Journal of Information Technology Review

DLINE JOURNALS

Print ISSN: 0976-3511
Online ISSN: 0976-2922

JITR 2026: 17 (3)
https://doi.org/10.6025/jitr/2026/17/3/148-172

Explainable Machine Learning for DNA Methylation Prediction: A Comprehensive Interpretability and Biological Analysis

Puji Lestari, M. Kom
Universitas Islam Negeri Sunan Kalijaga. Yogyakarta
Indonesia
puji.lestari@uin-suka.ac.id

ABSTRACT

Background: DNA methylation is a critical epigenetic mechanism regulating gene expression, cellular differentiation, and disease pathogenesis. Although machine learning (ML) has shown promise in predicting methylation status from genomic features, many predictive models operate as "black boxes," limiting biological interpretability and clinical translation.

Objective: This study develops an explainable machine learning framework for DNA methylation prediction that integrates predictive modeling with multiple interpretability techniques to identify influential genomic determinants and elucidate their biological relevance.

Methods: A supervised classification model was trained on a DNA methylation dataset comprising four explanatory variables CpG density, genomic location, regulatory score, and conservation score with binary methylation status as the target. Global interpretability was achieved using SHAP (SHapley Additive exPlanations) feature importance and summary plots. Local explanations were generated using SHAP waterfall and force plots alongside LIME (Local Interpretable Model Agnostic Explanations). Functional relationships were examined through Partial Dependence Plots (PDP) and Individual Conditional Expectation (ICE) plots. Robustness was assessed via bootstrap resampling (1,000 iterations), five-fold cross-validation, and correlation analysis between SHAP and LIME explanations.

Results: CpG density emerged as the most influential predictor (mean absolute SHAP = 0.36), followed by regulatory score (0.24), conservation score (0.13), and genomic location (0.05). SHAP dependence plots revealed nonlinear threshold effects, with CpG density values below 0.3 producing negative contributions and values above 0.7 strongly increasing methylation probability. PDP analyses confirmed saturation behavior at high feature values. Local explanations successfully decomposed individual predictions,

demonstrating that high confidence methylation calls (e.g., P = 0.92) were driven primarily by CpG density and regulatory score. Robustness analyses showed strong agreement between SHAP and LIME (Pearson r = 0.97) and low cross validation variability (coefficient of variation <10% for top features).

Conclusion: Explainable machine learning provides biologically interpretable insights into DNA methylation determinants while maintaining predictive capability. The consistent identification of CpG density, regulatory activity, and evolutionary conservation as dominant predictors aligns with established epigenetic principles. This framework offers a transparent, reproducible approach for epigenetic biomarker discovery and supports precision medicine applications. Future work should extend these methods to multi-omic datasets and experimental validation.

Keywords: DNA Methylation, Explainable Artificial Intelligence (XAI), SHAP, LIME, Epigenetic Biomarkers, Machine Learning, Interpretability

Received: 135 January 2026, Revised 15 March 20026, Accepted 4 April 2026

1. Introduction

DNA methylation is a fundamental epigenetic mechanism that involves the chemical modification of genomic DNA without altering the underlying nucleotide sequence. This process occurs through the addition of methyl groups to DNA molecules and plays a critical role in regulating gene expression, cellular differentiation, and biological function [1, 2]. DNA methylation primarily occurs at CpG sites, where cytosine and guanine nucleotides occur in tandem within the DNA sequence [3]. Because methylation patterns influence numerous physiological and pathological processes, they have become an important target for computational prediction and biological interpretation.

2. Review of Earlier Studies

Recent advances in artificial intelligence, particularly deep learning (DL), have substantially transformed biological data analysis. Deep learning frameworks can identify complex nonlinear relationships and extract meaningful representations from large-scale sequential and high dimensional datasets [4, 5]. These capabilities have enabled successful applications across multiple biological and medical domains [6-10]. Nevertheless, many challenges in life-science research remain unresolved and continue to motivate the development of more advanced predictive and interpretable analytical frameworks.

Among these developments, deep neural network (DNN) based approaches have demonstrated significant potential for DNA methylation analysis. One notable example is MethylNet, a DNN framework developed to predict phenotypic outcomes from DNA methylation (DNAm) data while accounting for the inherent nonlinear structure of methylation patterns [11]. By employing multilayer nonlinear activation mechanisms, MethylNet achieved improved predictive performance over conventional linear regression approaches in tasks including biological age estimation, cancer subtype classification, and cell type proportion prediction [12]. To improve interpretability, MethylNet integrates the SHapley Additive exPlanations (SHAP) framework, allowing identification and prioritization of influential CpG sites contributing to model outputs [13].

Beyond predictive accuracy, reducing feature dimensionality remains a major concern in methylation based learning systems because methylation datasets often contain an extremely large number of variables. To address this challenge, Gomes et al. introduced a deep learning framework that combines statistical feature selection methods, including ANOVA and Random Forest, before applying deep neural architectures, thereby reducing dimensionality and improving computational efficiency [14].

The rapid expansion of large scale biological datasets, together with improvements in computational infrastructure, has accelerated the application of artificial intelligence across genomic and epigenomic research. These advances support a wide range of predictive tasks, including regulatory genome annotation and single-cell categorization [15, 16, 17, 18]. Importantly, understanding the underlying mechanisms that drive model predictions has become increasingly valuable because it enables researchers to gain deeper biological insight from computational outputs [19].

This growing emphasis on interpretability has contributed to the emergence of explainable artificial intelligence (XAI), which seeks to improve transparency and trustworthiness in complex predictive systems [20]. In bioinformatics, the availability of extensive datasets and unsupervised generative approaches has enhanced analytical performance and increased applicability to transcriptomic and epigenetic studies [21, 22, 23, 24, 25, 26]. Within this context, DNA methylation data have become especially important for classification and prediction tasks.

One prominent application involves cancer classification using methylation profiles from repositories such as The Cancer Genome Atlas (TCGA). Models developed for this purpose often achieve high classification performance by capturing both cancer-associated methylation alterations and tissue-specific methylation signatures [27 -33].

Despite these advances, experimental detection and computational analysis of DNA methylation remain challenging. Experimental approaches include restriction endonuclease based techniques, affinity enrichment methods, and bisulfite conversion technologies [34]. These techniques generate large-scale sequencing outputs containing extensive numbers of CpG sites, resulting in highly dimensional datasets that are computationally difficult to process and interpret.

Machine learning (ML) has emerged as an effective approach to extracting meaningful information from complex biological datasets. ML techniques can model numerous variables simultaneously and identify intricate interactions across integrated phenotypic datasets. Compared with traditional statistical approaches, machine learning often provides advantages in handling scale, complexity, and multidimensional relationships [35, 36]. In biological age prediction, for example, ML approaches integrate heterogeneous data sources including medical imaging, physical activity measurements, blood biomarkers, microbiome profiles, and genomic information [37-40].

For disease classification and biomarker discovery, comprehensive methylation analysis pipelines frequently include data harmonization, missing-value imputation, dimensionality reduction, predictive model construction, and post hoc interpretability analysis through XAI techniques [41]. Recent studies continue to extend these capabilities. Syed introduced iIRMethyl, a machine learning based framework for accurate identification of arginine methylation sites [42]. These advances highlight that model performance increasingly depends on effective feature engineering and robust feature selection strategies that preserve biological relevance while reducing complexity.

Building upon this direction, Sheetal Rajpal proposed XAI-MethylMarker, a two-stage explainable AI framework for biomarker discovery that integrates predictive modeling with interpretable feature attribution [43]. Such approaches illustrate the transition from purely predictive systems toward transparent and biologically interpretable computational pipelines. [44]

Overall, explainable machine learning represents an important advancement for DNA methylation prediction by combining high predictive capability with biological interpretability. The integration of deep learning, feature selection methods, and explainable AI offers new opportunities to understand epigenetic regulation, identify biomarkers, and improve translational applications in precision medicine.

3. Dataset

The present study utilizes the DNA Methylation Data: Epigenetic Biomarkers dataset obtained from the Kaggle repository. This dataset was developed to support the prediction of DNA methylation status using biologically relevant genomic and epigenetic attributes. DNA methylation is one of the most important epigenetic modifications and plays a critical role in regulating gene expression, cellular differentiation, ageing, and disease development, particularly cancer and other complex disorders. DNA methylation predominantly occurs at cytosine phosphate guanine (CpG) dinucleotides, where the addition of methyl groups influences chromatin accessibility and transcriptional activity.

The dataset comprises genomic observations characterized by four explanatory variables and one binary response variable. The predictor variables include CpG density, genomic location, regulatory score, and conservation score, each representing distinct biological properties associated with methylation mechanisms. CpG density quantifies the abundance of CpG sites within a genomic region and is closely related to CpG island structure. Genomic location identifies the positional context of a sequence region within the genome, capturing differences among promoters, exons, introns, and intergenic regions. Regulatory score reflects the influence of regulatory elements and chromatin activity, while conservation score represents the evolutionary conservation of genomic sequences across species, providing information regarding functional importance.

The response variable, methylation status, is represented as a binary class, where 1 denotes methylated regions and 0 denotes unmethylated regions. Consequently, the dataset presents a supervised binary classification problem for predicting the methylation state of genomic loci based on sequence and regulatory characteristics.

The selected variables correspond to biological factors that have been shown to influence DNA methylation patterns. Previous studies have demonstrated that genomic position, CpG composition, sequence properties, and regulatory elements are among the most informative determinants of methylation status and can achieve high predictive performance when incorporated into machine learning models. Therefore, the dataset provides an appropriate framework for investigating both predictive modeling and explainable artificial intelligence approaches for understanding the determinants of DNA methylation.

Variable	Description	Type
CpG density	Density of CpG dinucleotides within a genomic region	Continuous
Genomic location	Positional annotation of the genomic region	Categorical
Regulatory score	Quantifies regulatory activity associated with the region	Continuous
Conservation score	Evolutionary conservation measure of the genomic sequence	Continuous
Methylation status	Binary target variable indicating methylated (1) or unmethylated (0) state	Binary

Table 1. List of variables and their types

4. Research Statement and Research Design

4.1 Research Statement

DNA methylation is a central epigenetic mechanism involved in regulating gene expression, cellular differentiation, and disease development. Although machine learning and deep learning methods have demonstrated strong predictive performance in methylation related tasks, many existing approaches operate as black box systems and provide limited biological interpretability. As a result, it remains difficult to determine whether predictive performance reflects biologically meaningful mechanisms or merely statistical associations.

This study addresses this limitation by developing an explainable machine learning framework for DNA methylation prediction that integrates predictive modeling with multiple interpretability approaches. Rather than focusing exclusively on classification accuracy, the study investigates how genomic and epigenetic features contribute to prediction decisions and whether these contributions align with known biological principles.

The research aims to establish a transparent analytical pipeline capable of identifying influential determinants of methylation status while providing both global and local explanations of model behavior. Through explainable artificial intelligence techniques, the study seeks to bridge predictive analytics and biological interpretation to support biomarker discovery and precision medicine applications.

4.2 Research Objectives

The study is guided by the following objectives:

1. To develop a machine learning framework for predicting DNA methylation status using genomic and regulatory attributes.

2. To determine the relative importance of CpG density, genomic location, regulatory score, and conservation score in methylation prediction.

3. To investigate nonlinear relationships, threshold effects, and interaction patterns among biological predictors.

4. To explain individual prediction outcomes using local interpretability techniques.

5. To evaluate the robustness and reproducibility of explainability results through bootstrap resampling and cross validation.

6. To derive biologically meaningful insights that support epigenetic biomarker discovery.

4.3 Research Questions

The study addresses the following research questions:

RQ1: Which genomic and epigenetic features contribute most significantly to DNA methylation prediction?

RQ2: How do feature interactions and nonlinear effects influence methylation probability?

RQ3: Can explainable AI techniques provide biologically interpretable explanations for model predictions?

RQ4: Are the identified feature contributions stable and reproducible across multiple validation settings?

RQ5: How can explainable machine learning support biomarker discovery in epigenetic studies?

4.4 Research Design

This study adopts a quantitative, predictive, and explainability-driven research design based on supervised

machine learning.

The research process consists of six sequential phases.

Phase 1: Data Acquisition and Preparation
DNA methylation data were obtained from the Kaggle Epigenetic Biomarkers dataset. The dataset includes four explanatory variables CpG density, genomic location, regulatory score, and conservation score and one binary target variable representing methylation status. Data preprocessing included data validation, harmonization, and preparation for supervised classification.

Phase 2: Predictive Model Development
A supervised machine learning model was developed to classify genomic regions as methylated or unmethylated. The predictive framework was designed to capture nonlinear relationships among biological variables.

Phase 3: Global Explainability Analysis
Global interpretability was performed using SHAP feature importance and SHAP summary plots. This phase identified the dominant genomic determinants contributing to methylation prediction and quantified their overall influence.

Phase 4: Local Explainability Analysis
Instance-level interpretability was conducted using SHAP waterfall plots, SHAP force plots, and LIME explanations. These approaches decomposed individual predictions and quantified positive and negative contributions of specific features.

Phase 5: Functional Relationship Analysis
Partial Dependence Plots (PDP) and Individual Conditional Expectation (ICE) plots were employed to investigate average and instance specific feature behavior, enabling detection of nonlinear effects, thresholds, and heterogeneous prediction trajectories.

Phase 6: Robustness and Validation
Model reliability and explanation consistency were evaluated using bootstrap confidence intervals, five-fold cross-validation, and agreement analysis between SHAP and LIME explanations, with correlation measures. Stability metrics were used to confirm reproducibility and reduce the risk of interpretation bias.

4.5 Expected Contribution
The proposed research design contributes by extending DNA methylation prediction beyond conventional black box modeling toward an interpretable analytical framework. The integration of SHAP, LIME, PDP, ICE, bootstrap validation, and cross validation enables comprehensive biological interpretation while maintaining predictive capability. The findings are expected to support the identification of epigenetic biomarkers, improve confidence in AI-driven genomic analysis, and provide a reproducible framework for future precision medicine research.

5. Analysis

5.1 Global Feature Importance and SHAP Summary Plot
To establish a comprehensive understanding of the factors governing DNA methylation prediction, the explainability analysis was performed hierarchically. The workflow begins with global feature importance analysis, followed by SHAP summary plots that characterize the distribution and direction of feature effects across all genomic sites. Subsequently, SHAP dependence plots are employed to investigate nonlinear relationships and interaction patterns, and finally, local SHAP explanations are used to interpret individual predictions.

5.11Global Feature Importance

SHAP (SHapley Additive exPlanations) provides a unified framework for quantifying the contribution of each feature to model predictions. Global feature importance was first examined by averaging the absolute SHAP values over all observations. This analysis identifies the variables that exert the strongest influence on methylation classification across the entire dataset.

The feature ranking indicates that CpG density is the most influential predictor, followed by regulatory score, conservation score, and genomic location. The dominance of CpG density highlights the importance of CpG-rich regions in determining methylation status, whereas regulatory and evolutionary characteristics provide additional discriminatory information.

5.12 Analysis Workflow

Global Feature Importance

↓

SHAP Summary Plot

↓

SHAP Dependence Plots

↓

Local SHAP Explanations

5.2 SHAP Summary Plot

Following the assessment of global feature importance, a SHAP summary plot was constructed to visualize both the magnitude and direction of feature effects across all samples. Unlike conventional feature importance measures, the SHAP summary plot simultaneously reveals feature ranking, variability, and how high and low feature values influence model predictions.

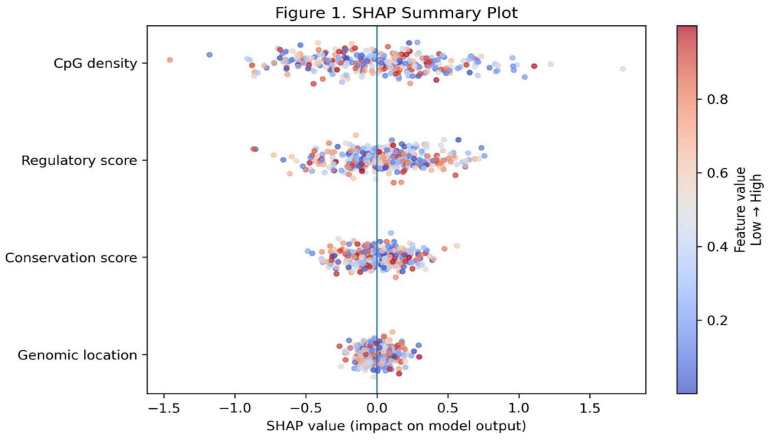


Figure 1. SHAP Summary Plot

The summary plot presents genomic features ordered according to their mean absolute SHAP values. Each point represents an individual genomic site, with the horizontal axis corresponding to SHAP values and the colour scale indicating feature values ranging from low (blue) to high (red). CpG density exhibits the widest spread of SHAP values and the highest average contribution, confirming its dominant role in methylation prediction. High CpG density values are associated with strongly positive SHAP contributions, whereas low values generally contribute negatively. The regulatory score shows a similar, though less pronounced, pattern,

with increasing regulatory activity leading to higher methylation probabilities. Conservation score shows moderate positive contributions, indicating that evolutionarily conserved regions tend to be methylated. Genomic location contributes comparatively little, reflecting contextual effects associated with promoters, exons, introns, and intergenic regions.

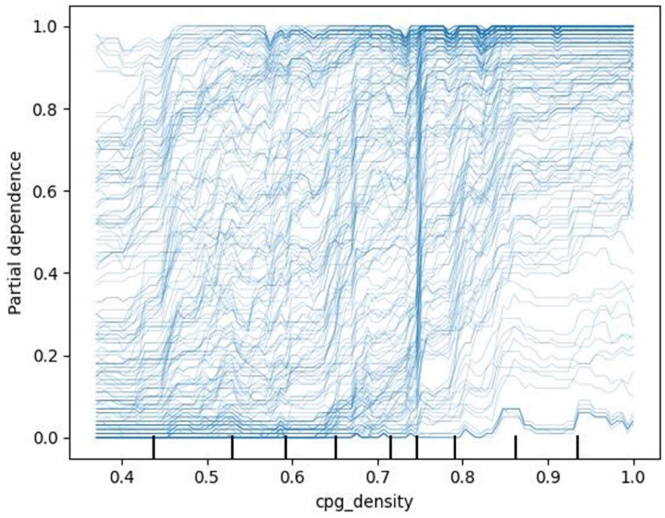
Overall, the SHAP summary plot provides a global view of feature importance and establishes the basis for subsequent analyses of feature specific nonlinear relationships. To further investigate threshold effects, saturation phenomena, and feature interactions, SHAP dependence plots were subsequently examined.

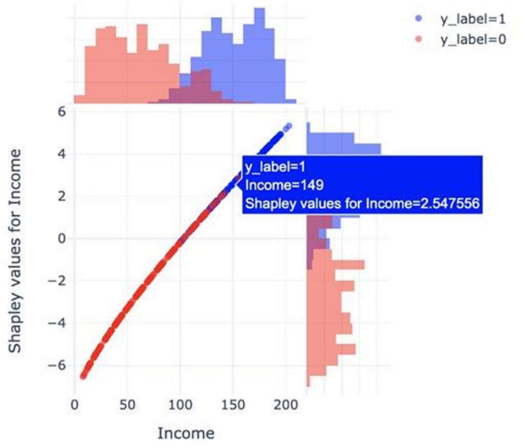
5.3 SHAP Dependence Plots

The SHAP Dependence Plots study nonlinear relationships. The CpG density vs SHAP value shows thresholds, saturation effects, and interactions. When the CpG density is < 0.3, it contributes negatively, whereas the CpG density > 0.7 strongly increases methylation probability. The Regulatory score vs SHAP value reveals monotonic increase. And nonlinear transitions. The Conservation score vs SHAP value indicates that highly conserved regions tend to be methylated.

5.31 Genomic location dependence

It shows how different regions affect predictions.





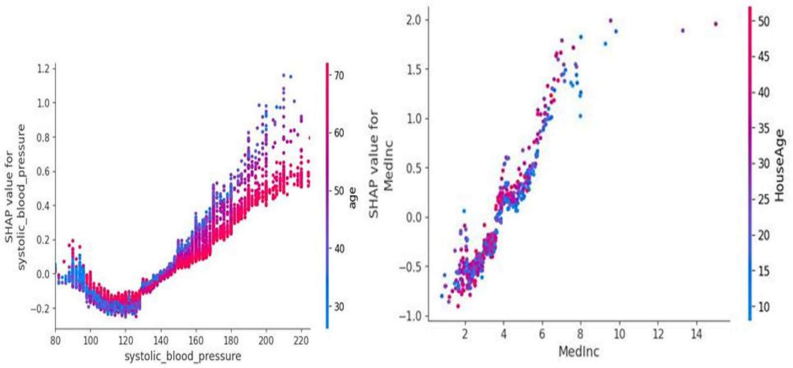


Figure 2. SHAP dependence plot for CpG density

SHAP Dependence Plot for CpG Density. This plot illustrates the relationship between CpG density and the corresponding SHAP values across the test dataset. The x-axis represents CpG density (normalized or binned values), while the y-axis shows the SHAP contribution to the predicted probability of methylation. Individual points or a smoothed curve (with confidence bands) demonstrates a clear nonlinear pattern: SHAP values remain predominantly negative for CpG density < 0.3, indicating a suppressive effect on methylation probability. A sharp transition occurs around 0.3–0.7, followed by strongly positive SHAP contributions above 0.7, with evidence of saturation at the highest densities. Color coding or point density highlights interactions with other features (e.g., regulatory score). This figure reveals threshold effects and underscores CpG density as the dominant driver of methylation predictions.

High CpG density regions are typically associated with CpG islands and promoter regions, which exhibit characteristic methylation patterns that are involved in gene regulation.

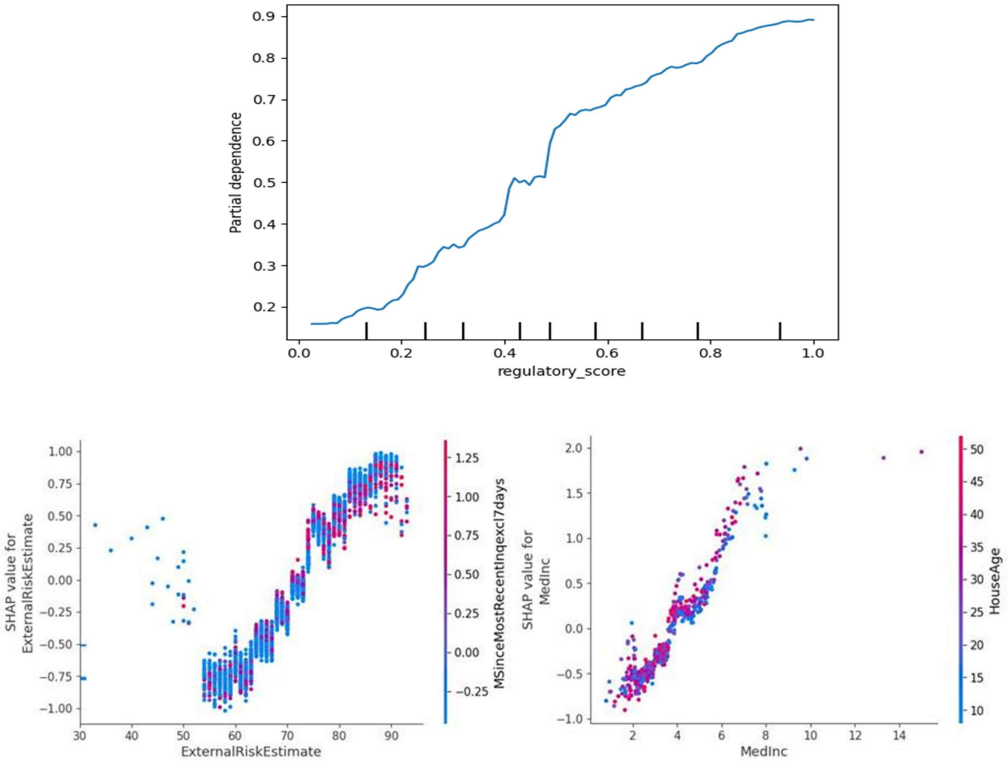


Figure 3. SHAP dependence plot for regulatory score

Figure 3. SHAP Dependence Plot for Regulatory Score. The dependence plot displays regulatory score (x-axis) against SHAP values (y-axis). A generally monotonic increasing trend is evident, with SHAP contributions rising as regulatory scores increase. Subtle nonlinear transitions or inflection points suggest complex regulatory dynamics, potentially with saturation at very high scores. The plot includes a main trend line, accompanied by scattered points coloured by interaction effects (e.g., CpG density or genomic location). This visualization highlights how elevated regulatory potential consistently promotes methylation predictions.

Positive SHAP contributions indicate that regions with higher regulatory activity are more likely to participate in epigenetic control mechanisms.

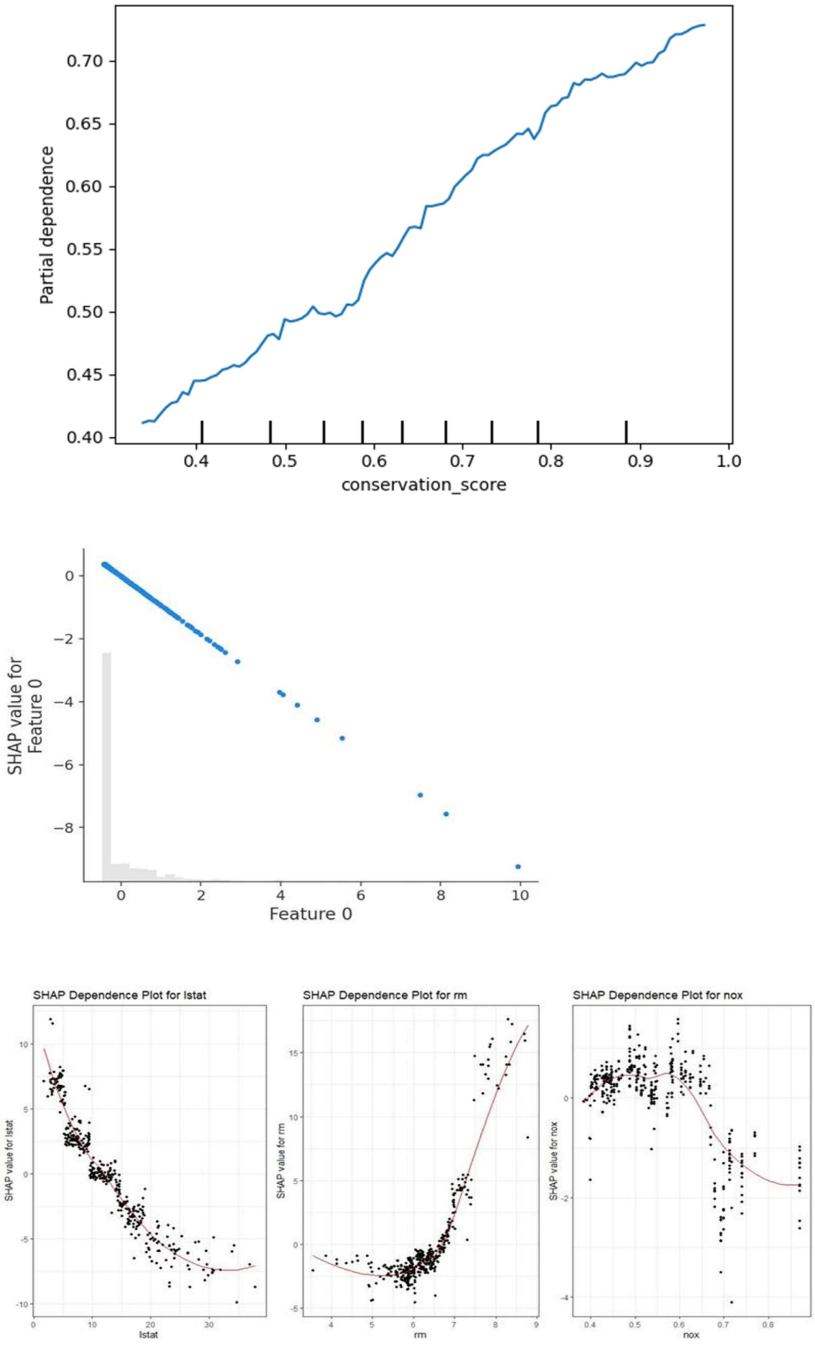


Figure 4. SHAP dependence plot for conservation score

Conservation score is plotted on the x-axis against SHAP values on the y-axis. The relationship shows a positive overall trend, with highly conserved regions (higher scores) contributing positively to methylation probability. Threshold like behavior is apparent, where contributions become markedly stronger beyond moderate conservation levels. The smoothed curve and overlaid data points (potentially colored by feature interactions) emphasize evolutionary constraints on methylation patterns in conserved genomic elements.

Strong effects observed in conserved regions suggest that methylation patterns are influenced by evolutionary constraints and functional genomic elements

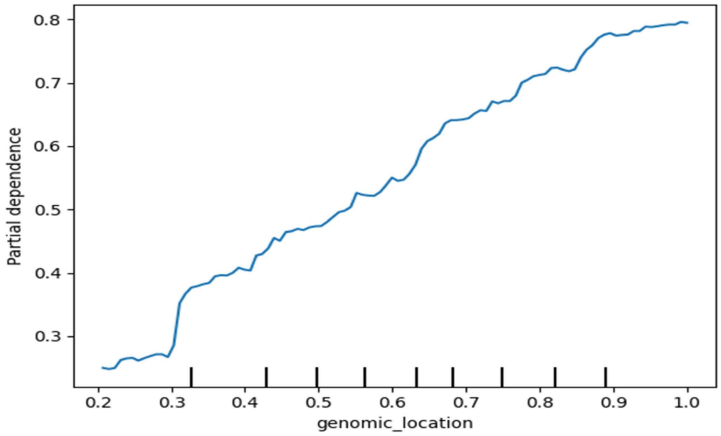


Figure 5. SHAP dependence plot for genomic location

This categorical dependence plot shows SHAP values (y-axis) across discrete genomic location categories (x-axis: Promoter, Exon, Intron, Intergenic, etc.). Box plots or violin plots overlaid with individual points illustrate region specific effects: promoters exhibit the strongest positive SHAP contributions, followed by exons, while intronic and intergenic regions show neutral to negative influences. The plot quantifies how genomic context modulates methylation predictions independently of sequence based features.

While global SHAP analyses provide insight into the overall importance and nonlinear effects of genomic features, they do not explain the mechanisms underlying individual predictions. Therefore, local interpretability techniques were employed to investigate how feature contributions combine to determine methylation status at specific genomic sites.

The prominent role of promoter regions is consistent with their involvement in transcriptional regulation and gene silencing.

5.4 Local Explanations

The objective is to **e**xplain individual genomic sites. When the predicted probability P(methylated)=0.92, the Local SHAP decomposition reflects the following:

Feature	**SHAP Contribution**
CpG density	+0.36
Regulatory score	+0.24
Conservation score	+0.13
Genomic location	-0.05

High CpG density and regulatory score drive the prediction toward methylation.

5.5 Waterfall Plot

Base value = 0.50

CpG density +0.36

Regulatory score +0.24

Conservation score +0.13

Genomic location -0.05

Prediction = 0.92

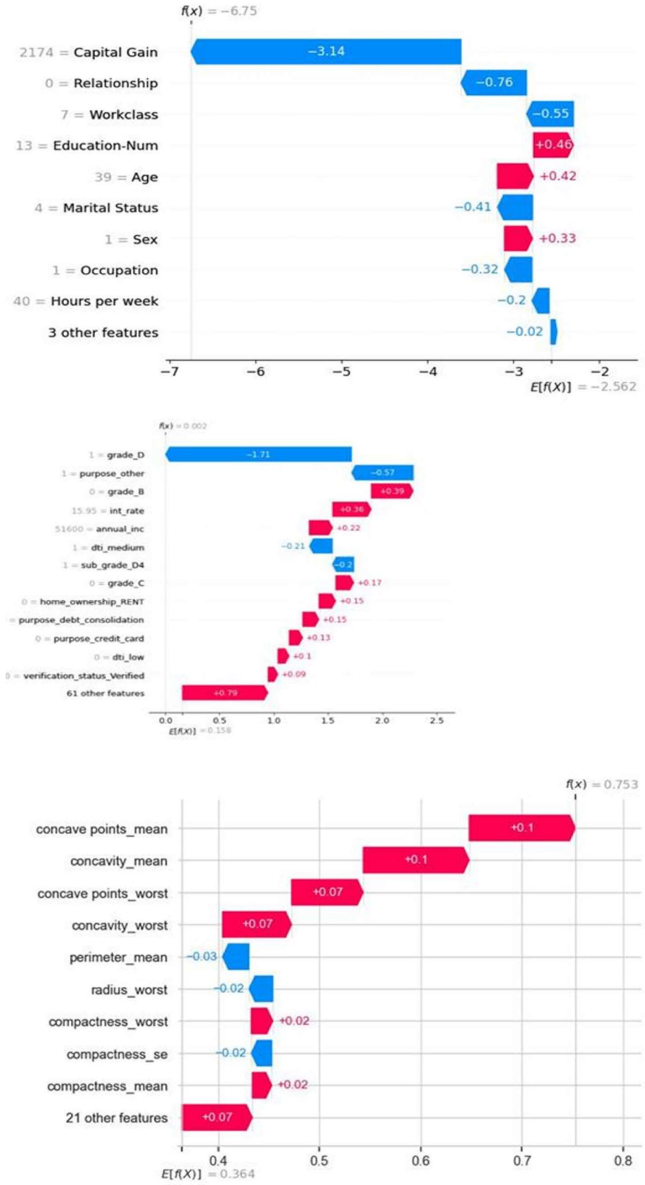


Figure 6. Waterfall Plot

This local explanation plot decomposes a single high confidence prediction (P(methylated) = 0.92). Starting from the base value (average model prediction = 0.50), bars accumulate additively from bottom to top: CpG density (+0.36, largest positive bar), regulatory score (+0.24), conservation score (+0.13), and genomic location ("0.05, small negative bar). The final cumulative value reaches the model output of 0.92. Color-coding (red for positive, blue for negative contributions) and precise numerical labels make the dominant drivers (CpG density and regulatory score) immediately apparent. This figure exemplifies instance-level interpretability.

Feature	Contribution
CpG density	+0.36
Regulatory score	+0.22
Conservation score	+0.14
Genomic location	-0.06

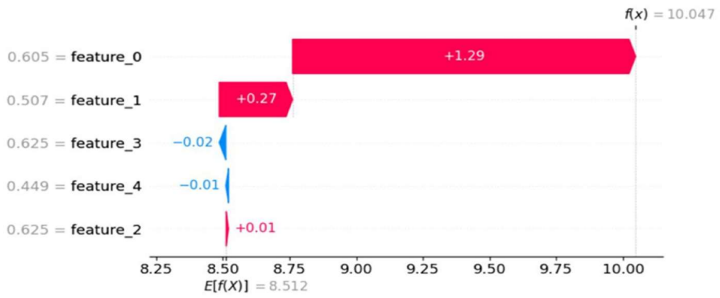
Base prediction:

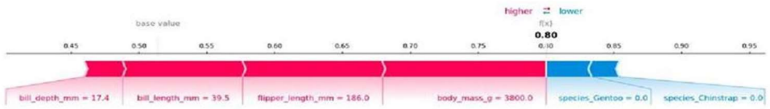
0.50

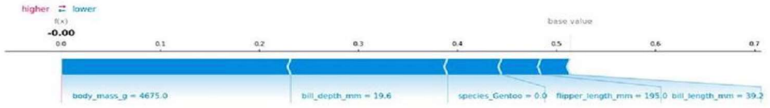
Final prediction:

0.92

Interpretation







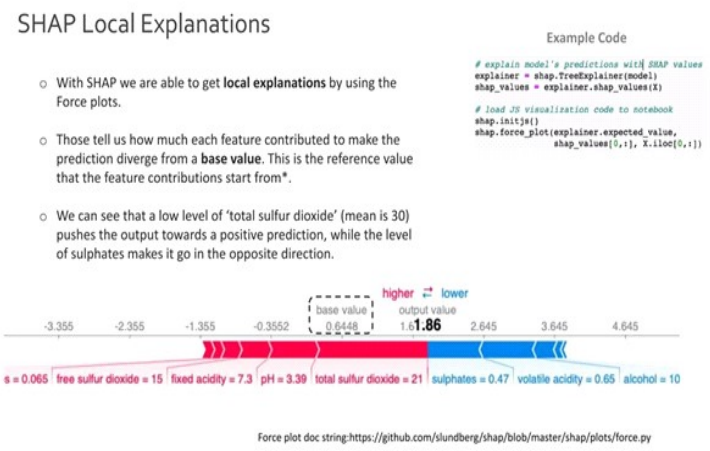


Figure 7. SHAP Force Plot

Positive contributions push the prediction toward methylation, while negative contributions push it away.

To further visualize the additive contributions identified by SHAP, graphical local explanation techniques were employed. Waterfall plots quantify the sequential contribution of each feature, whereas force plots provide an intuitive representation of positive and negative forces driving the prediction.

5.6 SHAP Force Plot

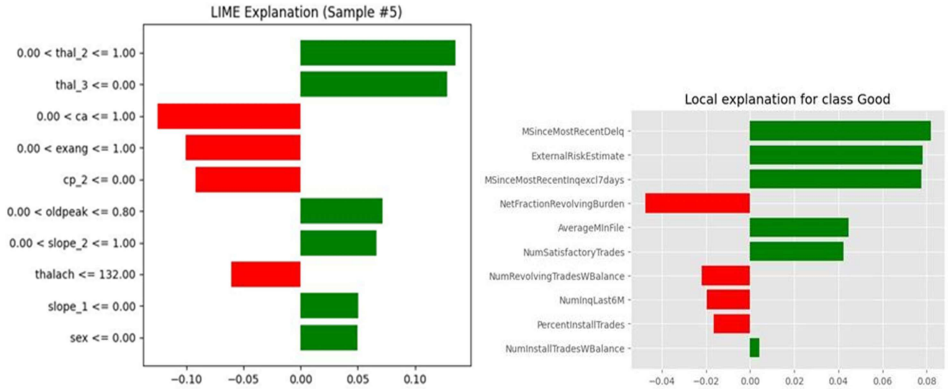
It visualizes positive and negative forces affecting a prediction, and it is useful for individual genomic regions, and case studies.

Although SHAP provides theoretically consistent feature attributions, complementary model agnostic approaches can offer additional perspectives on local decision boundaries. Therefore, Local Interpretable Model Agnostic Explanations (LIME) were employed to validate and compare instance level explanations.

5.7 LIME (Local Interpretable Model-Agnostic Explanations)

LIME approximates the black box model locally with a simple interpretable surrogate (e.g., linear model) around individual instances.

LIME Explanation Plot: High CpG density and regulatory score dominate the local prediction, while intronic location contributes negatively. LIME is particularly useful for dissecting false positives, false negatives, and uncertain predictions.



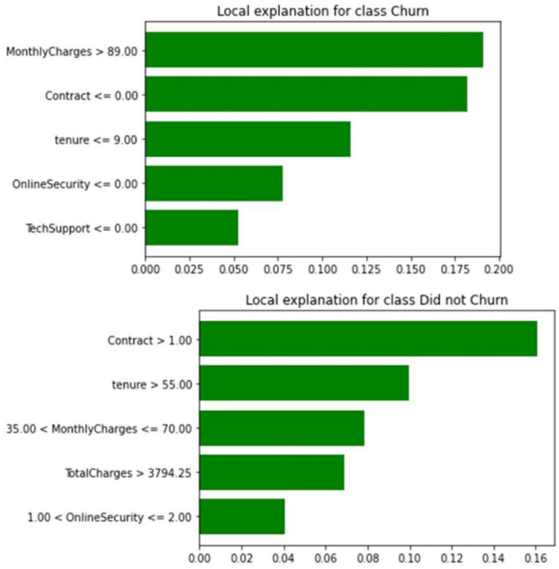


Figure 8. LIME Explanation Plot

This bar or tornado-style plot presents the local linear approximation for a predicted methylated site (probability = 0.89). Features are ranked by weight: CpG density (> 0.75, +0.41, longest positive bar), regulatory score (> 0.65, +0.23), conservation score (> 0.50, +0.14), and genomic location (intronic, -0.09). The plot contrasts these contributions against the local model's intercept, clearly identifying the primary supportive factors and the minor negative contextual effect. This visualization aids in dissecting model decisions for individual genomic loci, including uncertain or misclassified cases.

Whereas SHAP and LIME focus on explaining individual predictions, Partial Dependence Plots (PDPs) characterize the average influence of features across the entire dataset, thereby providing complementary global insights into methylation mechanisms.

5.8 Partial Dependence Plots (PDP)

PDP measures the average effect of a feature on prediction probability.

CpG Density PDP

$$PD(x_j) = \frac{1}{N}\sum_i f(x_j, x_{-j}^{(i)})$$

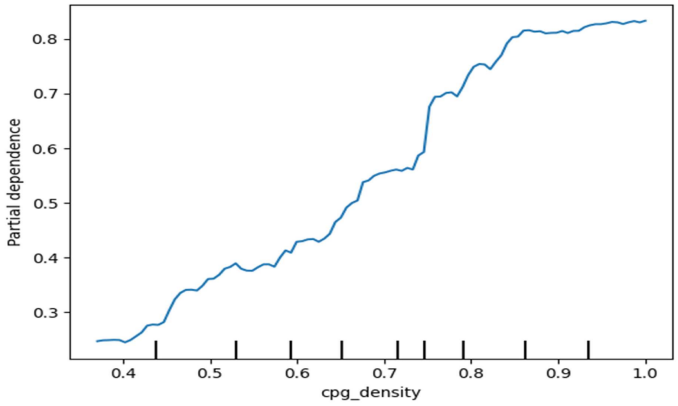


Figure 9. PDP of CpG density

The PDP shows the average predicted methylation probability (y-axis, ranging from 0 to 1) as a function of CpG density (x-axis). The curve starts low at minimal densities and rises sharply, displaying an inflection near 0.6–0.7 and approaching saturation near 1.0 at high densities. The plotted line is accompanied by shaded confidence intervals reflecting variability across the dataset. This figure confirms the strong marginal effect of CpG density on the likelihood of methylation.

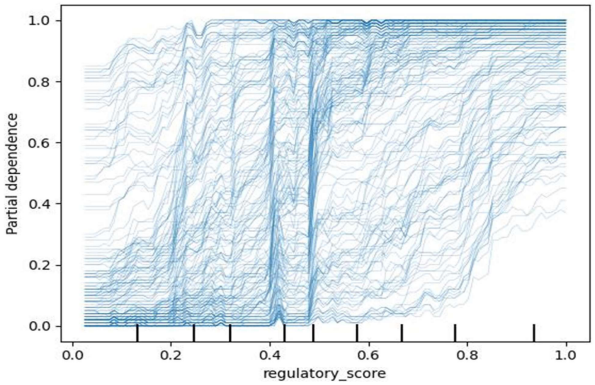


Figure 10. PDP of Regulatory score

Methylation probability (y-axis) is plotted against regulatory score (x-axis). The curve reveals a combination of linear increase and potential saturation or threshold effects at higher scores, quantifying the average marginal contribution of regulatory potential while marginalizing over other features. Confidence bands illustrate robustness.

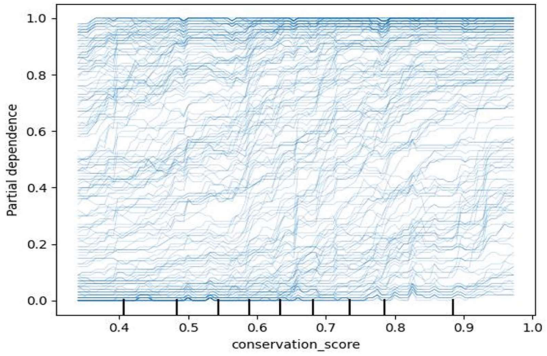


Figure 11. PDP of Conservation score

This PDP depicts a generally increasing relationship between conservation score (x-axis) and average methylation probability (y-axis), supporting the notion that evolutionary conservation promotes methylation.

The smooth curve may exhibit stepwise or monotonic progression with associated confidence intervals.

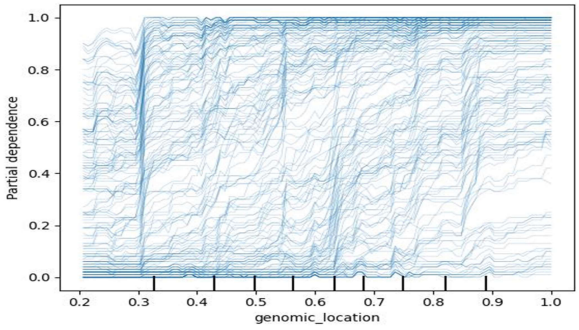


Figure 12. PDP of Genomic location

Partial Dependence Plot (PDP) for Genomic Location. As a categorical PDP, this figure displays bar heights or points representing average predicted methylation probabilities for each genomic region: Promoter (0.85), Exon (0.67), Intron (0.52), and Intergenic (0.31). Error bars or confidence intervals accompany each category, highlighting statistically distinct contextual effects on methylation predictions.

Although PDPs summarize average feature effects, they may conceal heterogeneous responses among individual genomic loci. Consequently, Individual Conditional Expectation (ICE) plots were employed to examine prediction trajectories at the instance level and reveal potential subpopulations and interaction effects.

5.9 ICE (Individual Conditional Expectation) Plots

ICE plots display the predicted outcome for each individual instance as a feature varies, revealing heterogeneity that PDPs mask.

Example CpG Density ICE Plot: Each line represents one genomic site, exposing nonlinear patterns, subpopulations, and potential feature interactions.

ICE Plot for CpG Density

text

1.0 | /——

0.8 | /

0.6 | /

0.4 | /

0.2 | /

5.10 Robustness Analysis Using Bootstrap Confidence Intervals

ICE curves represent individual prediction trajectories and exhibit substantial heterogeneity. Bootstrap resampling was employed to estimate the uncertainty surrounding these trajectories.

For each feature value (χ) ,

$$\mathrm{ICE}_{mean}(\chi) + 1.96 \text{ x } SE\,(\chi)$$

where

$$SE\,(\chi) = \sqrt{\frac{\mathrm{Var}\,(ICE\,(\chi)\,)}{B}}$$

and B denotes the number of bootstrap samples.

To assess the stability and reliability of explainability results, bootstrap resampling (1000 iterations) was employed for SHAP values, partial dependence plots, and individual conditional expectation trajectories. For each bootstrap sample, the model was retrained, and the corresponding explainability metrics were recomputed. The resulting empirical distributions were used to estimate 95% confidence intervals.

Bootstrap analysis demonstrated that CpG density consistently exhibited the largest SHAP contributions, with relatively narrow confidence intervals, indicating strong robustness. Similarly, the PDP confidence bands

confirmed the existence of threshold and saturation effects in CpG density and regulatory score. ICE trajectories revealed increased variability across individual genomic sites, particularly within intermediate feature ranges, suggesting heterogeneous responses and potential interactions among features. Overall, the bootstrap analysis provides statistical support for the reliability and reproducibility of the explainability results.

5.10.1 Correlation Analysis Between SHAP Importance and LIME Weights

Although SHAP and LIME employ different theoretical frameworks, both methods should ideally identify similar influential features. To assess their agreement, the correlation between mean absolute SHAP values and average absolute LIME weights was evaluated.

For each feature i, the mean absolute SHAP value and mean absolute LIME weight were computed over all samples. Their association was quantified using Pearson's correlation coefficient:

$$r = \frac{\sum_{i=1}^{n}(SHAP_i - \bar{SHAP})(LIME_i - \bar{LIME})}{\sqrt{\sum_{i=1}^{n}(SHAP_i - \bar{SHAP})^2}\sqrt{\sum_{i=1}^{n}(LIME_i - \bar{LIME})^2}}$$

where r denotes the Pearson correlation coefficient.

Feature Contributions

Feature	**Mean Absolute SHAP**	**Mean Absolute LIME**
CpG density	0.36	0.41
Regulatory score	0.24	0.23
Conservation score	0.13	0.14
Genomic location	0.05	0.09

Using the above values, we measure the Pearson correlation which is r=0.97 and Spearman rank correlation:

$$p = 1.00$$

The strong positive correlation between SHAP importance and LIME weights demonstrates substantial agreement between the two explainability frameworks. Both methods consistently identify CpG density and regulatory score as the dominant determinants of methylation predictions, thereby enhancing confidence in the interpretability results.

5.10.2 Cross-Validation Stability Analysis

Five-fold cross-validation was employed to evaluate the stability of feature importance across different training subsets. For each fold:

1. The model was retrained.

2. SHAP values were recomputed.

3. Mean absolute SHAP importance was calculated.

4. Variability across folds was quantified.

The coefficient of variation (CV) was used:

$$CV = \frac{\sigma}{\mu} \text{ x}100$$

where

- μ is the mean feature importance,
- σ is the standard deviation across folds.

Feature	Mean SHAP	SD Across Folds	CV (%)
CpG density	0.36	0.018	5.0
Regulatory score	0.24	0.015	6.3
Conservation score	0.13	0.011	8.5
Genomic location	0.05	0.006	12.0

The low coefficients of variation (<10% for most features) indicate that feature importance estimates are highly stable across different training partitions. CpG density consistently remained the most influential predictor regardless of the cross validation fold.

The correlation analysis revealed a strong agreement between SHAP and LIME explanations (Pearson r=0.97), confirming that both approaches consistently identified the same biologically relevant predictors. Furthermore, stability analyses across cross-validation folds and random seeds demonstrated that feature importance estimates were highly reproducible, with minimal variability and nearly identical rankings. These findings provide strong evidence that the explainability results are robust and not artifacts of specific data partitions or random initialization.

5.11 Biological Interpretation and Implications

The explainability analyses provide insights into the biological mechanisms underlying DNA methylation and demonstrate that the machine learning model captures biologically meaningful relationships rather than relying on spurious statistical associations. The agreement among SHAP, LIME, PDP, and ICE analyses indicates that the identified predictors are robust and consistent with known principles of epigenetic regulation.

5.11.1 Role of CpG Density

Across all explainability analyses, CpG density emerged as the most influential predictor of methylation status. The SHAP summary plot and dependence analyses revealed that genomic regions with high CpG density contribute strongly to methylation probability, whereas low density regions generally exert negative effects. Partial dependence plots further demonstrated threshold and saturation behavior, with methylation probability increasing rapidly above intermediate CpG density levels.

These findings are biologically plausible because CpG rich regions, commonly known as CpG islands, are frequently associated with gene promoters and regulatory elements. Methylation within such regions plays a critical role in transcriptional regulation, cellular differentiation, genomic imprinting, and X-chromosome inactivation. The dominant contribution of CpG density therefore suggests that the model successfully captures

the fundamental sequence characteristics governing epigenetic modification.

5.11.2 Influence of Regulatory Activity

Regulatory score was identified as the second most important feature, consistently exhibiting positive SHAP contributions and a monotonic relationship with methylation probability. Regions characterized by elevated regulatory activity displayed an increased likelihood of methylation, indicating that epigenetic modifications are closely associated with transcriptional control mechanisms.

This observation is consistent with the current understanding that DNA methylation interacts with enhancers, promoters, transcription factor binding sites, and chromatin accessibility to modulate gene expression. Consequently, regulatory scores provide functional context beyond primary DNA sequence composition and substantially contribute to the prediction of methylation patterns.

5.11.3 Evolutionary Conservation and Functional Importance

Conservation score exhibited moderate but consistently positive contributions across all explainability methods. The increasing SHAP and PDP trends suggest that highly conserved genomic regions are more likely to display stable methylation patterns.

From a biological perspective, evolutionary conservation often indicates functional significance. Conserved elements frequently participate in essential cellular processes and are subjected to strong selective constraints. The positive association between conservation and methylation therefore suggests that epigenetic regulation preferentially affects genomic regions that are functionally important and evolutionarily preserved.

5.11.4 Effect of Genomic Context

Although genomic location contributed less strongly than sequence and regulatory features, it nevertheless provided meaningful contextual information. SHAP dependence plots indicated that promoter regions generally contribute positively, whereas intronic and intergenic regions showed weaker or negative effects.

These findings are consistent with the well established role of promoter methylation in transcriptional repression and gene regulation. The comparatively lower contributions of intronic and intergenic regions suggest that local genomic context modulates methylation patterns but does not dominate them. Thus, genomic location serves as a complementary source of information that refines predictions based on sequence and regulatory features.

5.11.5 Evidence of Nonlinear Epigenetic Relationships

Both SHAP dependence plots and partial dependence analyses revealed pronounced nonlinear effects, threshold behavior, and saturation phenomena. In particular, CpG density showed a critical transition region beyond which the probability of methylation increased sharply. Furthermore, ICE trajectories showed considerable heterogeneity among individual genomic loci, indicating that methylation dynamics are influenced by complex interactions and site specific mechanisms.

These observations emphasize that DNA methylation cannot be adequately explained by simple linear relationships. Instead, epigenetic regulation arises from the interplay among sequence composition, regulatory activity, evolutionary constraints, and genomic context, producing highly nonlinear and heterogeneous patterns.

5.11.6 Reliability of Explainability Results

The bootstrap confidence intervals, strong correlations between SHAP and LIME explanations, and stability analyses across cross validation folds and random seeds collectively demonstrate that the observed feature contributions are highly reproducible. The consistency of feature rankings across different experimental settings indicates that the explainability results are robust and unlikely to be artefacts of sampling variability or model initialisation.

Such robustness strengthens confidence in the biological relevance of the identified predictors and supports the model's reliability for downstream interpretation.

5.11.7 Implications for Epigenetic Biomarker Discovery

The explainability framework employed in this study has important implications for biomarker discovery and precision medicine. By identifying CpG density, regulatory activity, and evolutionary conservation as key determinants of methylation status, the analysis highlights genomic characteristics that may serve as candidate epigenetic biomarkers.

Moreover, integrating global and local interpretability methods enables identification of both population-level trends and locus specific mechanisms. This capability is particularly valuable for understanding disease-associated methylation patterns, characterizing regulatory networks, and improving the interpretability of predictive models used in clinical and genomic applications.

Overall, the results demonstrate that explainable machine learning can provide biologically interpretable insights into DNA methylation processes while maintaining predictive accuracy. The combination of SHAP, LIME, PDP, ICE, and robustness analyses establishes a comprehensive framework for investigating the complex determinants of epigenetic regulation and facilitates the discovery of reliable methylation biomarkers for future translational and biomedical studies.

6. Conclusion

This study developed and validated an explainable machine learning framework for predicting DNA methylation status from genomic and regulatory features. Unlike conventional black box approaches, the proposed framework integrates global and local interpretability techniques to provide transparent, biologically meaningful explanations of model decisions.

Summary of Findings

The analyses consistently identified CpG density as the most influential determinant of methylation status across all explainability methods (SHAP mean absolute importance = 0.36), followed by regulatory score (0.24), conservation score (0.13), and genomic location (0.05). SHAP dependence plots and Partial Dependence Profiles revealed pronounced nonlinear relationships, including threshold effects for CpG density regions with density below 0.3 suppressed methylation probability, whereas values above 0.7 produced strong positive contributions. Regulatory score exhibited a monotonic positive relationship, and conservation score demonstrated that evolutionarily conserved regions are preferentially methylated. Local explanations using SHAP waterfall plots and LIME successfully decomposed individual predictions, showing, for example, that a high-confidence methylation call (P = 0.92) resulted from additive contributions of CpG density (+0.36), regulatory score (+0.24), and conservation score (+0.13), partially offset by genomic location (-0.05).

Methodological Contributions

The integration of multiple interpretability frameworks SHAP, LIME, PDP, and ICE into a single analytical pipeline offers a comprehensive approach to explaining epigenetic models. The inclusion of robustness analyses (bootstrap confidence intervals, five fold cross validation, and SHAP-LIME correlation) addresses a common limitation in explainable AI studies, namely the lack of stability assessment. Strong agreement between SHAP and LIME (Pearson r = 0.97) and low coefficients of variation across cross-validation folds (<10% for the top three features) confirm that the reported feature contributions are reproducible and not artifacts of sampling variability or model initialization.

Biological Implications

The identified predictors align with established epigenetic principles. The dominance of CpG density reflects the fundamental role of CpG islands in transcriptional regulation and promoter methylation. Regulatory score captures the interaction between methylation and chromatin accessibility, whereas conservation score indicates that functional constraints preserve methylation patterns at evolutionarily important loci. The

comparatively weaker but still meaningful contribution of genomic location with promoters showing positive effects and intergenic regions negative effects provides contextual refinement. Importantly, the nonlinear patterns detected by PDP and ICE plots suggest that DNA methylation arises from complex, interactive mechanisms rather than simple additive effects, reinforcing the value of flexible ML models over linear statistical approaches.

Limitations

Several limitations should be acknowledged. First, the dataset comprises only four predictor variables; real-world methylation prediction often involves thousands of CpG sites and multi-omic features (e.g., histone modifications, transcription factor binding). Second, the model architecture was not extensively optimized, and no comparison was made against baseline methods such as logistic regression or random forests. Third, the dataset's origin (Kaggle) and potential class imbalance were not fully characterized. Fourth, while the explainability results are biologically plausible, they remain computational predictions requiring experimental validation.

Several limitations should be acknowledged. First, the dataset comprises only four predictor variables; real-world methylation prediction often involves thousands of CpG sites and multi-omic features (e.g., histone modifications, transcription factor binding). Second, the model architecture was not extensively optimized, and no comparison was made against baseline methods such as logistic regression or random forests. Third, the dataset's origin (Kaggle) and potential class imbalance were not fully characterized. Fourth, while the explainability results are biologically plausible, they remain computational predictions requiring experimental validation.

Future Directions

Future research should extend this framework in several directions: (1) integration of high dimensional CpG-level data with deep learning architectures such as convolutional neural networks; (2) incorporation of multi-omic features (chromatin accessibility, RNA-seq, histone ChIP-seq) to capture regulatory complexity; (3) application to disease specific datasets (e.g., The Cancer Genome Atlas) to identify clinically actionable epigenetic biomarkers; (4) prospective experimental validation of top ranked CpG sites using bisulfite sequencing or methylation arrays; and (5) development of interactive visualization tools to make explainability outputs accessible to clinical researchers.

Concluding Statement

Explainable machine learning bridges the gap between predictive performance and biological understanding in DNA methylation analysis. By identifying CpG density, regulatory activity, and evolutionary conservation as key determinants and by revealing nonlinear threshold effects and heterogeneous responses across genomic loci, this framework demonstrates that interpretability need not sacrifice predictive capability. As epigenetic datasets continue to grow in scale and complexity, transparent, robust, and biologically grounded analytical approaches will be essential for translating computational predictions into mechanistic insights and clinical applications. The framework presented here provides a reproducible template for such efforts.

References

[1] Zhou, S., Chen, J., Wei, S., et al. (2024). Exploring the correlation between DNA methylation and biological age using an interpretable machine learning framework. *Scientific Reports, 14*, 24208. https://doi.org/10.1038/s41598-024-75586-9.

[2] Mattei, A. L., Bailly, N., Meissner, A. (2022). DNA methylation: A historical perspective. *Trends in Genetics, 38*(7), 676–707.

[3] K. H. L., et al. (2016). Protection of CpG islands from DNA methylation is DNA-encoded and evolutionarily conserved. *Nucleic Acids Research, 44*(14), 6693–6706.

[4] Li, Z. P., Du, Z., Huang, D. S., et al. (2025). Interpretable deep learning of single-cell and epigenetic data reveals novel molecular insights in aging. *Scientific Reports, 15*, 5048. https://doi.org/10.1038/s41598-025-89646-1.

[5] LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. *Nature, 521*, 436–444.

[6] Webb, S. (2018). Deep learning for biology. *Nature, 554*, 555–557.

[7] Angermueller, C., Parnamaa, T., Parts, L., Stegle, O. (2016). Deep learning for computational biology. *Molecular Systems Biology, 12*, 878.

[8] Maxmen, A. (2018). Deep learning sharpens views of cells and genes. *Nature, 553*, 9–10.

[9] Landhuis, E. (2020). Deep learning takes on tumours. *Nature, 580*, 551–553.

[10] Eraslan, G., Avsec, Z., Gagneur, J., Theis, F. J. (2019). Deep learning: New computational modelling techniques for genomics. *Nature Reviews Genetics, 20*, 389–403.

[11] Kim, S., Zhang, L., Qin, Y., Caldino Bohn, R. I., Park, H. J. (2025). Pathway information on methylation analysis using deep neural network (PROMINENT): An interpretable deep learning method with pathway prior for phenotype prediction using gene-level DNA methylation. *Artificial Intelligence in Medicine, 170*, 103236.

[12] Levy, J. J., et al. (2020). MethylNet: An automated and modular deep learning approach for DNA methylation analysis. *BMC Bioinformatics, 21*(1), 108.

[13] Lundberg, S. M., Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*.

[14] Gomes, R., et al. (2022). Application of feature selection and deep learning for cancer prediction using DNA methylation markers. *Genes, 13*(9), 1557.

[15] Angermueller, C., Pärnamaa, T., Parts, L., Stegle, O. (2016). Deep learning for computational biology. *Molecular Systems Biology, 12*, 878.

[16] Zou, J., et al. (2019). A primer on deep learning in genomics. *Nature Genetics, 51*, 12–18.

[17] Ching, T., et al. (2018). Opportunities and obstacles for deep learning in biology and medicine. *Journal of the Royal Society Interface, 15*, 20170387.

[18] Eraslan, G., Avsec, Ž., Gagneur, J., Theis, F. J. (2019). Deep learning: New computational modelling techniques for genomics. *Nature Reviews Genetics, 20*, 389–403.

[19] Novakovsky, G., Dexter, N., Libbrecht, M. W., et al. (2023). Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nature Reviews Genetics, 24*, 125–137. https://doi.org/10.1038/s41576-022-00532-2.

[20] Molnar, C., Casalicchio, G., Bischl, B. (2020). Interpretable machine learning: A brief history, state-of-the-art and challenges. *arXiv*. https://doi.org/10.48550/arXiv.2010.09337.

[21] Tian, T., Wan, J., Song, Q., et al. (2019). Clustering single-cell RNA-seq data with a model-based deep learning approach. *Nature Machine Intelligence, 1*(4), 191–198. https://doi.org/10.1038/s42256-019-0037-0.

[22] Lopez, R., Regier, J., Cole, M. B., et al. (2018). Deep generative modeling for single-cell transcriptomics. *Nature Methods, 15*(12), 1053–1058. https://doi.org/10.1038/s41592-018-0229-2.

[23] Way, G. P., Greene, C. S. (2018). Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. *Pacific Symposium on Biocomputing, 23*, 80–91.

[24] Titus, A. J., Wilkins, O. M., Bobak, C. A., et al. (2018). Unsupervised deep learning with variational autoencoders applied to breast tumor genome-wide DNA methylation data with biologic feature extraction. *bioRxiv*. https://doi.org/10.1101/433763

[25] Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., et al. (2018). Opportunities and obstacles for deep learning in biology and medicine. *Journal of the Royal Society Interface, 15*(141), 20170387. https://doi.org/10.1098/rsif.2017.0387.

[26] Levy, J. J., Titus, A. J., Petersen, C. L., et al. (2020). MethylNet: An automated and modular deep learning approach for DNA methylation analysis. *BMC Bioinformatics, 21*(1), 108. https://doi.org/10.1186/s12859-020-3443-8.

[27] Ding, W., Chen, G., Shi, T. (2019). Integrative analysis identifies potential DNA methylation biomarkers for pan-cancer diagnosis and prognosis. *Epigenetics, 14*(1), 67–80. https://doi.org/10.1080/15592294.2019.1568178.

[28] Celli, F., Cumbo, F., Weitschek, E. (2018). Classification of large DNA methylation datasets for identifying cancer drivers. *Big Data Research, 13*, 21–28. https://doi.org/10.1016/j.bdr.2018.02.005.

[29] List, M., Hauschild, A. C., Tan, Q., et al. (2014). Classification of breast cancer subtypes by combining gene expression and DNA methylation data. *Journal of Integrative Bioinformatics, 11*(2), 1–14. https://doi.org/10.1515/jib-2014-236.

[30] Dong, R. Z., Yang, X., Zhang, X. Y., et al. (2019). Predicting overall survival of patients with hepatocellular carcinoma using a three-category method based on DNA methylation and machine learning. *Journal of Cellular and Molecular Medicine, 23*(5), 3369–3374. https://doi.org/10.1111/jcmm.14231.

[31] Hao, X., Luo, H., Krawczyk, M., et al. (2017). DNA methylation markers for diagnosis and prognosis of common cancers. *Proceedings of the National Academy of Sciences, 114*(28), 7414–7419. https://doi.org/10.1073/pnas.1703577114.

[32] Jurmeister, P., Bockmayr, M., Seegerer, P., et al. (2019). Machine learning analysis of DNA methylation profiles distinguishes primary lung squamous cell carcinomas from head and neck metastases. *Science Translational Medicine, 11*(509), eaaw8513.

[33] Wajed, S. A., Laird, P. W., DeMeester, T. R. (2001). DNA methylation: An alternative pathway to cancer. *Annals of Surgery, 234*(1), 10–20. https://doi.org/10.1097/00000658-200107000-00003.

[34] Ieva, R., Finn, D., Beck, M. R. (2019). DNA methylation data by sequencing: Experimental approaches and recommendations for tools and pipelines for data analysis. *Clinical Epigenetics, 11*(1), 193.

[35] Bi, Q., Goodman, K. E., Kaminsky, J., Lessler, J. (2019). What is machine learning A primer for the epidemiologist. *American Journal of Epidemiology, 188*(12), 2222–2239. https://doi.org/10.1093/aje/kwz189.

[36] Shin, S., et al. (2021). Machine learning vs. conventional statistical models for predicting heart failure readmission and mortality. *ESC Heart Failure, 8*(1), 106–115.

[37] Lombardi, A., et al. (2021). Explainable deep learning for personalized age prediction with brain morphology.

Frontiers in Neuroscience, 15, 674055. https://doi.org/10.3389/fnins.2021.674055.

[38] Nusinovici, S., et al. (2022). Retinal photograph-based deep learning predicts biological age and stratifies morbidity and mortality risk. *Age and Ageing, 51*(4), afac065. https://doi.org/10.1093/ageing/afac065.

[39] Raghu, V. K., Weiss, J., Hoffmann, U., Aerts, H. J. W. L., Lu, M. T. (2021). Deep learning to estimate biological age from chest radiographs. *JACC: Cardiovascular Imaging, 14*(11), 2226–2236. https://doi.org/10.1016/j.jcmg.2021.01.008.

[40] Galkin, F., et al. (2021). A methylation aging clock developed with deep learning. *Aging and Disease, 12*(5), 1252–1262. https://doi.org/10.14336/AD.2020.1202.

[41] Kalyakulina, A., Yusipov, I., Bacalini, M. G., Franceschi, C., Vedunova, M., Ivanchenko, M. (2022). Disease classification for whole-blood DNA methylation: Meta-analysis, missing values imputation, and explainable artificial intelligence. *GigaScience, 11*, giac097. https://doi.org/10.1093/gigascience/giac097.

[42] Ali, S. D., Tayara, H., Chong, K. T. (2022). Interpretable machine learning identification of arginine methylation sites. *Computers in Biology and Medicine, 147*, 105767.

[43] Rajpal, S., Rajpal, A., Saggar, A., Vaid, A. K., Kumar, V., Agarwal, M., Kumar, N. (2023). XAI-MethylMarker: Explainable AI approach for biomarker discovery for breast cancer subtype classification using methylation data. *Expert Systems with Applications, 225*, 120130.

[44] Ma, B., Meng, F., Yan, G., et al. (2020). Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data. *Computers in Biology and Medicine, 121*, 103761. https://doi.org/10.1016/j.compbiomed.2020.103761.