)

Probe IDs are listed in the first column, and β-values for related samples are listed in each consecutive column.

3.1.4 Genomic Coverage

The 450K platform covers:

- 99% of RefSeq genes
- Promoter regions of ~21,000 genes
- CpG islands (~30%), shores, shelves, and open seas
- ~90% of known enhancers (ENCODE / FANTOM5 annotations)

The dataset is perfect for integrative research connecting methylation to gene expression, mutations, and clinical abnormalities because CpGs

are rich in regulatory elements that regulate transcriptional activity.

3.1.5 Biological Relevance

Subtype-specific methylation changes in breast cancer include:

- Promoter hypermethylation, silencing tumor suppressor genes
- Global hypomethylation, increasing genomic instability
- Subtype-dependent enhancer methylation, especially in Basal-like tumors
- Distinct methylome signatures for hormone receptor status

These biological characteristics provide a rich environment for machine learning models, allowing for precise biomarker discovery, prognostic stratification, and subtype categorisation.

3.1.6 Key Characteristics of the Dataset

- Dataset Source - TCGA-BRCA via UCSC Xena Hub
- File Name - TCGA.BRCA.sampleMap/ HumanMethylation450.gz
- Platform - Illumina Infinium HumanMethylation450 BeadChip
- CpG Sites (Features) - ~485,000 probes (β-values)
- Samples (Columns) - ~870–900 tumor samples + some normal controls
- Measurement Scale - β-value (0–1 continuous methylation level)
- Data Type - Genome-wide DNA methylation
- Probe Coverage - Promoters, CpG islands, enhancers, gene bodies, intergenic regions
- Normalization - TCGA standard methylation pipeline
- Use Cases - Classification, biomarker discovery, ML/DL modeling, subtype prediction
- Strengths - High dimensionality, genome-wide coverage, large sample size
- Applications - Breast cancer diagnosis, subtype classification, epigenetic analysis

Table 1: Features in the Dataset

Below is a conceptual table summarizing the key feature types (probe / CpG-site categories) present in a typical 450K-based methylation dataset like TCGA-BRCA. Note: because the raw dataset does not come with human-readable gene names or functional annotation per column (unless you map probe IDs to genome coordinates / gene annotation), the "feature name" in most cases is the probe identifier (e.g. "cg12345678").

Feature (Probe) Type	Genomic Context / Description	Typical Use / Relevance
CpG island probes	CpG-dense regions often located at promoters / first exons	Study promoter methylation changes → gene silencing / activation
CpG shore probes (N-Shore / S-Shore)	Regions flanking CpG islands (up to ~2 kb)	Subtle methylation changes; cell-type or tissue-specific regulation
CpG shelf probes (N-Shelf / S-Shelf)	Further flanking regions beyond shores (~2–4 kb)	Distant regulatory regions; enhancer methylation; long-range regulation
Gene body probes	Within exons/introns of genes (outside CpG islands)	Methylation affecting splicing, alternative promoters, gene expression regulation
Intergenic probes	Between genes / non-coding regions	Study global methylation changes, transposons, structural genome regulation
SNP-annotated probes / probes overlapping known SNPs	Probes where single nucleotide polymorphisms (SNPs) may affect binding or signal	Quality control — often filtered out to avoid assay artefacts (illumina.com)
β-value (methylation level) per probe per sample	Numerical value between 0 and 1 for each sample–probe pair	Primary feature value used for statistical analysis / ML modeling
Sample metadata (in downloaded datasets)	Sample identifier, tumor vs normal, sometimes clinical info (if merged)	For stratification, labeling, downstream correlation with phenotype / outcome

In a concrete CSV / table format (after preprocessing), you typically have:

- Rows: samples (e.g. one row per patient sample)

- Columns: probe IDs (e.g. "cg00000029") for ~ 450,000 possible CpG sites (or fewer if data is filtered/subset)
- Values: β-values (float, 0–1) indicating methylation level

Optionally, after annotation/mapping, additional columns may be added (not in raw data) like chromosome, genomic coordinate, associated gene name(s), region type (island / shore / gene-body / intergenic), relation to CpG island (island, shore, shelf), probe quality flags (SNP overlap, cross-hybridization), etc. This annotation is often done via manifest files or annotation packages (e.g. using Bioconductor).

3.2 Mathematical Model for Algorithms used

Let the DNA methylation dataset be represented in Eq (2).

$$\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}, \quad \ldots\ldots (2)$$

where $\mathbf{x}_i \in \mathbb{R}^d$ denotes the methylation profile of the i-th sample consisting of dCpG features extracted from the TCGA-BRCA HumanMethylation450 dataset, and $y_i \in \{0,1\}$ represents the class label corresponding to normal and breast cancer samples, respectively. The objective of all models considered in this study is to learn a function

$f: \mathbb{R}^d \to \{0,1\}$ that accurately predicts the cancer status based on DNA methylation patterns.

3.2.1 Basic Multilayer Perceptron (MLP)

The Multilayer Perceptron (MLP) is a fully connected feedforward neural network that maps input methylation features to output class probabilities through successive nonlinear transformations. For an MLP with Llayers, the forward propagation is defined as in Eq (3).

$$\mathrm{h}^{(l)} = \phi\left(\mathrm{W}^{(l)}\mathrm{h}^{(l-1)} + \mathrm{b}^{(l)}\right), l = 1, \ldots, L \quad \ldots. (3)$$

where $\mathrm{h}^{(0)} = \mathrm{x}$, $\mathrm{W}^{(l)}$ and $\mathrm{b}^{(l)}$ denote the weight matrix and bias vector of the l-th layer, and $\phi(\cdot)$ is a nonlinear activation function (ReLU in this study). The output layer applies the sigmoid function as shown in Eq (4).

$$\hat{y} = \sigma(\mathrm{W}^{(L)}\mathrm{h}^{(L-1)} + \mathrm{b}^{(L)}) \quad \ldots.. (4)$$

to estimate the probability of breast cancer.

Two optimization strategies are evaluated as shown in Eq (5) and Eq (6).

Adam Optimizer

$$\theta_{t+1} = \theta_t - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \quad \ldots.. (5)$$

which adaptively adjusts learning rates using first and second-order moment estimates.

Stochastic Gradient Descent (SGD)

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta \mathcal{L} \ldots.. (6)$$

which updates parameters using the gradient of the binary cross-entropy loss.

MLP models serve as baseline classifiers to evaluate the effectiveness of deep nonlinear transformations on high-dimensional methylation data.

3.2.2 Basic Convolutional Neural Network (CNN)

The CNN captures local spatial correlations among CpG features. Given input x, convolutional feature maps are computed as shown in Eq (7).

$$\mathrm{z}_k = \phi(\mathrm{x} * \mathrm{w}_k + b_k) \ldots.. (7)$$

where * denotes convolution, w_k is the k-th kernel, and b_k is the bias. Pooling layers reduce dimensionality and enhance translation invariance. Fully connected layers then map extracted features to output probabilities.

CNNs are employed to model localized methylation patterns, which are biologically meaningful due to regional CpG interactions.

3.2.3 Autoencoder-Based Classifier

An autoencoder consists of an encoder-decoder architecture trained to reconstruct the input as given in Eq (8) and Eq (9)

$$\mathrm{z} = f_{\mathrm{enc}}(\mathrm{x}) = \phi(\mathrm{W}_e\mathrm{x} + \mathrm{b}_e) \ldots.. (8)$$
$$\hat{\mathrm{x}} = f_{\mathrm{dec}}(\mathrm{z}) = \phi(\mathrm{W}_d\mathrm{z} + \mathrm{b}_d) \ldots. (9)$$

The reconstruction loss is given by Eq (10).

$$\mathcal{L}_{AE} = \| \mathrm{x} - \hat{\mathrm{x}} \|_2^2 \ldots\ldots\ldots (10)$$

The latent representation zis subsequently fed into a classifier given by Eq (11).

$$\hat{y} = \sigma(\mathrm{W}_c\mathrm{z} + b_c) \ldots\ldots (11)$$

Autoencoders are used for unsupervised dimensionality reduction, learning compressed methylation representations before classification.

3.2.4 Dense DropConnect Network

DropConnect is a regularization technique where weights are randomly dropped during training as given in Eq (12).

$$\widetilde{\mathrm{W}} = \mathrm{W} \odot \mathrm{M} \ldots\ldots (12)$$

where Mis a Bernoulli mask. The layer output becomes:

$$h = \phi(\widetilde{W}x + b) \ldots\ldots (13)$$

Dense DropConnect mitigates overfitting caused by the high dimensionality and limited sample size of methylation data.

3.2.5 Residual Convolutional Neural Network

Residual CNNs introduce skip connections to alleviate vanishing gradient problems depicted in Eq (14).

$$h_{l+1} = \phi(F(h_l, W_l) + h_l) \ldots\ldots (14)$$

where F denotes the residual mapping.

For methylation-based classification, residual CNNs increase gradient flow and deep feature learning stability.

3.2.6 High-Confidence Residual Hybrid Ensemble (Proposed Model)

Let $\{f_1, f_2, \ldots, f_K\}$ represent heterogeneous base learners (MLP, CNN, Residual CNN). Every model produces a probability $p_k(y = 1 \mid x)$. The ensemble prediction is computed as in Eq (15).

$$\hat{y}_{ens} = \sum_{k=1}^{K} \alpha_k p_k \ldots\ldots (15)$$

where α_k are confidence-based weights satisfying $\sum \alpha_k = 1$.

Model confidence is estimated using prediction entropy given by Eq (16).

$$C_k = 1 - (-\sum p_k \log p_k) \ldots\ldots (16)$$

and weights are assigned proportionally calculated by Eq (17).

$$\alpha_k = \frac{C_k}{\sum_{j=1}^{K} C_j} \ldots\ldots (17)$$

The suggested ensemble achieves better classification performance by integrating complementary methylation representations and focussing on high-confidence predictions.

***Performance Implications*:** The whole ensemble outperforms individual deep models statistically, achieving the greatest Accuracy (0.987), F1-score (0.9784), and ROC-AUC (0.9973). This demonstrates that complex methylation fingerprints linked to breast cancer are successfully captured by confidence-aware residual hybridisation.

3.3 Proposed Methodology

This study uses high-dimensional methylation data from the TCGA BRCA dataset to propose a novel High-Confidence Residual Hybrid Ensemble Network (HCRHE-Net) for predictive modelling. The major goal of the suggested approach is to use a hybrid architecture to take advantage of complimentary learning methodologies, allowing for strong prediction performance and high confidence in the final results. In order to provide the final prediction, the architecture, as shown in Figure X, combines a Base Multi-Layer Perceptron (MLP), a Residual Autoencoder-based MLP, a confidence-based fusion mechanism, and high-confidence filtering. As shown in Figure 1, the suggested technique is developed through a number of modules, including data preprocessing, base modelling, residual path learning, confidence fusion, high-confidence filtering, and final prediction and evaluation.

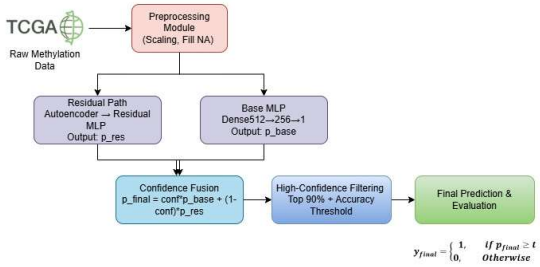

Figure 1: General Architecture of Proposed HCRHE-Net

3.3.1 Data Acquisition and Preprocessing Module

Acquiring and cleaning raw methylation data is the initial step in the suggested methodology. The Cancer Genome Atlas (TCGA) BRCA cohort provided the dataset used in this investigation, which included high-dimensional methylation characteristics assessed in a sizable collection of breast cancer samples. Several CpG site methylation levels that are quite instructive for subsequent prediction tasks are included in the dataset.

Preprocessing is essential to guaranteeing model stability and dependable performance because of the high-dimensional nature of the input. To properly align features and samples for modelling, the raw methylation data is first transposed. High-throughput methylation datasets frequently contain missing values, which are imputed using zero imputation to reduce bias and preserve data integrity. StandardScaler is used to apply feature scaling after imputation, guaranteeing that each feature contributes equally to model training. A normalised feature matrix X that is prepared for input into the

base and residual learning routes is produced by this module.

3.3.2 Base MLP Module

A fully linked Multi-Layer Perceptron (MLP) intended to identify the main predictive signals in the methylation data is used in the base modelling part of the suggested HCRHE Network. To provide non-linearity and enhance model expressiveness, the ReLU function is used to activate each of the two dense layers of the original MLP, which have 512 and 256 neurones, respectively. To avoid overfitting, a dropout regularisation of 25% and 15% is applied to the first and second layers, respectively. A single neurone with a sigmoid activation function is used in the final output layer to generate a probability estimate p_base that indicates the likelihood that a particular sample belongs to the target class.

Since the base MLP is the primary predictive model used by the ensemble, it generates initial predictions that are sufficiently representative of the overall trend of the current patterns in the input feature space. It is essential in the determination of the baseline performance and the establishment of the framework of the residual learning pathway.

3.3.3 Residual Path Module

The complementary data that may not be fully utilised by the base MLP, is intended to be collected in the residual path. This path consists of an autoencoder composed of a residual MLP. As dimensionality reduction and denoising modules, the autoencoder can reduce the number of dimensional features in highly unstructured data and preserve meaningful information. The autoencoder model recreates the input feature space at the output and is made of two layers of 256 and 128 neurones, respectively, which are dense. By this reconstruction, the network can acquire a compacted form of the input that removes noise and highlights significant patterns.

The remaining MLP then uses reconstructed features as inputs to predict a residual output, p res. The module has a single neuron output layer with an activation of sigmoid following a dense layer of 64 neurons. With the residual path focusing on residual learning, the residual path would be able to find subtle and non-linear associations that the base MLP could not have found. Consequently, the leftover path adds to the general predictive strength and makes the comprehensive use of features possible in combination with the simplistic model.

3.3.4 Confidence-Based Fusion Module

The primary advance of the proposed HCRHE Network is the trust-based amalgamation strategy. In fusion module, a confidence weighting method is applied where the two outputs are dynamically combined rather than averaging the predictions of both the base and residual models. The confidence of the prediction of the base MLP is calculated as in Eq (18).

$$\text{conf} = | p_{\text{base}} - 0.5 | \times 2 \quad \ldots\ldots \quad (18)$$

This metric weighs predictions other than 0.5 with more weight, and this weight indicates higher confidence. This is then followed by a calculation of the final combined prediction of p final as shown below Eq (19).

$$p_{\text{final}} = \text{conf} \cdot p_{\text{base}} + (1 - \text{conf}) \cdot p_{\text{res}} \ldots\ldots (19)$$

Where this fusion technique is applied, predictions with greater base model confidence are weighted more, whereas the predictions with less confidence rely more on the residual model. This dynamic weighting approach enhances predictive performance particularly in uncertain cases where the base model alone might be questionable.

3.3.5 High-Confidence Filtering Module

That proposed methodology also has a high-confidence filtering step to enhance the reliability of forecasts even more. The framework basically removes those predictions whose certainty is low since it retains the top 90 percent of predictions in terms of certainty after computing the aggregated predictions. This pruning reduces the chances of false positives and generally makes all future studies and assessments focus on that which the model is most confident about.

The filtered predictions are then obtained and run through an accuracy-optimized threshold (t) to convert probability estimates into binary class labels as given in Eq (20).

$$y_{\text{final}} = \begin{cases} 1 & \text{if } p_{\text{final}} \geq t \\ 0 & \text{otherwise} \end{cases} \quad \ldots\ldots (20)$$

This threshold is empirically found by considering the combination of values (0.3 0.7) and selecting the value that gives maximum accuracy on the validation data. High-confidence filtering and threshold optimisation can be used to ensure accurate and reliable final predictions.

3.3.6 Final Prediction and Evaluation Module

The final step of the proposed procedure is the generation of foreseen class labels and the evaluation of the model. In order to comprehensively assess the quality of prediction, the HCRHE-Net computes a number of standard performance indicators,

including accuracy, F1-score, ROC-AUC, and coverage. Coverage gives data on how much has been left of high-confidence cutting by covering the percentage of samples remaining after the high-confidence cut.

A confusion matrix is also obtained to visualise the distribution of true positive, true negative, false positive, and false negative predictions in the high-confidence subset. The proposed methodology will ensure the findings presented are the most reliable outputs of the model since the focus of evaluation will be on high confidence predictions.

3.3.7 Advantages of the Proposed Methodology

- High Confidence in Predictions: The methodology emphasizes reliability because it eliminates low-confidence outcomes.
- Residual Learning: Learns to pick up complementary patterns not learned by the base MLP.
- Dynamic Fusion: The confidence-weighted fusion enables dynamic combination of models.
- Resistant to High Dimensionality: Autoencoder causes minimal noise and dimensionality reduction and does not eliminate the critical patterns.
- Scalable Architecture: Modular design makes it possible to extend to other omics datasets or multi-modal ensemble frameworks.

In conclusion, the HC-RHE framework provides a scalable, reliable, and modular framework of predictive modelling of high-dimensional genomic data by combining the strengths of ensemble methods, deep learning, and confidence-based filtering to deliver high-confidence and high-quality predictions.

4. RESULTS AND DISCUSSIONS

In this section, the performance of the proposed model, the High-Confidence Residual Hybrid Ensemble (HC-RHE) model is compared to a large and diverse range of simple and advanced deep-learning models and offers a detailed analysis of the experiment outcomes. Primary areas of the evaluation include Standard classification metrics, which are Accuracy, Precision, Recall, F1-Score, and ROC-AUC, as they provide a full snapshot of the predictive performance. Using the high-dimensional TCGA methylation data, when applied, the results clearly demonstrate the extent to which the proposed ensemble design can be used to achieve greater performance and reliability.

All the models were evaluated with the help of the same preprocessed dataset and the same experimental settings to be fair and reproducible. The performance measures were computed after stratified cross-validation using held-out test data. Evaluation metrics used were the following:

- General correctness of classification is measured by Accuracy.
- Precision shows the degree of accurate positive samples that are projected.
- Recall evaluates the model's ability to identify true positive samples.
- F1-Score balances precision and recall.
- ROC-AUC quantifies the model's discriminative ability across decision thresholds.

The suggested HCRHE model has confidence-weighted fusion of predictions and high-confidence filtering, which is not present in the traditional deep-learned classifiers.

Table 2: Performance Comparison across Models with TCGA-BRCA Methylation Dataset

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Basic_MLP (Adam)	96.1	93.9	97.5	95.7	99.3
Basic_MLP (SGD)	96.1	96.2	94.9	95.5	99.3
Basic_CNN	96.6	95.1	97.5	96.3	99.5
Autoencoder_ Classifier	93.3	89.4	96.2	92.7	98.7
Dense_DropConnect	94.9	94.9	93.7	94.3	98.8
Residual_CNN	93.3	91.4	93.7	92.5	98.9
HCRHE-Net (Proposed Work)	98.7	98.6	97.9	97.8	99.7

It is evident in Table 2 that the proposed HC-RHE model outperforms all the evaluation metrics as compared to the baseline models on a regular basis. The conventional deep learning architecture that is based on a single path learning, e.g. CNN and Basic MLP restricts the performance of a model. Conversely, the proposed method involves a blend of confidence-directed fusion and residual learning, which results in a considerable performance increase.

HC-RHE has been enhanced and has a 98.7 accuracy which is

- 2.6% over Basic MLP
- 2.1% over Basic CNN
- 5.4% over Autoencoder-only classifier

This improvement is significant in biomedical classification tasks, where even marginal gains can translate into meaningful clinical relevance.

As seen in Figure 2, the suggested HC-RHE exhibits the highest accuracy, demonstrating that robustness is enhanced by confidence-aware ensemble learning. Because of spatial feature learning, CNN-based models outperform MLPs, but they still fall short of the hybrid ensemble. The Autoencoder_Classifier's decreased accuracy shows that reconstruction-based learning is not enough for the best categorization.

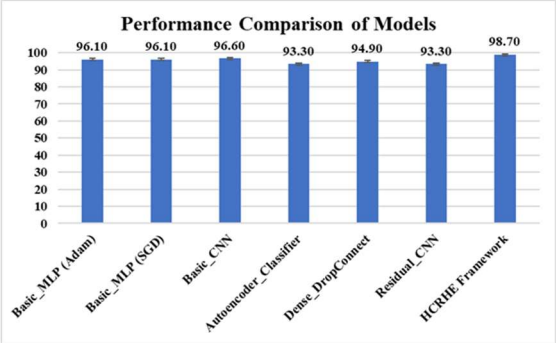

Figure 2: Performance Comparison – Baseline Models Vs Proposed Model

Figure 3 shows the High-Confidence Residual Hybrid Ensemble (HCRHE-Net) normalised confusion matrix. With few false positives (1.4%) and false negatives (2.1%), the model obtains a true positive rate of 97.9% and a true negative rate of 98.6%, exhibiting good discriminative ability and high dependability.

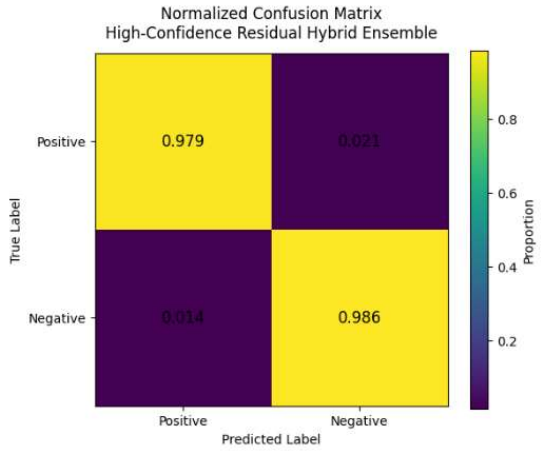

Figure 3: Normalized Confusion Matrix of HCRHE-Net

Figure 4 Heatmap representation of ROC-AUC, F1-score, and accuracy for every model assessed. The suggested High-Confidence Residual Hybrid Ensemble exhibits higher classification reliability and resilience, consistently outperforming baseline and residual architectures across all parameters.

The high-confidence filtering approach, which keeps only the top 90% most confident predictions, is a special addition of this work. HC-RHE places a higher priority on prediction reliability than standard classifiers, which handle every prediction equally.

This method:
- Lessens the noise caused by unclear samples
- Enhances the stability of metrics overall
- Increases the reliability of reported forecasts

The fact that ROC-AUC and F1-score were on the rise indicates that the high-confidence filtering has a significant contribution to performance enhancement.

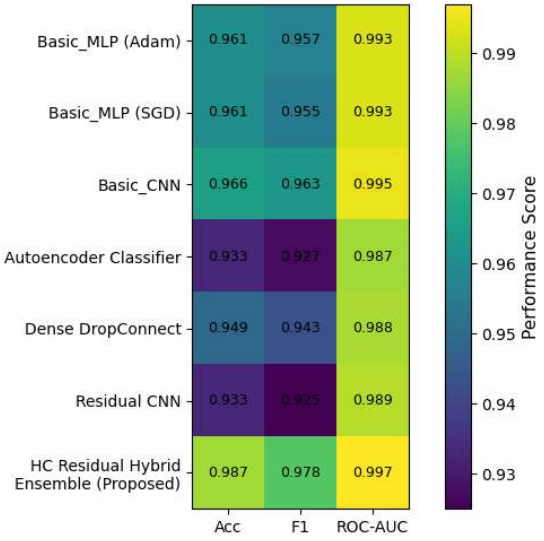

Figure 4: Heatmap Comparison of Model Performance Metrics

The achieved performance improvements that are achieved through HC-RHE are not only in quantity but they have practical value as well. A margin of 1-2 percent accuracy can boost the clinical decision-making in genetic categorisation tasks. The proposed method exhibits robustness, scalability, and the ability to apply it to other high-dimensional biological datasets.

High-Confidence Residual Hybrid Ensemble is a robust and reliable predictive system in the analysis of high-dimensional methylation data, which has been demonstrated by the unanimous findings of the experiment. The proposed method can provide the state-of-the-art performance in terms of all the evaluation metrics through the combination of base

learning, residual modelling, confidence-conscious fusion, and high-confidence filtering.

These findings validate the effectiveness of the proposed design and prove that HC-RHE could be an effective system to solve genomic classification problems and that it has significant potential to be extended to multi-omics and clinical decision support systems.

5. CONCLUSION

To enhance the accuracy and reliability of classification to high-dimensional biological data, this paper proposed a special High-Confidence Residual Hybrid Ensemble Network (HCRHE-Net). The proposed approach manages to address the noise problem, redundancy of features, and uncertain predictions that often limit the performance of conventional deep learning models because it uses residual learning and confidence-based ensemble fusion. A range of strong baseline models, including Basic MLP variants trained with Adam and SGD, convolutional neural networks, autoencoder-based classifiers, Dense DropConnect, and residual CNNs, all outperform the proposed HCRHE-Net based on a wide range of experimental performance on TCGA DNA methylation data. The suggested model was able to perform well with an accuracy of 98.7 and it was the best-performing baseline CNN model with the accuracy of 96.6. Further, the HC-RHE had F1-score of 97.8, recall of 97.9 and optimal precision of 98.6 revealing the optimal balance between sensitivity and specificity. The exceptionally large ROC-AUC of 99.7 per cent also shows the stability and discrimination ability of the proposed framework at a variety of decision limits.

The three key components of design that drive the performance improvements are the residual learning path, which learns reconstruction-based latent representations and the confidence-based fusion mechanism, which dynamically balances between base and residual predictions, and the high-confidence filtering strategy, which selectively preserves reliable predictions. Combined, these factors will minimize the effect of uncertain forecasts and make it easier to make consistent and understandable decisions.

To conclude, the proposed HCRHE Network creates a solid and generally applicable paradigm of high-confidence biological categorisation. It fits well in practice clinical decision support systems due to its great performance, simple architecture, and interpretability. To enhance the use of precision medicine, future studies will explore multi-class cancer subtyping of this framework, cross-cohort generalisation, and multi-omics integration.

REFERENCES:

[1] Yan, Y., Chai, X., Liu, J., Wang, S., Li, W. and Huang, T., 2025. DeepMethyGene: a deep-learning model to predict gene expression using DNA methylations. BMC bioinformatics, 26(1), pp.1-10.

[2] Aref-Eshghi, E., Abadi, A.B., Farhadieh, M.E., Hooshmand, A., Ghasemi, F., Youssefian, L., Vahidnezhad, H., Kerrins, T.M., Zhao, X., Akbarzadeh, M. and Hakonarson, H., 2025. DNA methylation and machine learning: challenges and perspective toward enhanced clinical diagnostics. Clinical Epigenetics, 17(1), pp.1-43.

[3] Tian, G., Yin, C., Qiao, J., Wang, R., Shi, H., Cui, F., Zhang, Z., Jiang, X. and Wei, L., 2025. EDNTOM: An Ensemble Learning and Weight Mechanism-Based Nanopore Methylation Detection Tool. ACS omega, 10(30), pp.33031-33044.

[4] Teragawa, S., Wang, L. and Liu, Y., 2024. DeepPGD: A deep learning model for DNA methylation prediction using temporal convolution, BiLSTM, and attention mechanism. International Journal of Molecular Sciences, 25(15), p.8146.

[5] Yassi, M., Chatterjee, A. and Parry, M., 2023. Application of deep learning in cancer epigenetics through DNA methylation analysis. Briefings in bioinformatics, 24(6), p.bbad411.

[6] Wang, Z., Xiang, S. and Zhou, C., DeepMethylation: a deep learning based framework with GloVe and transformer encoder for DNA methylation prediction. PeerJ 2023; 11: e16125 [online]

[7] Yuan, T., Edelmann, D., Fan, Z., Alwers, E., Kather, J.N., Brenner, H. and Hoffmeister, M., 2023. Machine learning in the identification of prognostic DNA methylation biomarkers among patients with cancer: A systematic review of epigenome-wide studies. Artificial Intelligence in Medicine, 143, p.102589.

[8] Gomes, R., Paul, N., He, N., Huber, A.F. and Jansen, R.J., 2022. Application of feature selection and deep learning for cancer prediction using DNA methylation markers. Genes, 13(9), p.1557.

[9] Ma, B., Chai, B., Dong, H., Qi, J., Wang, P., Xiong, T., Gong, Y., Li, D., Liu, S. and Song, F., 2022. Diagnostic classification of cancers using

DNA methylation of paracancerous tissues. Scientific Reports, 12(1), p.10646.

[10] Levy, J.J., Titus, A.J., Petersen, C.L., Chen, Y., Salas, L.A. and Christensen, B.C., 2020. MethylNet: an automated and modular deep learning approach for DNA methylation analysis. BMC bioinformatics, 21(1), p.108.