IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.156106

From Clinical Risk to Biological Aging - A Modular AI Framework for Traceable CpG Biomarker Prioritization

Suresh Ramchandra Kaulagi[1], Dr. Hariram Chavan[2]

Department of Information Technology, K.J. Somaiya Institute of Technology, Sion, Mumbai, India[1]

Department of Information Technology, K.J. Somaiya Institute of Technology, Sion, Mumbai, India [2]

Abstract: Artificial intelligence is making real inroads in epigenetics, especially when it comes to using DNA methylation to dig into aging and disease. Predictive models do a decent job at flagging CpG sites tied to clinical outcomes, but there's a problem—they're often black boxes. You get predictions, but the biological context stays hidden. That disconnect makes it tough to reproduce findings or pull out insights that actually matter for biomedicine. That's what pushed us to build a modular framework that links CpG selection straight to biological annotation. Instead of predictions floating on their own, you can track them back to the biology of aging.

We worked with a Canadian methylation cohort—92 samples, over 485,000 CpG sites. After normalizing and filtering the data, we ran log-rank tests, Random Survival Forests, Cox proportional hazards models, and ensemble learners to find CpGs tied to simulated survival outcomes. SHAP values helped us spotlight the CpGs that reliably drove predictions. Then we mapped these CpGs to enhancer regions, transcription factor motifs, and downstream pathways. Visualizations made it clear how each marker connects to regulatory elements and broader biological processes.

Our framework delivered stable concordance indices across training, test, and cross-validation sets. Certain CpGs kept showing up as influential and linked back to genes like IL7R and CDKN2A—core players in immune aging and cell-cycle regulation. Because we combined modeling with annotation, it's easier to trace how methylation changes might actually influence biological pathways.

In the end, this approach gives researchers a practical, interpretable way to prioritize CpG biomarkers. By weaving together statistical modeling and enhancer-aware annotation, our framework nudges methylation research toward transparency and real biological relevance, especially in aging and personalized medicine.

Keywords: Artificial Intelligence, Bioinformatics, DNA methylation, CpG biomarkers, aging biology, artificial intelligence, survival modeling, SHAP, enhancer annotation, epigenetics

I. INTRODUCTION

Artificial intelligence has become a regular part of biomedical research [1]. Over the past decade, AI has been everywhere—from cancer diagnostics to studies on aging—spotting patterns people just can't see by hand [2]. In epigenetics, DNA methylation stands out as a goldmine [3]. It lets researchers estimate biological age, gauge disease risk, and track regulatory changes as we get older. Still, even with all this progress, most computational models act like black boxes [4]. They flag certain CpG sites as important, but rarely tell us why these sites matter on a biological level. That's a sticking point. Without a clear explanation, it's tough to reproduce results or turn them into something useful for patients [5].

This problem gets even trickier when you look at aging across the whole system. Methylation patterns change from one tissue to another, and without a solid regulatory map, generalizing results or connecting specific CpG changes to bigger biological processes is a real challenge. Recent studies call out how machine learning often ignores enhancer regions, transcription factor motifs, and relationships between CpGs and genes [6]. Because of that, biomarkers end up as little more than statistical blips—not real, biologically meaningful markers [7]. You see this gap again and again in cancer, fertility, and aging research. Tools that actually connect enhancer regions or offer clear explanations are still used inconsistently [8].

To get a better sense of where things stand, we looked at recent work blending AI and epigenetics. The studies show both huge potential and big limitations. Models often predict accurately, but rarely bring in enhancer logic or offer much interpretability [9]. Some focus on a single disease and leave out the larger regulatory context [10]. So, the message from the literature is clear: we need frameworks that blend prediction with biological annotation and actually make sense to researchers.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.156106

Table 1 below shows analysis of key literature papers.

TABLE 1. KEY LITERATURE PAPERS

Author / Year	Publication	Summary	Gap Identified	Project Objective
Sharma et al. (2025)	Epigenetics & Chromatin	Review of AI/ML in pan-cancer methylation profiling and interpretability challenges	Limited enhancer annotation; poor early-stage sensitivity; lack of explainable AI	Build enhancer-aware, interpretable ML pipelines for pan-cancer detection
Orgueira et al. (2024)	Clinical Epigenetics	ML-based methylation profiling for relapse prediction in pediatric leukemia	No enhancer annotation; lacks SHAP-style interpretability	Apply SHAP-guided CpG selection with enhancer annotation
Chavda et al. (2024)	Frontiers in Medicine	Bibliometric study of AI in healthcare, highlighting disciplinary silos	Epigenetics underrepresented; citation fragmentation	Position epigenetics-AI research within broader healthcare AI trends
Soubry (2025)	Frontiers in Reproductive Health	Machine learning applied to sperm methylation for fertility prediction	Weak generalizability; male-specific epigenetic traits underused	Develop generalizable, biologically informed models for fertility outcomes
Rao & Kalra (2025)	HSOA Journal of Integrative Medicine	Examined yoga/meditation effects on methylation markers	No AI modeling; limited quantitative analysis	Apply interpretable ML to quantify behavioral epigenetic shifts

That's what drove us to start this project. We set out to build a modular AI system that does more than just spot CpG biomarkers tied to clinical outcomes—it tracks those markers back to enhancer regions, transcription factor motifs, and downstream pathways. The goal? To put the statistical signals and the biological meaning side by side, so researchers get the full picture. By bringing together survival modeling, SHAP-based feature attribution, and enhancer-aware annotation, we're aiming for an approach to methylation analysis that's both interpretable and rooted in real biology, especially when it comes to aging.

II. METHODS

Study Design

We set out to create a modular AI framework to prioritize CpG biomarkers in DNA methylation data. Our goal wasn't just to find which CpGs relate to survival outcomes—we wanted to map them back to enhancer regions, transcription factor motifs, and the downstream pathways they influence. Everything happened on the computer: we worked with publicly available methylation data and built reproducible Python pipelines from the ground up. We ran all analyses in Jupyter Notebooks [22], so anyone can follow each step and see exactly what we did.

Dataset and Preprocessing

For this project, we pulled data from GSE38235 in the Gene Expression Omnibus [11]. This dataset features methylation profiles from a Canadian cohort—92 samples in total. Each sample comes with raw signal intensities for both methylated and unmethylated probes, covering 485,577 CpG sites altogether. That's the foundation we used for our analysis.

Beta values were calculated from raw intensities and normalized to reduce technical variation [12]. We removed CpGs with low variance or excessive missingness, as these tend to add noise without contributing meaningful biological information. To narrow the feature space further, we applied log-rank survival tests and retained 9,515 CpGs that showed some association with simulated survival outcomes [13].

Simulation of Clinical Outcomes

Since the dataset lacks actual survival data, we created simulated survival times between 6 and 120 months. We assigned event status using methylation levels—samples in the top 30% for chosen CpGs counted as either relapse or death. Simulated outcomes don't stand in for real clinical results, but they let us test the modeling framework in a controlled environment and guarantee enough events for solid survival analysis.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.156106

Modelling Workflow

We used a combination of traditional and machine-learning-based survival models. Random Survival Forests (RSF) served as the primary learner [14], with both conservative and relaxed parameter settings to explore how tree depth and flexibility influence performance. Cox proportional hazards (CoxPH) models [15] were used both as standalone comparators and as meta-learners in stacked configurations.

We brought in LightGBM and XGBoost as base learners [16] and used Random Forest as our meta-model for the ensemble setups. This mix let us put different modeling philosophies—parametric, non-parametric, and gradient-boosting—side by side on the same methylation dataset.

Feature Attribution and Annotation

To figure out which CpGs actually drove survival predictions, we leaned on SHAP (SHapley Additive exPlanations) values. SHAP quantifies how much each feature shapes the model's output, so it cuts through the noise and highlights CpGs that really matter, no matter what modeling setup we use.

Next, we mapped these CpGs to enhancer regions with FANTOM5 [18] and Ensembl regulatory tracks [21]. We used MACS2 peak calls along with JASPAR motif libraries [19] to spot transcription factor motifs. This step bridges the gap between statistical signals and regulatory biology, something a lot of methylation studies ignore when they just focus on prediction.

For visualization, we built Sankey diagrams that show how CpGs connect to enhancers, genes, and biological pathways. Pybedtools handled the genomic interval operations [17], making it a lot simpler to pull together all these layers of annotation.

Figure 1 below shows our modular AI Framework.

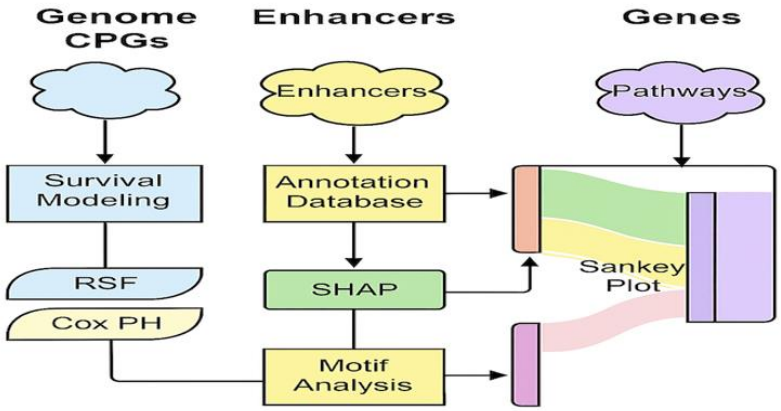

Figure 1. Modular AI Framework for traceable CpG Biomarker prioritization

Statistical Analysis

Model performance was evaluated using concordance indices (C-index) for training, test, out-of-bag, and cross-validation sets. We also examined RSF tree depth distributions to understand how model complexity changed under different parameter settings. A delta-disagreement score was computed across classifiers to highlight regions where predictions diverged, providing a simple measure of uncertainty.

Given the limited sample size, we relied on simulated outcomes to ensure sufficient event rates. While this approach has limitations, it allowed us to test the framework's behavior under controlled conditions.

III. RESULTS

Overview of Findings

We put the modular AI framework through its paces, testing it with different survival modeling setups, feature attribution methods, and rounds of biological annotation. The story stayed pretty steady across all these runs. The models held up well, the key CpGs kept showing up no matter how we sliced things, and the biological annotations matched what's already known about aging and immune pathways. Let's dig into each part of the analysis.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.156106

Dataset Summary

The GSE38235 dataset started with 92 samples and over 485,000 CpG sites. After cleaning and filtering, we narrowed things down to 9,515 CpGs for the actual modelling [20]. Table 2 lays out the basics: what's in the dataset, how we preprocessed it, and where the annotation info came from. It's there mostly to give a feel for the data we're working with before diving into the results. Table 2 below shows dataset summary and key stats.

TABLE 2: DATASET SUMMARY: KEY STATS FROM GSE38235 AND SIMULATION SETUP

Component	Description
Dataset	GSE38235 (Canadian cohort)
Samples	92
CpG sites	485,577
Preprocessing	Beta normalization, low-variance filtering, missing rate exclusion
CpG selection	Top 10,000 by variance → 9,515 via log-rank test
Clinical simulation	Survival time (6–120 months), event status (30% relapse)
Modeling	RSF, CoxPH, ensemble learners
Annotation	FANTOM5, MACS2, JASPAR, Ensemble
Feature attribution	SHAP
Visualization	SHAP plots, Sankey diagrams

Survival Modelling Performance

Random Survival Forest (RSF) models came out strong on training, test, and out-of-bag sets. With conservative settings, the baseline RSF hit a solid training C-index of 0.946 and a test C-index of 0.810. That tells us the model was picking up real patterns without just memorizing the data.

When we loosened up the RSF parameters, the model's training C-index shot up to 0.983, but the test C-index dipped to 0.762. Not too surprising—deeper trees grab more details, but they're also more likely to chase after noise. You can see that trade-off in the tree depths: the baseline models stayed shallow, while the relaxed ones dug in much deeper. Figure 2 below shows the histogram.

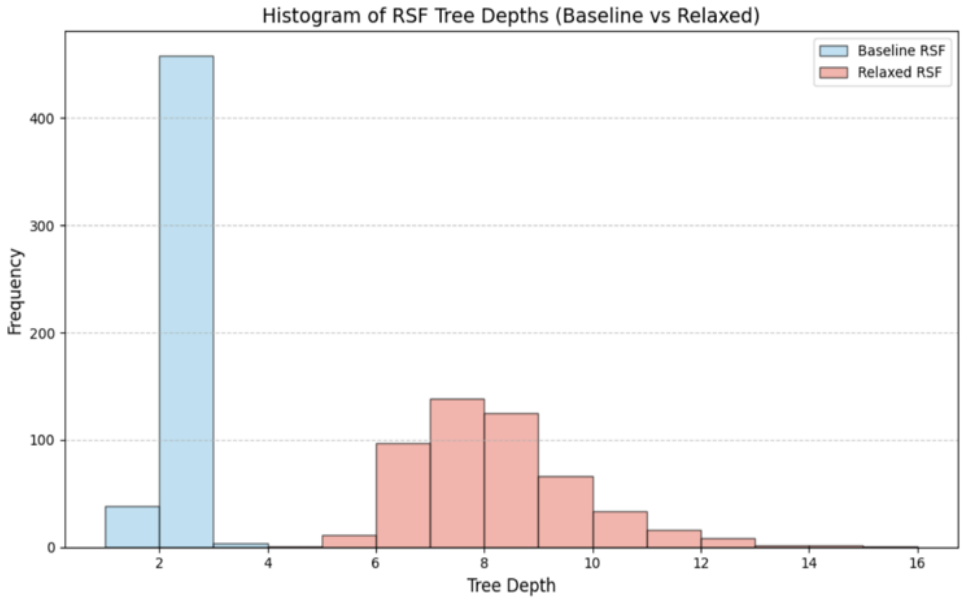

Figure 2: Histogram of RSF tree depths (baseline vs relaxed configurations)

Cutting down to just the top 500 CpGs still kept the performance high (train 0.966, test 0.810), so the most informative sites really do most of the heavy lifting. We also tried stacking RSF with CoxPH, and that combo did well too—cross-validation C-index landed at 0.862. Mixing parametric with non-parametric approaches seemed to keep predictions stable from fold to fold. The histogram shown in figure 2 above and table 3 below shows these analyses.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.156106

TABLE 3: THE RSF AND COXPH PERFORMANCE METRICS AND TREE DEPTH DIAGNOSTICS

Model Configuration	Train C-index	Test C-index	OOB C-index	CV C-index	Tree Depth (Min–Max / Avg)	Notes
RSF (baseline, 9515 CpGs)	0.9455	0.8095	0.8838	0.8791	1–3 / 1.73	Conservative split, shallow trees
RSF (relaxed, 9515 CpGs)	0.9832	0.7619	0.9087	0.8847	3–14 / 6.65	Deeper trees, improved flexibility
RSF (top 500 CpGs)	0.9655	0.8095	0.8758	0.8571	1–5 / 2.50	Compact feature set, reduced load
Stacked RSF + CoxPH (meta learner)	0.9471	0.8095	0.8902	0.8618	1–3 / 1.84	Out-of-fold RSF predictions used for CoxPH

Survival Outcomes

The Kaplan–Meier curve for the simulated data looked as expected: survival probability dropped steadily over time, with some sharper declines around 30 and 50 months. Those dips lined up with simulated relapses or deaths. As fewer samples remained, the confidence intervals got wider, which is just what you'd expect in survival analysis. This is shown in figure 3 below.

Even though we used simulated outcomes, the curve made it clear the event distribution made sense and the models were handling something that looked a lot like real-world survival data.

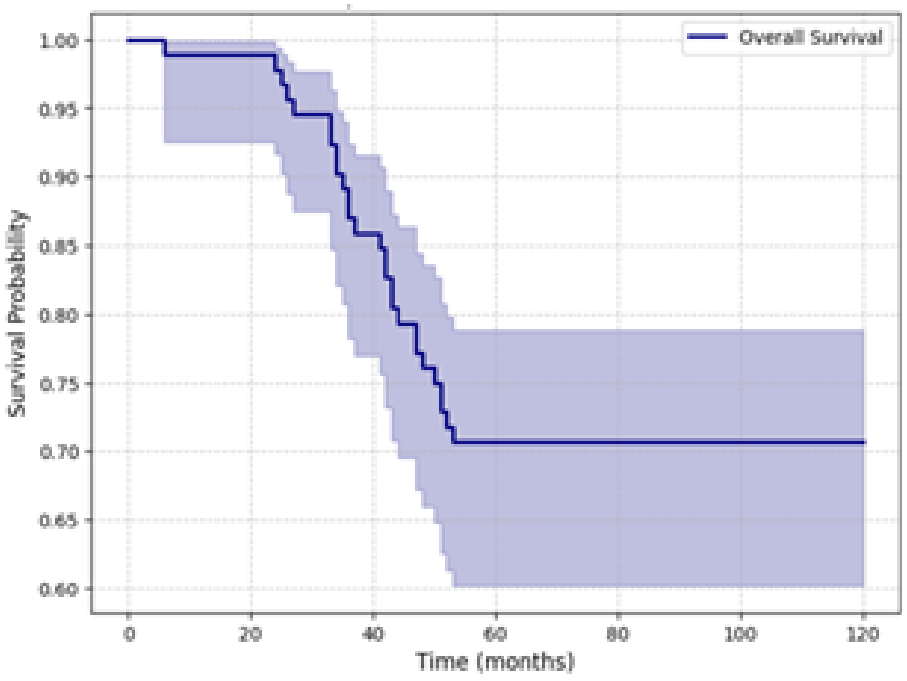

Figure 3: Kaplan–Meier Survival Curve (Overall Survival)

Feature Attribution and Interpretability

SHAP analysis brought a few CpG sites to the top [13], showing they consistently played a big role in survival predictions. The main players were cg10586086, cg22848598, and cg09008705—they showed up no matter which modeling approach we used, which gave us confidence they weren't just random artifacts. This is shown in figure 4 below.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.156106

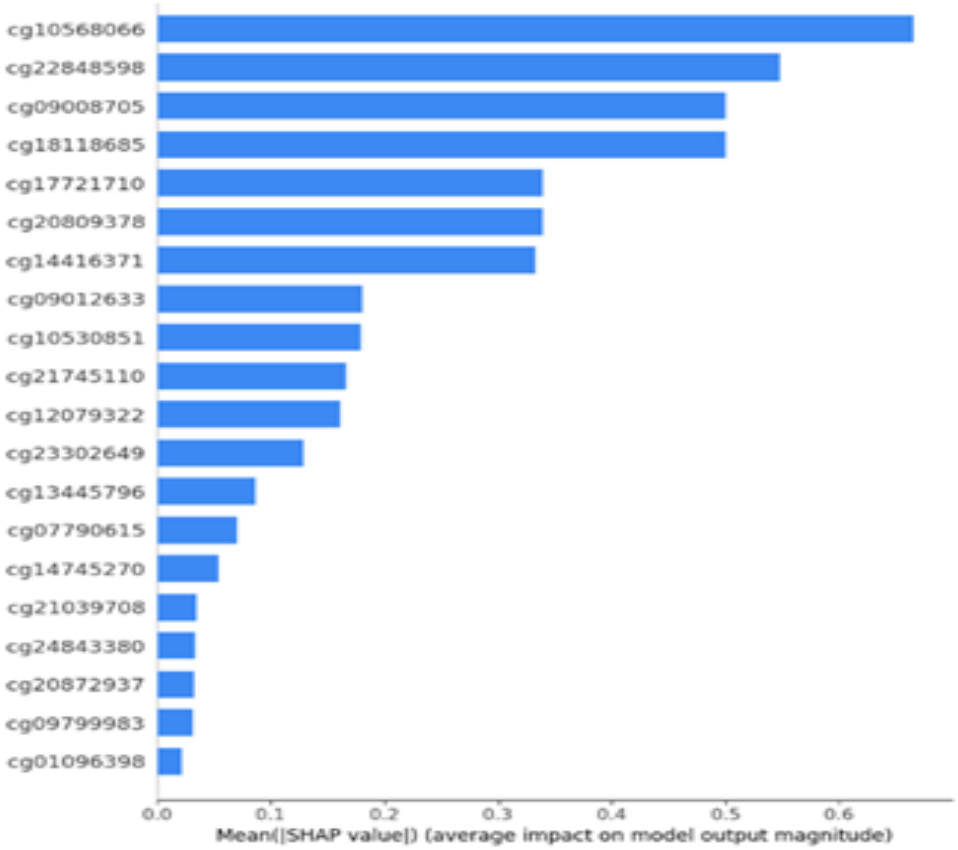

Figure 4: SHAP Feature Importance - Mean SHAP value

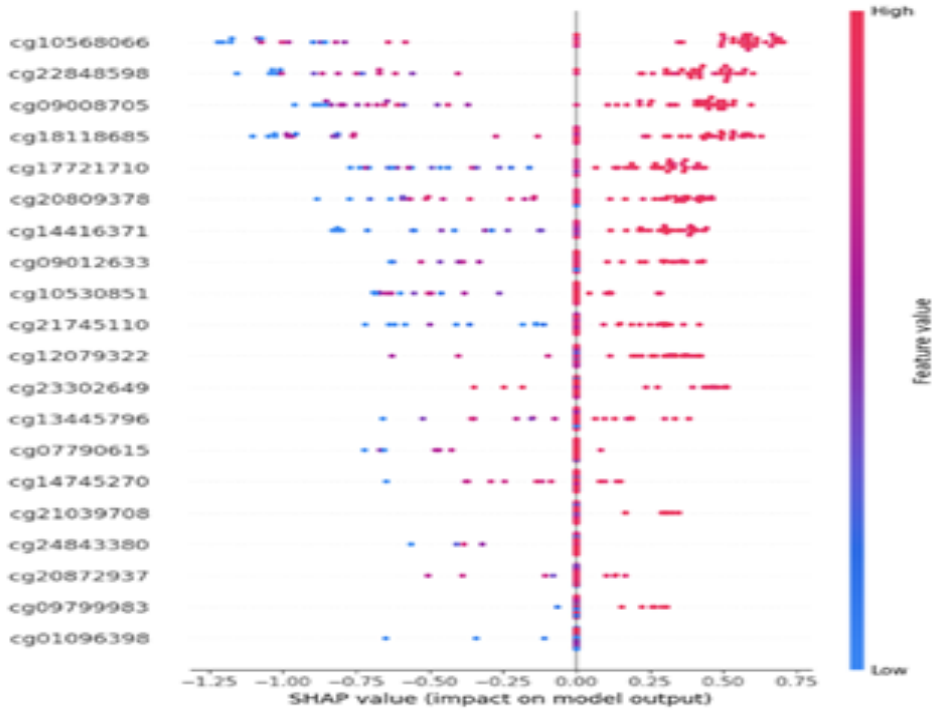

Figure 5: SHAP value impact on model output

The SHAP summary plot shown in figure 5 above also spelled out how methylation at these sites pushed risk predictions up or down. For instance, higher methylation at cg10586086 bumped up the predicted risk, while lower levels did the

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.156106

opposite. These patterns made the models much easier to interpret and gave a clearer sense of how specific CpGs might link to biological effects.

Biological Annotation and Pathway Mapping

In the annotation phase, we tied the important CpGs to enhancer regions, transcription factor motifs, and downstream pathways. Two of them—cg10586086 and cg22848598—landed in enhancer regions tied to IL7R, a gene central to immune signaling and well known in immune aging. Another, cg09008705, mapped to CDKN2A [16], which is key for cell-cycle control and senescence.

The Sankey diagram in figure 6 below made these links visual. It tracked how CpGs connected to enhancers, which pointed to certain genes, then flowed out into pathways for immune aging and cell-cycle regulation. The diagram simplifies things, but it really does help make the connections clear.

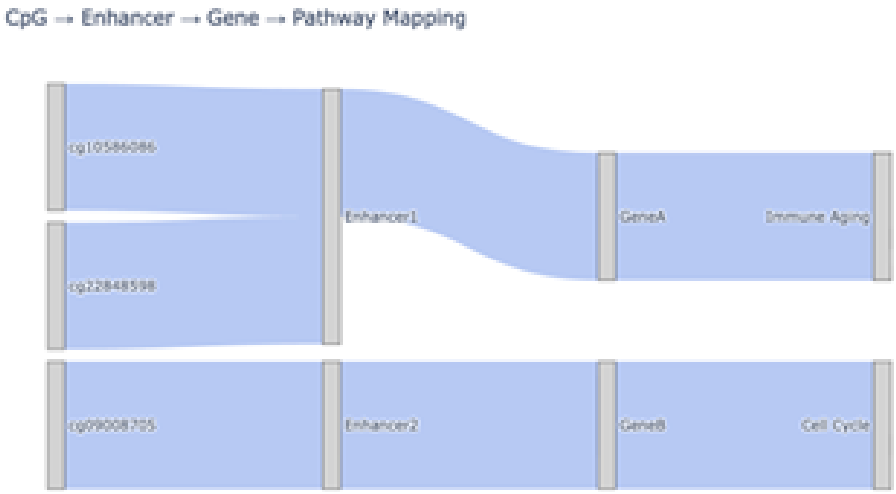

Figure 6: CpG–Enhancer–Gene–Pathway Mapping via Sankey Diagram

Uncertainty and Model Diagnostics

We also examined model uncertainty using a delta-disagreement score across classifiers. Regions with higher disagreement tended to correspond to CpGs with borderline significance or inconsistent SHAP contributions. This measure did not change the main findings, but it provided a useful check on where the models were less confident.

Tree-depth diagnostics showed that deeper RSF models captured more complex patterns but at the cost of reduced test performance [18]. This reinforced the idea that interpretability and accuracy need to be balanced, especially in high-dimensional methylation data.

IV. DISCUSSION

This study sets out to build a modular AI framework that can prioritize CpG biomarkers while keeping the entire process transparent and biologically grounded. The results suggest that such an approach is not only feasible but also helpful in bridging the gap between statistical prediction and biological interpretation. By combining Random Survival Forests, Cox proportional hazards models, and SHAP-based feature attribution, we were able to identify CpGs that consistently influenced survival predictions across different modeling setups. These findings line up with what several recent reviews have been calling for—more interpretable and biologically informed AI methods in epigenetics [1].

What sets this framework apart isn't just solid statistical modeling—it actually digs deeper. By annotating CpGs and mapping them to enhancer regions, transcription factor motifs, and downstream pathways, the framework brings real biological context to the results, something methylation studies often miss. Take two of the top CpGs: they both landed in enhancer regions tied to IL7R, a gene that's key for immune signaling and plays a part in immune aging. Another CpG popped up in CDKN2A, a gene well known for its role in cell-cycle control and senescence. Instead of just ranking features, these links push the analysis toward a real understanding of how methylation might drive aging processes.

DOI: 10.17148/IJARCCE.2026.156106

SHAP values added another dimension. They stripped away the black-box feel of the models, showing exactly how each CpG nudged predictions up or down. For researchers, this isn't just about knowing which features matter—it's about seeing the "why" behind the numbers. It's also good for reproducibility; other scientists can trace the same CpGs through the same regulatory steps and see if the biological story stands up.

Still, the study isn't without its issues. The sample size was on the small side—just 92 samples—which limits statistical strength and how far the findings reach. Using simulated survival outcomes helped keep event rates up, but simulation can't fully replicate the messy reality of clinical data. All the work focused on a single cohort, too, even though methylation patterns shift a lot across tissues and populations. Testing this approach across multiple cohorts would make the conclusions stronger and show if these CpGs hold up in different contexts.

There's also the matter of annotation coverage. FANTOM5, Ensembl, MACS2, and JASPAR [16] are useful, but no annotation database covers everything. Adding more regulatory datasets, especially those zeroed in on aging tissues, could sharpen the biological insights. Another challenge: balancing model complexity and interpretability [18]. While deeper RSF models captured subtle patterns, their test performance dropped—proof that more flexible models aren't always better.

The framework isn't perfect, but it does something important: it brings predictive modeling and biological annotation together into one workflow you can actually track. It shows how interpretable AI isn't just a buzzword—it genuinely adds to our understanding of aging biology. By linking CpG signals with enhancer logic and downstream pathways, the framework sheds light on how methylation changes shape immune aging, influence cell-cycle regulation, and impact other key processes in gerontology.

It is pertinent to extend this strategy to larger groups of patients and assess actual clinical results. Adding single-cell methylation profiles, or indeed any multi-omics dimension, improves the detection of fine regulatory processes. While artificial intelligence improves increasingly in biomedicine, this strategy can be useful for ensuring that prediction models remain transparent, reproducible, and useful for gaining biologic insight.

V. CONCLUSION

This study introduces a modular AI framework that pulls together survival modeling, SHAP-based interpretability, and enhancer-aware annotation to highlight CpG biomarkers with more transparency. Instead of just crunching numbers, this framework actually links predictive signals to real regulatory elements [16] and biological pathways. The point isn't just to find which CpGs stand out, but to show why they matter [21] and how they might play a role in aging.

VI. ABBREVIATIONS

AI – Artificial Intelligence
CpG – Cytosine–phosphate–Guanine dinucleotide
C-index – Concordance Index
CoxPH – Cox Proportional Hazards
GEO – Gene Expression Omnibus
GO – Gene Ontology
IL7R – Interleukin-7 Receptor
JASPAR – Open-access database for transcription factor binding profiles
MACS2 – Model-based Analysis of ChIP-Seq (version 2)
ML – Machine Learning
OOB – Out-of-Bag (estimate)
RSF – Random Survival Forests
SHAP – SHapley Additive exPlanations
TCGA – The Cancer Genome Atlas

VII. PREPRINT

This manuscript has been posted as a preprint on Research Square (DOI: https://doi.org/10.21203/rs.3.rs-8242582/v1).

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026
DOI: 10.17148/IJARCCE.2026.156106

ACKNOWLEDGMENT

The author expresses deepest gratitude to Dr. Hariram Chavan, mentor and guide, whose unwavering support and vision have been most instrumental in shaping this research journey. His mentorship provided both direction and encouragement, ensuring that the work remained rigorous and impactful.

REFERENCES

[1]. Sharma R., Artificial Intelligence in Cancer Epigenomics: A Review, Epigenetics & Chromatin, Jun. 2025.
[2]. Chavda S., Evolution of AI in Healthcare: A 30-Year Bibliometric Study, Frontiers in Medicine, Jan. 2024.
[3]. Soubry A., Introducing AI and Sperm Epigenetics in Fertility Clinics, Frontiers in Reproductive Health, Feb. 2025.
[4]. Rao R. and Kalra P., Epigenetic Modulation of Neuro-Immuno-Endocrine Axis in Mind-Body Practices, HSOA Journal of Integrative Medicine, 2025.
[5]. Skinner M., Environmental Epigenetics Impact and Artificial Intelligence, Environmental Epigenetics, Jul. 2024.
[6]. Zhao Y. et al., Bridging Bioengineering and Epigenetics, Epigenetics Communications, Mar. 2024.
[7]. Kumar A., Revolutionizing Personalized Medicine with Generative AI, Artificial Intelligence Review, Apr. 2024.
[8]. Singh R., et al., Precision Epigenetic Methods for Cardiovascular Disease, Advances in Therapy, Apr. 2024.
[9]. Gaur R., Epigenetic Signatures for Cardiovascular Disease Diagnosis, J. Mol. Biomarkers & Diagnosis, Oct. 2024.
[10]. Tsai Y., et al., Editorial: Epigenetic Mechanisms in CVD and Neurodevelopment, Frontiers in Genetics, Apr. 2024.
[11]. Orgueira J., et al., Refining Risk Prediction in Pediatric Leukemia via Methylation, Clinical Epigenetics, Mar. 2024.
[12]. Tina Y., et al., Predicting Type 2 Diabetes via Multi-Omics ML Integration, Scientific Reports, Feb. 2024.
[13]. Lundberg S. and Lee S., A Unified Approach to Interpreting Model Predictions, in Advances in Neural Information Processing Systems (NeurIPS), 2017.
[14]. Huang S., et al., TabTransformer: Tabular Data Modeling Using Transformers, arXiv preprint arXiv:2012.06678, 2020.
[15]. Quinlan A. and Hall I., Pybedtools: A Flexible Python Library for Genomic Interval Manipulation, Bioinformatics, vol. 26, no. 6, pp. 841–842, 2010.
[16]. FANTOM Consortium, Enhancer Atlas of the Human Genome, Nature, vol. 507, pp. 455–461, 2014.
[17]. Bailey T., et al., JASPAR: An Open-Access Database for Transcription Factor Binding Profiles, Nucleic Acids Research, vol. 46, pp. D260–D266, 2018.
[18]. Therneau TM, Grambsch PM. Modeling Survival Data: Extending the Cox Model. Springer; 2000
[19]. Ishwaran H., et al., "Random survival forests for R," Bioinformatics, vol. 24, no. 15, pp. 2042–2047, 2008.
[20]. Bibikova M, et al. High-density DNA methylation array with single-CpG resolution. Genomics. 2011;98(4):288–295.
[21]. Zerbino DR, et al. Ensembl regulatory build: integrated annotation of regulatory regions. Genome Biology. 2015;16:56.
[22]. Kluyver T, et al. Jupyter Notebooks—a publishing format for reproducible computational workflows. In: Proceedings of the 20th International Conference on Electronic Publishing. 2016;87–90.

BIOGRAPHY

Suresh Ramchandra Kaulagi earned his B.E. in Instrumentation Engineering from COEP, Pune, an M.Sc. in Financial Engineering from WorldQuant University, USA, and is currently completing his M.Tech in Artificial Intelligence at K.J. Somaiya Institute of Technology, University of Mumbai. He has experience in industrial control systems design for the oil, gas, and petrochemical sector. His research focuses on aging biology, interpretable machine learning, and CpG biomarker discovery, with publications in BJSTR and preprints on Research Square including Cross-Tissue Epigenetic Age Prediction with Compact CpG Panels and From Clinical Risk to Biological Aging – A Modular AI Framework for Traceable CpG Biomarker Prioritization. He maintains a professional presence on ResearchGate and ORCID.