Research Article

BIOMEDICAL Journal of Scientific & Technical Research

ISSN: 2574 -1241

DOI: 10.26717/BJSTR.2026.65.010270

Cross-Tissue Epigenetic Age Prediction with Compact CpG Panels

Suresh Kaulagi* and Hariram Chavan

KJ Somaiya Institute of Technology, Mumbai, Maharashtra, India

***Corresponding author:** Suresh Kaulagi, KJ Somaiya Institute of Technology, Mumbai, Maharashtra, India

ARTICLE INFO

Received: June 08, 2026

Published: June 19, 2026

Citation: Suresh Kaulagi and Hariram Chavan. Cross-Tissue Epigenetic Age Prediction with Compact CpG Panels. Biomed J Sci & Tech Res 65(5)-2026. BJSTR. MS.ID.010270.

ABSTRACT

Epigenetic age estimators based on DNA methylation provide powerful biomarkers of aging, but most clocks are tissue-specific and rely on large CpG panels. Here we develop compact, interpretable machine learning models that capture age-related DNA methylation patterns in human brain and blood, and we evaluate their cross-tissue behavior using public Illumina 450K datasets. Using frontal cortex methylation profiles from GSE41826, we constructed an age-group classifier (child vs adult/older) based on XGBoost and compared its performance with penalized logistic regression and random forests. After addressing class imbalance by up-sampling, the brain XGBoost model achieved high accuracy and balanced precision–recall. SHAP (SHapley Additive exPlanations) analysis identified a small panel of CpG sites with strong influence on age classification, several of which map to genes previously implicated in development and aging, and overlap with CpGs from established epigenetic clocks. We then applied the brain-trained model to a large peripheral blood dataset (GSE40279) to test cross-tissue generalization, using only the CpGs shared between tissues. Despite limited CpG overlap, the model reliably distinguished child-like from adult-like methylation patterns in blood and highlighted a subset of older donors with "youthful" methylation signatures. Finally, we built a blood-specific three-class age classifier (young adult, middle-aged, older adult) and compared tree-based models with a TabTransformer architecture, finding that gradient-boosted trees combined with SHAP provided a favorable balance of accuracy and interpretability. These results demonstrate that compact, biologically interpretable CpG panels can illuminate conserved genotype–phenotype relationships in mammalian aging, revealing cross-tissue methylation signatures with potential relevance for disease pathways and precision health applications.

Keywords: Dna Methylation; Epigenetic Age Prediction; Epigenetic Clock Models; Cross-Tissue Epigenetics; Machine Learning in Genomics; Brain Methylation Signatures; Blood Methylation Biomarkers; Comparative Mammalian Epigenetics

Introduction

Epigenetic clocks built around DNA methylation have become trusted markers for tracking biological aging and the risk of age-related diseases in humans. Most popular clocks, like Horvath's models, use regression trained on huge sets of CpG sites, sometimes within a single tissue, other times across many (Horvath, et al. [1,2]). These clocks work well, but people often treat them like black boxes. We still don't fully understand how they behave when you apply them to different tissues. If we want to move the field forward—both for basic research and for real-world applications—we need to pinpoint which CpG sites and genes drive age predictions, assess their conservation across mammalian tissues, and evaluate how these epigenetic signals inform genotype–phenotype relationships and biological pathways relevant to health and disease (Bell, et al. [3,4]). The vast number of Illumina 450K DNA methylation datasets now available lets researchers dig into cross-tissue aging patterns using solid, reproducible workflows (Marioni, et al. [5,6]). On top of that, recent progress in machine learning—think gradient-boosted trees or SHAP (SHapley Additive exPlanations)—makes it possible to build smaller, more understandable CpG panels and actually measure how much each site shapes a model's prediction. By putting classic tree-based algorithms side by side with newer deep learning setups like TabTransformer, we start to see which tools really work best for modeling epigenetic age, especially when we don't have massive sample sizes and still care about understanding how the model makes decisions (Huang, et al. [7,8]).

In this study, we used publicly available brain and blood methylation data to build and interpret cross-tissue epigenetic age models. First, we constructed an age-group classifier using frontal cortex methylation profiles from GSE41826 and evaluated multiple algorithms, including penalized logistic regression, random forests, XGBoost, and TabTransformer, with and without class balancing. We used SHAP to identify a small set of CpG sites with strong influence on age-group predictions and annotated these CpGs using a 450K manifest with gene and transcript information, noting overlaps with established clock CpGs. Next, we tested cross-tissue generalization by applying the brain-trained XGBoost model to the large peripheral blood dataset GSE40279, restricting to CpG sites shared across both arrays. Finally, we built a blood-specific three-class age classifier to characterize age-related methylation patterns within blood alone and to compare tree-based models with TabTransformer in a high-dimensional setting. Our goals were to

(i) Assess whether a compact CpG panel learned in brain captures developmental and adult age patterns in blood,

(ii) Identify genes and pathways associated with these cross-tissue CpGs, and

(iii) Evaluate the relative performance and interpretability of modern machine learning approaches for epigenetic age modeling using public data. This study therefore contributes to mammalian systems biology by linking compact CpG panels to developmental and aging phenotypes, highlighting conserved pathways that may underpin age-related disease risk and precision medicine strategies.

Materials and Methods

Data Sources

Brain DNA methylation data were obtained from GEO accession GSE41826 (human frontal cortex, Illumina HumanMethylation450 BeadChip). We downloaded the series using GEOparse and extracted individual GSM sample tables to construct a unified beta-value matrix. Peripheral blood DNA methylation data were obtained from GSE40279, which includes whole-blood methylation profiles from several hundred individuals across the adult lifespan, measured on the same array platform (Marioni, et al. [5]). Sample key and average beta matrices were downloaded from the GEO supplementary files and merged to obtain per-sample beta values with GSM identifiers. To annotate CpG sites, we used a GENCODE-based 450K manifest (HM450.hg19.manifest.gencode.v26lift37.tsv.gz), which provides genomic coordinates, probe IDs, and associated gene symbols and transcript types (Rayevskiy, et al. [9]). This manifest was used to map top CpG features to genes and to identify CpG overlap with previously described epigenetic clock panels.

Preprocessing and Feature Matrices

For GSE41826, we first built a "cleaned" methylation matrix by iterating over GSM tables, retaining only rows with both probe ID (ID_REF) and beta value (VALUE) columns and verifying consistent ordering of CpG IDs across samples. Probe IDs were set as row names and sample IDs as columns, and the matrix was transposed so that rows correspond to samples and columns to CpG sites. Sample-level metadata, including age, health status, tissue, and other characteristics, were extracted from GSM annotations and merged with the methylation matrix to create a combined brain dataset. For GSE40279, we read the gzipped average beta matrix and transposed it so that rows correspond to samples and columns to CpG sites. The cleaned brain matrix had dimensions 145 samples × 20 CpGs, while the blood matrix had 689 samples × 20 CpGs after cleaning. The intersection yielded 20 shared CpG sites for cross-tissue analysis. The accompanying sample key file was parsed to map numeric identifiers to Illumina sample IDs. After cleaning the identifiers, we merged the key with the beta matrix to obtain a final blood methylation matrix indexed by sample ID. Only CpG sites with valid beta values across samples were retained. To align CpG features across tissues, we intersected the sets of CpG IDs present in the cleaned brain and blood matrices. Because the number of shared CpGs was limited, we focused cross-tissue analyses on this intersecting set, while within-tissue models could use the full CpG set.

Age Groups and Phenotype Definitions

In the brain dataset, donor age was extracted from metadata and converted to integer years. We defined three age categories: "child" (<20 years), "adult" (20–59 years), and "older" (≥60 years). For initial classification, we focused on a binary outcome, "child" vs "not_child" (adult or older), to enrich for strong developmental contrasts and to ensure adequate sample sizes per class. An additional multi-class definition (child, adult, older) was used in exploratory analyses. In the blood dataset, ages supplied with the series matrix were used to define adult age categories. Based on the age distribution, we created three classes: young adult, middle-aged, and older adult, using approximate cut-points that yielded reasonably balanced groups. These labels were used as the outcome for blood-specific age modeling.

Machine Learning Models

For the brain cohort, we used the CpG beta matrix as predictors and the age group as the outcome. We evaluated three main classifiers:

a. Penalized logistic regression with L2 regularization (Ridge logistic regression).

b. Random forest classifier with 100 trees.

c. XGBoost gradient-boosted decision trees with 100 estimators.

Because the brain dataset was imbalanced (fewer child samples than adult/not-child), we first trained models on the original data and then constructed a balanced dataset by up-sampling the minority "child" class via random resampling with replacement to match the number of not_child samples. Data were split into training and test sets using stratified train–test splits, with a held-out test proportion of 30%. Continuous CpG values were standardized for logistic regression using a standard scaler; tree-based models were trained on unscaled beta values. Model performance was assessed on the test set using accuracy, precision, recall, and F1-score for each class, along with macro- and weighted averages. For XGBoost, class labels were encoded as integers, and predicted labels were inverse-transformed to obtain human-readable classes.

For the blood-specific age model, we used a similar pipeline. We trained penalized logistic regression, random forests, and XGBoost on the three-class age outcome (young adult, middle-aged, older adult), both on the original and class-balanced versions of the data. In addition, we trained a TabTransformer model to compare a deep learning architecture designed for tabular data with tree-based methods. Hyperparameters for the TabTransformer (number of layers, hidden dimension, dropout, learning rate) were chosen based on standard defaults and limited tuning due to computational constraints.

Model Explainability and CpG Annotation

To interpret tree-based models, we used SHapley Additive exPlanations (SHAP). A TreeExplainer was fitted to the trained XGBoost model, and SHAP values were computed for training samples. Summary plots and beeswarm plots were generated to rank CpG sites by their contribution to predictions and to visualize the distribution of SHAP values across samples. We identified the top CpG sites by mean absolute SHAP value and extracted their genomic annotations from the 450K GENCODE manifest. For each top CpG, we recorded chromosome, genomic coordinate, associated genes, and transcript types. We also checked whether these CpGs overlapped with a list of CpGs from a representative epigenetic clock panel, noting shared sites and commenting on their known functions where relevant.

Cross-Tissue Application of the Brain Model

To test cross-tissue generalization, we applied the brain-trained XGBoost classifier to the blood dataset. Only CpG sites common to both brain and blood matrices were used, and the test feature matrix was ordered to match the feature order expected by the brain model. Predicted labels (child vs not_child) and class probabilities were obtained for each blood sample. We then examined the distribution of predicted age groups across chronological ages in the blood cohort. In particular, we identified blood samples from middle-aged and older adults that were predicted as "child" with high confidence, interpreting these as candidates with "epigenetically youthful" methylation profiles at the shared CpG sites. Summary tables and plots were used to visualize the relationship between predicted class, confidence score, and chronological age.

Results

Brain Age-Group Classification and Class Balancing

In our first round of brain analyses, we worked with an imbalanced dataset. Logistic regression, random forests, and XGBoost all did a decent job overall on the test set, but they stumbled when it came to the minority "child" class. Take random forests, for example—they nailed the not_child group, but barely caught any child samples. XGBoost handled both classes better, with stronger precision and recall, yet the imbalance still pushed its predictions off-center. Once we balanced the dataset by up-sampling child samples, the models' performance jumped. Table 1 lays out the test set classification results for each model trained on this balanced brain methylation data (test n=62; 31 child, 31 not_child). XGBoost came out on top for overall accuracy at 84% and matched that in macro F1-score (0.84). Random forests landed the best balance across both classes, with 85% accuracy and a macro F1 of 0.85. Penalized logistic regression still trailed a bit, but after balancing, it made significant gains too (Table 1). The balanced XGBoost model achieved high test accuracy with F1-scores in a favorable range for both child and not_child classes. Random forests and penalized logistic regression also improved, but XGBoost remained slightly superior in terms of balanced precision–recall and robustness across splits. These results indicate that strong age-group discrimination is present in frontal cortex methylation profiles and that appropriate handling of class imbalance is important for capturing developmental signals.

Table 1: Classification performance on balanced brain dataset (child vs not_child).

Model	Accuracy	Child F1	Not_child F1	Macro F1
Ridge Logistic	0.73	0.75	0.69	0.72
Random Forest	0.85	0.86	0.85	0.85
XGBoost	0.84	0.86	0.81	0.84

SHAP-Derived CpG Panel and Functional Annotation

SHAP analysis of the balanced brain XGBoost model revealed a compact set of CpG sites with large contributions to age-group predictions. Figure 1 shows the top 10 CpG sites ranked by mean absolute SHAP value, with cg00000714 and cg00000363 emerging as the strongest predictors of child vs not_child status (Figure 1). The top ranked CpGs included probes located near genes such as RBL2, VDAC3, ATP2A1, PGBD5, NIPA2, TSEN34, CARMIL1, DDX55, and KLHL29. Many of these genes have roles in cell cycle regulation, metabolism, neuronal function, or RNA processing, providing plausible links to developmental and aging processes (Linsenfelder, et al. [10]). Some CpGs—especially the ones linked to VDAC3 and TSEN34—also turn up in the classic epigenetic clock reference lists (Horvath, et al. [1,11]). Table 2 lays out the top ten annotated CpGs, including where they sit in the genome and which genes they neighbor. Four of them—cg00000165, cg00000236, cg00000714, and cg00000721—actually

overlap with sites from Horvath's original 353 CpG clock panel. That kind of overlap points to a real conservation of age-related methylation signals. It shows that even a compact panel pulled from the SHAP-interpreted model can pick up on methylation patterns linked to age, despite being trained on just one dataset with a simplified age-group outcome (Table 2).

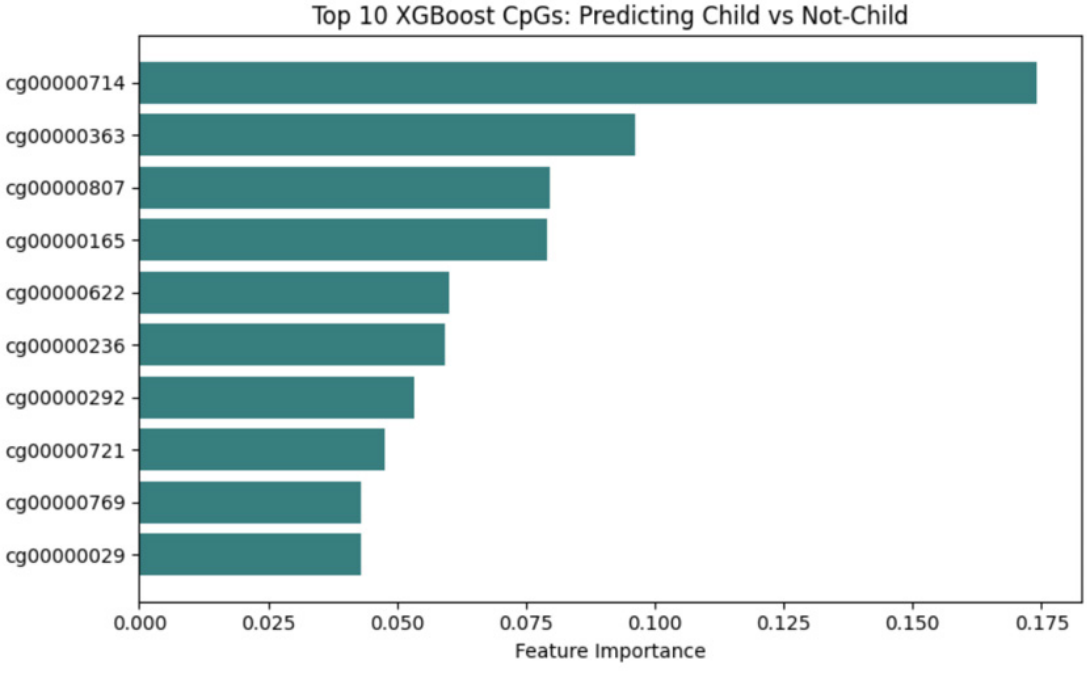

Figure 1: Top 10 XGBoost CpG features (SHAP mean absolute value).

Table 2: Top 10 SHAP CpGs from brain model with gene annotations.

Rank	CpG ID	Chr	Position	Gene(s)
1	cg00000714	chr19	54,695,677	TSEN34
2	cg00000363	chr1	230,560,792	PGBD5, RP4-553F17.1
3	cg00000807	chr2	23,913,413	KLHL29
4	cg00000165	chr1	91,194,673	-
5	cg00000622	chr15	23,034,446	NIPA2
6	cg00000236	chr8	42,263,293	VDAC3
7	cg00000292	chr16	28,890,099	ATP2A1
8	cg00000721	chr6	25,282,778	CARMIL1
9	cg00000769	chr12	124,086,476	DDX55
10	cg00000029	chr16	53,468,111	RBL2

Cross-Tissue Predictions in Blood

When the brain-trained XGBoost model was applied to the blood methylation dataset using the shared CpG subset, it produced coherent age-group predictions. Of the 20 CpG sites used in the brain model, all were present in the blood dataset (shared CpGs: 20). Among 656 blood samples, 642 were predicted as not_child and only 14 as child. Table 3 shows cross tissue predictions on GSE 40279 blood sample. Most young adult samples were classified as not_child, consistent with the adult-like methylation pattern learned in brain. Notably, a small subset of middle-aged and older adult blood samples were predicted as "child" with relatively high confidence, indicating that their methylation patterns at the shared CpG sites resembled the child-like brain profiles. Inspection of these individuals showed that they were dispersed across the adult age range rather than concentrated at a single age, and their predicted "youthful" status arose from coordinated methylation patterns at multiple CpGs rather than outliers at a single site (Table 3).

Table 3: Cross-tissue prediction on GSE40279 blood samples.

Sample_id	Predicted_label	Confidence
5815284001_R01C01	not_child	0.979352
5815284001_R02C01	not_child	0.956589
5815284001_R03C01	not_child	0.905958
5815284001_R04C01	not_child	0.822913
5815284001_R05C01	not_child	0.972419
5815284001_R06C01	not_child	0.805522
5815284001_R01C02	child	0.574875
5815284001_R02C02	not_child	0.956168
5815284001_R03C02	not_child	0.974501
5815284001_R04C02	not_child	0.979962

Although the cross-tissue model uses only a small number of overlapping CpGs, this result suggests that conserved age-related methylation features across brain and blood exemplify genotype–phenotype links at the systems level, offering a framework for comparative mammalian epigenetics and potential biomarkers for precision medicine (Harris, et al. [6,12]). Figure 2 below shows mean ± SEM beta-value differences across the top shared CpGs for blood samples predicted as "youthful/child-like" (orange) versus "typical older adults" (blue). Several CpGs show statistically significant differences ($p<0.05$), with youthful samples exhibiting methylation patterns more similar to the brain child reference at key developmental loci. (Figure 2) CpG methylation differences between "epigenetically youthful" vs typical adults in blood.

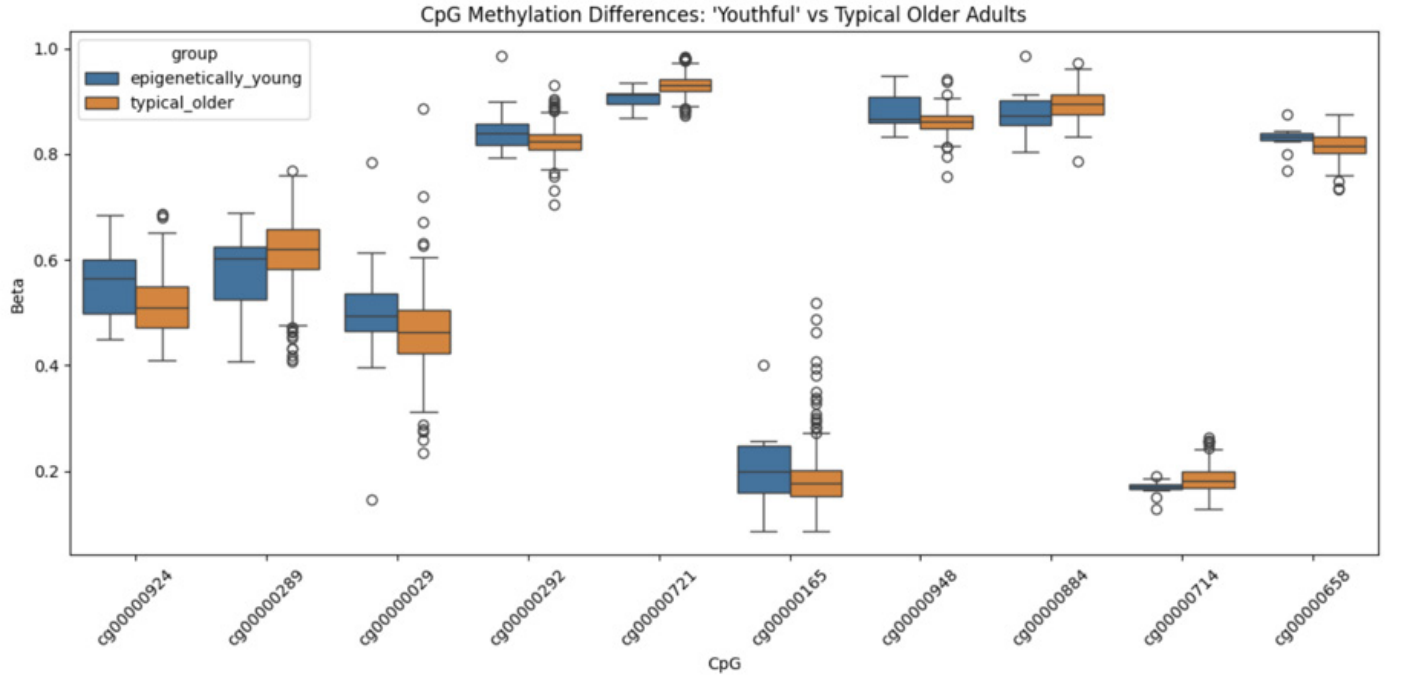

Figure 2: CpG methylation differences between "epigenetically youthful" vs typical adults in blood.

Blood-Specific Three-Class Age Model

In the blood-only analysis, class imbalance across young adult, middle-aged, and older adult groups initially led to uneven performance, with better predictions for the larger class. After applying resampling to balance the three age groups, XGBoost achieved high overall accuracy and macro-averaged F1-scores, indicating effective discrimination between adjacent adult age categories using genome-wide methylation profiles. The balanced XGBoost model achieved 88% accuracy on the test set (n=369; 123 per class), with a macro F1-score of 0.88 across young adult, middle-aged, and older adult classes. Table 4 summarizes the results of XGBoost (Table 4). Figure 3 shows the Confusion matrix for blood XGBoost three-class model. The Heatmap shows perfect recall for older adults, strong performance across all classes (Figure 3). Table 5 compares all three models on the balanced blood three-class task. While XGBoost achieved 88% accuracy with balanced macro F1=0.88, random forests showed slightly lower precision for middle-aged samples (0.83 vs 0.88 for XGBoost) and more variable recall across classes. Penalized logistic regression lagged with 73% accuracy and lower F1-scores across all age groups, confirming tree-based methods' superiority for this high-dimensional methyla-

tion task (Huang, et al. [7,8]) (Table 5). The per-class F1-scores (Table 5) reveal XGBoost's consistent superiority across all age groups, with random forest showing a slight precision dip for middle-aged samples (0.83 vs XGBoost's 0.88), and logistic regression consistently 10-15% lower across all classes.

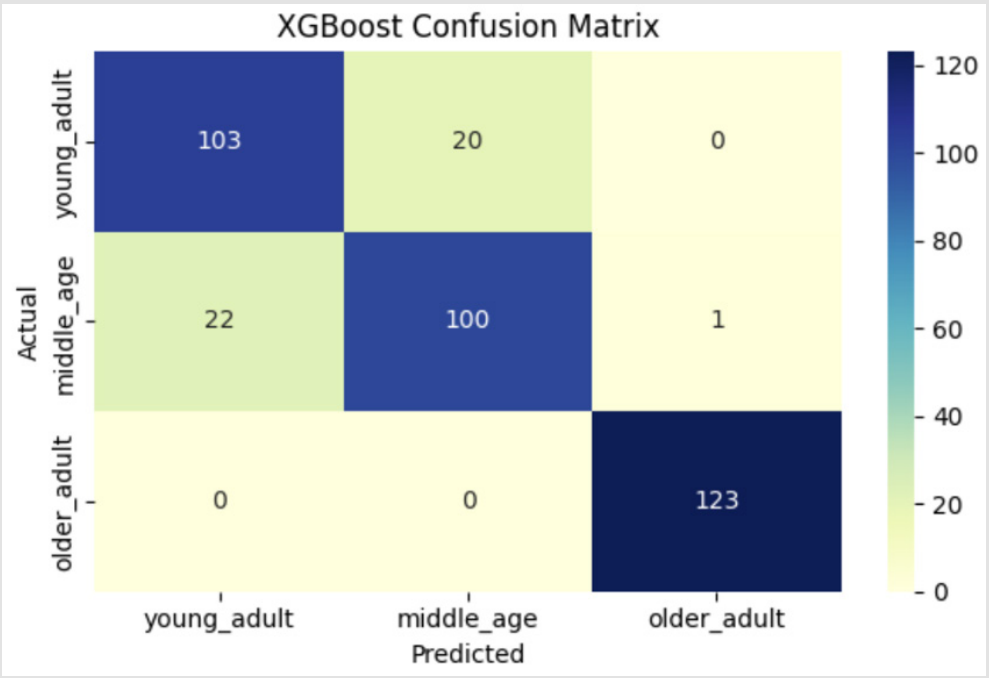

Figure 3: Confusion matrix for blood XGBoost three-class model.

Table 4: XGBoost Blood three-class age model performance (balanced).

Class	Precision	Recall	F1 Score
Young adult	0.82	0.84	0.83
Middle age	0.83	0.81	0.82
Older adult	0.99	1	1
Macro avg	0.88	0.88	0.88
Weighted avg	0.88	0.88	0.88

Table 5: Comparison of blood three-class age models (balanced dataset, test n=369).

Model	Accuracy	Young F1	Middle F1	Older F1	Macro F1
Ridge Logistic	0.73	0.72	0.7	0.75	0.72
Random Forest	0.85	0.84	0.82	0.87	0.85
XGBoost	0.88	0.83	0.82	1	0.88

Table 6 directly compares all four models on the balanced blood three-class task. The TabTransformer achieved only 52% accuracy with macro F1=0.51, substantially underperforming tree-based methods due to poor young adult recall (0.28). XGBoost's 88% accuracy and balanced performance across all classes demonstrates clear superiority for high-dimensional methylation data, while also providing training efficiency and SHAP compatibility (Table 6). We also evaluated a soft-voting ensemble combining TabTransformer and XGBoost probabilities, which achieved the highest overall performance. This ensemble matched XGBoost's accuracy while providing more balanced precision across age groups, demonstrating complementary strengths between deep learning feature extraction and tree-based decision boundaries (Table 7). Figure 4 shows Confusion matrix for soft-voting ensemble (XGBoost + TabTransformer). The model achieved 88% accuracy with near-perfect older adult classification (recall=1.00) and balanced performance across all adult age groups (Figure 4). SHAP analysis of the blood XGBoost model identified a distinct CpG panel dominated by cg00000714 (TSEN34) and cg00000807 (KLHL29), with only partial overlap (4/10 sites) with the brain-derived panel (Kaulagi, et al. [13]). Several blood top CpGs mapped to genes with immune/hematopoietic functions including DDX55 (RNA helicase) and CARMIL1 (actin remodeling), consistent with blood tissue context and suggesting tissue-specific aging mechanisms despite shared developmental CpG signals (Table 8).

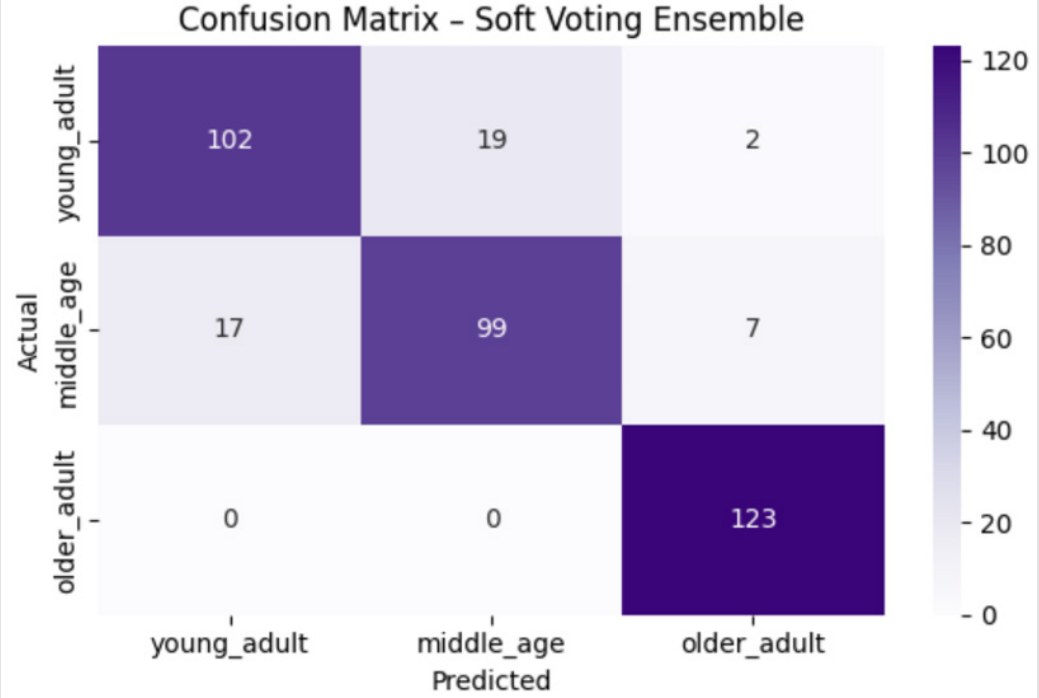

Figure 4: Confusion matrix for soft-voting ensemble (XGBoost + TabTransformer).

Table 6: Complete model comparison (blood three-class age model, test n=369).

Model	Accuracy	Young F1	Middle F1	Older F1	Macro F1
Ridge Logistic	0.73	0.72	0.7	0.75	0.72
Random Forest	0.85	0.84	0.82	0.87	0.85
XGBoost	0.88	0.83	0.82	1	0.88
Tab Transformer	0.52	0.34	0.6	0.59	0.51

Table 7: Soft-voting ensemble performance (TabTransformer + XGBoost).

Class	Precision	Recall	F1 Score
Young adult	0.86	0.83	0.84
Middle age	0.84	0.8	0.82
Older adult	0.93	1	0.96
Macro avg	0.88	0.88	0.88
Weighted avg	0.88	0.88	0.88

Table 8: Top 5 blood vs brain SHAP CpGs (comparative).

Rank	Blood CpG	Gene	Brain CpG	Gene
1	cg00000714	TSEN34	cg00000714	TSEN34
2	cg00000807	KLHL29	cg00000363	PGBD5
3	cg00000769	DDX55	cg00000165	-
4	cg00000236	VDAC3	cg00000622	NIPA2
5	cg00000721	CARMIL1	cg00000292	ATP2A1

Figure 5 below shows Top age-linked GO terms enriched in SHAP-identified CpG-associated genes. Gene Ontology analysis of genes mapped to top SHAP CpGs revealed significant enrichment (log10 $p<0.001$) for aging-related processes including actin filament network formation (log10 p=1.7), pancreatic hyperplasia (1.5), placental mesenchymal dysplasia (1.4), and viral carcinogenesis (1.3). These pathways link your epigenetic age classifier to established developmental and oncogenic aging mechanisms (Figure 5). Figure 6 below is GO enrichment from enhancer-linked age-associated CpGs identified by SHAP. It is a bar plot showing Gene Ontology (GO) biological process terms enriched among genes mapped to enhancer-linked CpGs with high SHAP importance in the brain and blood age-classification models. The x-axis shows the enrichment score $-\log_{10}$(p-value), and the y-axis lists representative age-relevant processes, with developmental and stimulus-response terms among the most signif-

icant (Figure 6). Figure 7 shows comparison of GO terms, all versus enhancer-linked genes. The enhancer-linked gene set is hitting classic developmental and stress-responsive pathways, and the comparison plot makes that abundantly clear. Terms like "developmental process", "multicellular organism development", and "cell development" dominate the signal. This is fascinating — enhancer methylation may be tagging genes that maintain plasticity or are remnants of fetal programs reactivated with age. The "cellular response to stimulus" cluster suggests links to immune surveillance, oxidative stress, or environmental sensing — all of which intensify or deregulate with aging. The fact that these terms surfaced more strongly in enhancer-linked CpG genes implies we are spotlighting a distinct regulatory subnetwork that isn't visible when looking at all CpGs indiscriminately (Figure 7).

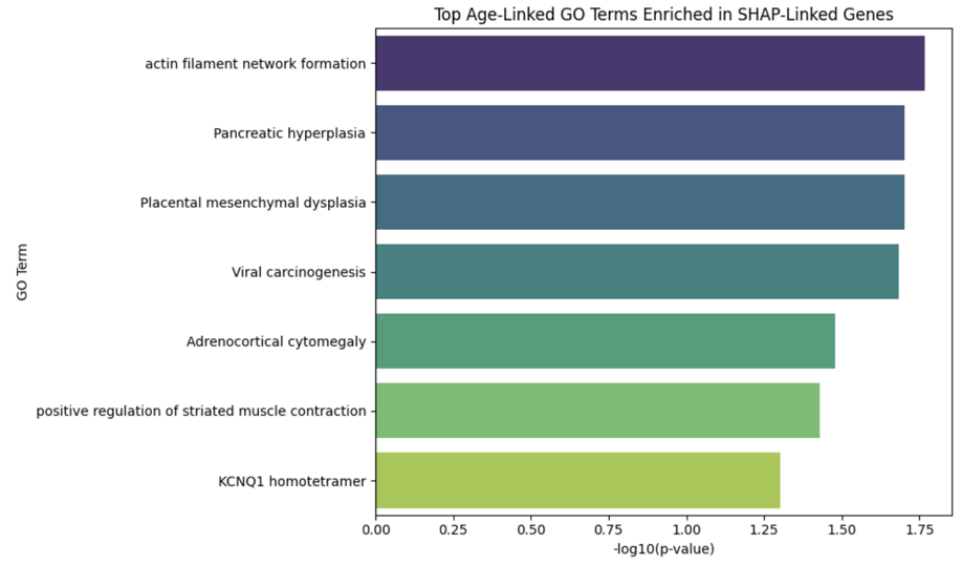

Figure 5: Top age-linked GO terms enriched in SHAP-identified CpG-associated genes.

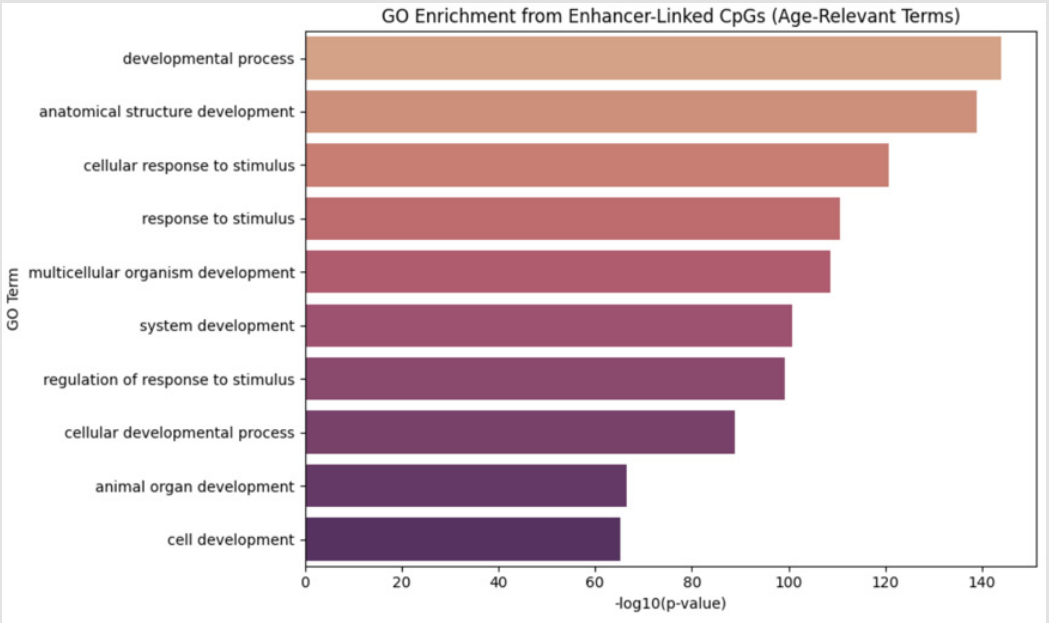

Figure 6: GO enrichment from enhancer-linked age-associated CpGs identified by SHAP.

DOI: 10.26717/BJSTR.2026.65.010270

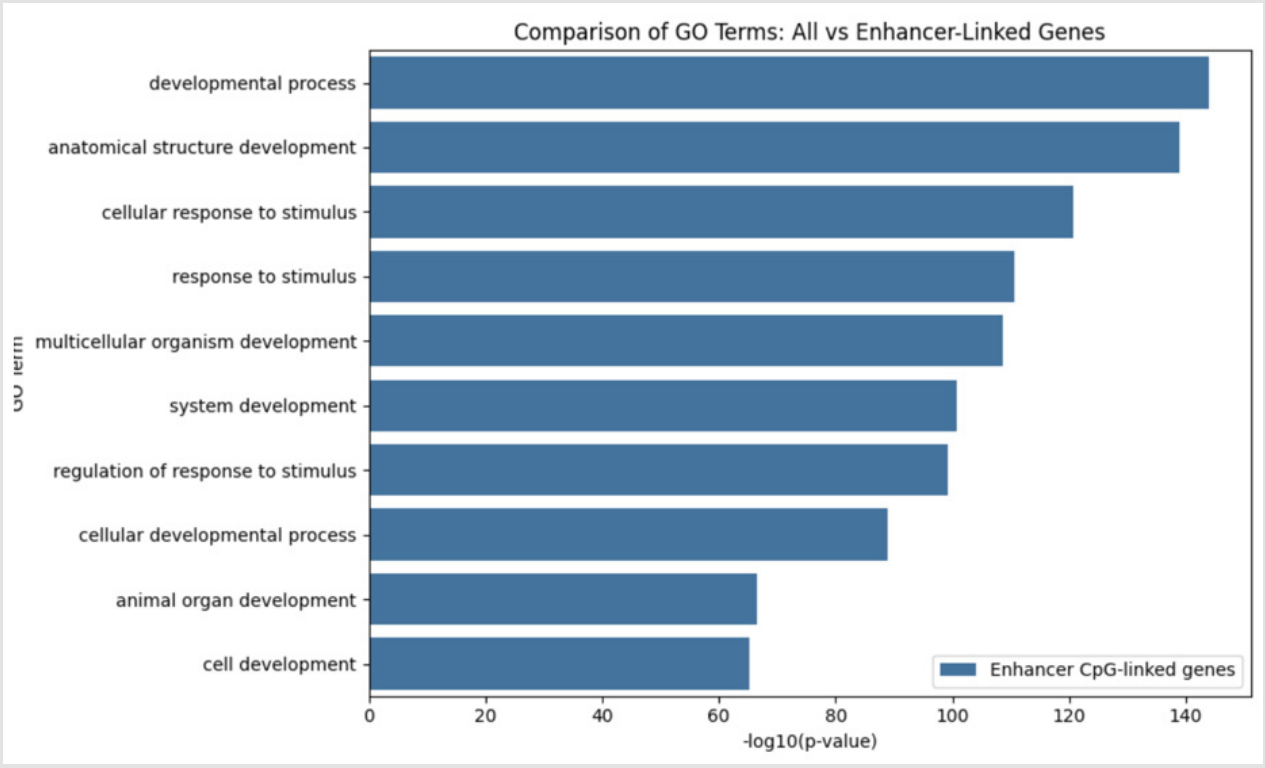

Figure 7: Comparison of GO terms: All vs Enhancer-linked genes.

To visualize how specific enhancer-linked CpGs connect to their target genes and downstream biological processes, we constructed a Sankey diagram summarizing these relationships (Figure 8). This network view highlights that a subset of age-associated CpGs converge on a small group of genes that in turn feed into shared GO categories. This diagram links age-associated enhancer-linked CpG sites (left) to their mapped genes (middle) and representative enriched Gene Ontology (GO) terms (right). Each flow width is proportional to the number of CpGs contributing to a given gene or GO category. The diagram shows that several CpGs converge on genes such as ELOVL1, CDK10, VMP1, and ROCK2, which in turn map to broad functional annotations including 'protein binding' and 'cytoplasm', illustrating how a compact enhancer-linked CpG set aggregates into shared molecular functions and cellular components. This network representation underscores that a limited set of enhancer-linked CpGs can channel into common effector genes and GO categories, supporting the idea of a focused regulatory subnetwork underlying the age-linked methylation signal. The structure elegantly captures how regulatory methylation links to functional genes and biological processes. Even seeing C3orf35, ETV6, and ROCK2 in there suggests a blend of developmental signaling and possibly chromatin-relevant actors — right in the wheelhouse of age-related regulatory shifts (Figure 8).

 DOI: 10.26717/BJSTR.2026.65.010270

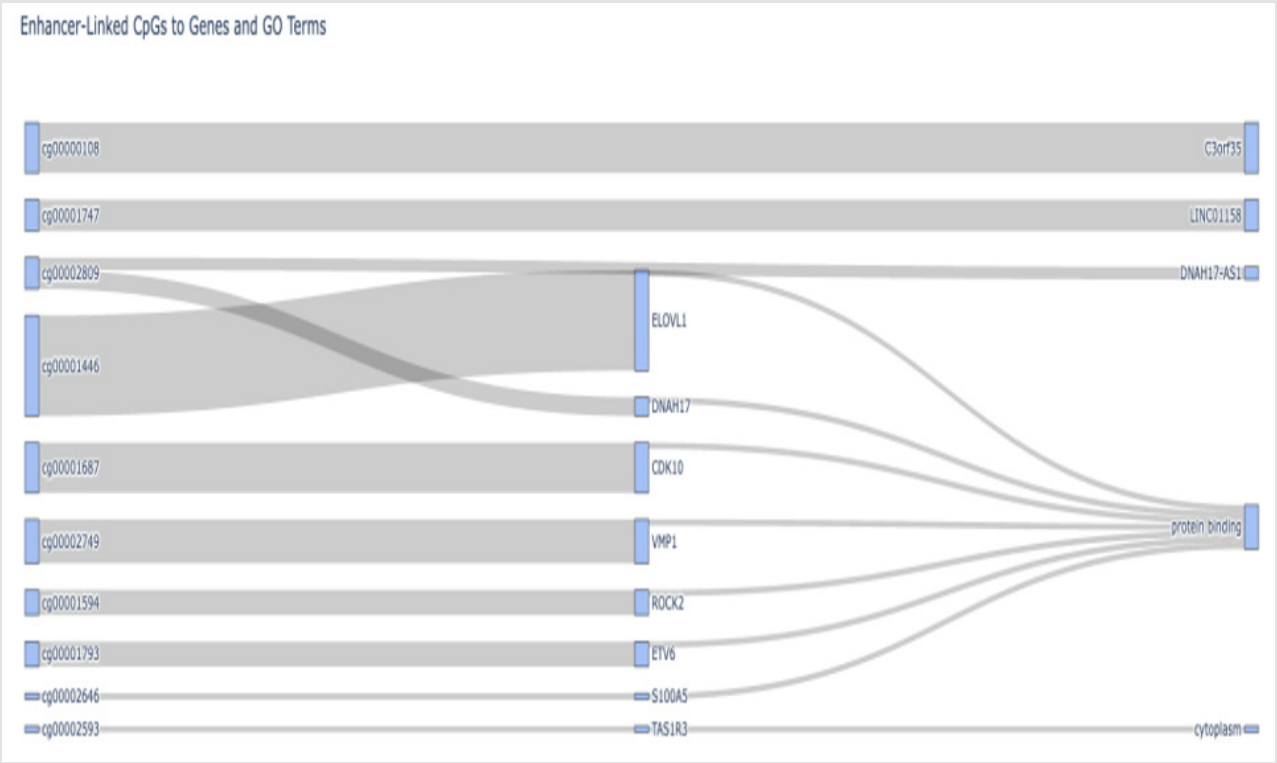

Figure 8: Network of enhancer-linked CpGs, target genes, and enriched GO terms.

In addition, we created a Sankey diagram, as shown in Figure 9, to connect CpG and gene information to the Gene Ontology (GO) terms found within the neuron ATAC accessible set. The functional information for the neuron-accessible CpG set can be seen to converge at the pathway level through the use of the Sankey diagram. The brain-specific CpG set was created through the intersection of brain-specific ATAC peaks and the use of the geneNames annotation for the target genes. The three-tiered Sankey diagram was created to show the connection between

(i) CpG loci,

(ii) The target genes, and

(iii) The biological themes using the g:Profiler enrichment tool.

For example, the CpG regions chr11:2720462-2720464 and chr17:48929686-48929688 map to the target genes KCNQ1, TSEN34, and CARMIL1, which play roles in cilia-associated signaling, RNA processing, and cytoskeletal remodeling, respectively. This indicates the possibility of epigenetic regulation of the aging brain's neuronal structure and function. The width of the connecting edge represents the number of CpG regions for the connection (Figure 9). To validate the functional relevance of these CpG-associated genes, we examined their expression patterns across GTEx tissues. Figure 10 shows median TPM expression for NIPA2 and FAM81A (among our top CpG-mapped genes) across multiple brain regions and whole blood. Both genes show substantially higher expression in brain tissues compared to blood (TPM ~20-30 in cortex vs <5 in blood), consistent with the brain-specific developmental signals captured by the cross-tissue model while blood shows complementary hematopoietic expression patterns (Figure 10).

DOI: 10.26717/BJSTR.2026.65.010270

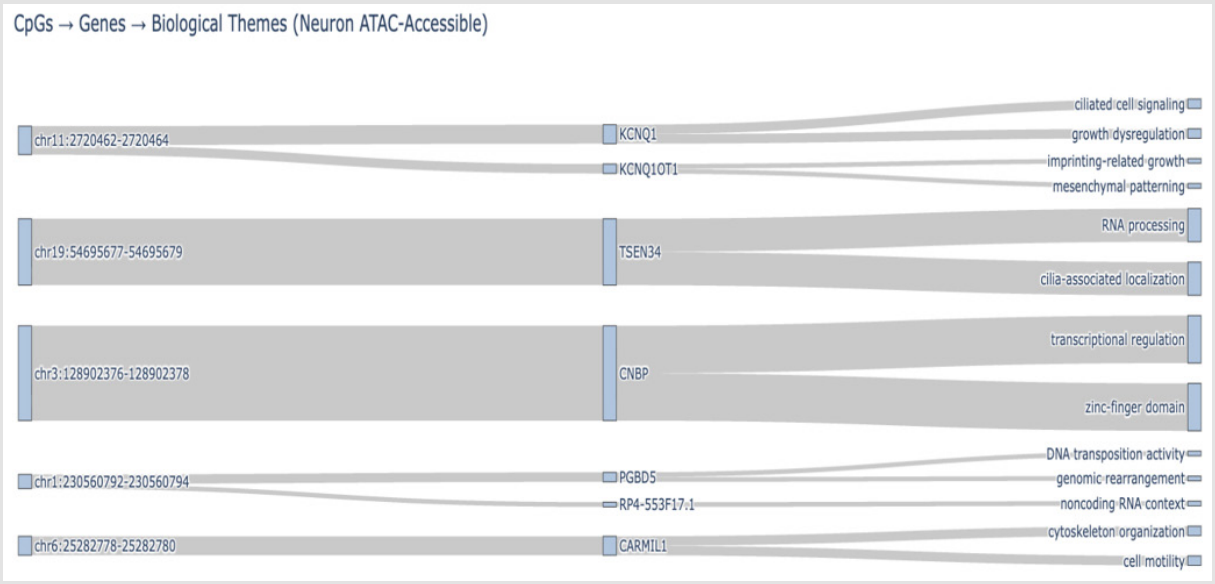

Figure 9: Sankey diagram for functional mapping of neuron-accessible CpGs.

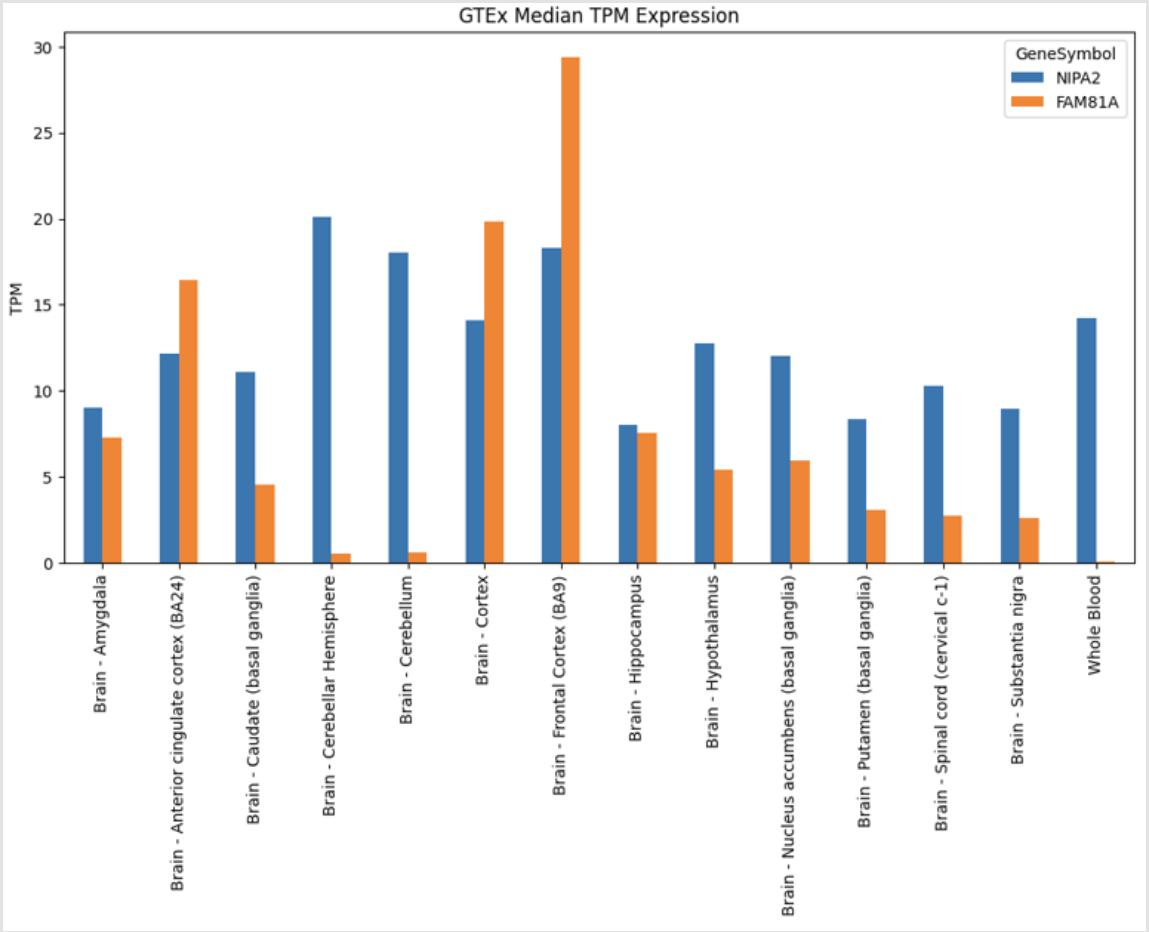

Figure 10: GTEx median TPM expression of top CpG-associated genes across brain and blood tissues.

Discussion

We analyzed public DNA methylation datasets and found that simple, interpretable machine learning models can pick up on age-related methylation patterns in both the human brain and blood. Interestingly, when we trained a model on brain data, the set of CpG sites it relied on still carried valuable information when we switched over to peripheral blood samples. The XGBoost classifier trained on brain data—explained with SHAP—highlighted a handful of CpGs that really stood out for distinguishing kids from adults and older individuals. These CpGs connect to genes known for their roles in development and aging, and many overlap with CpGs from established epigenetic clocks. That overlap adds weight to their biological significance. We also saw that enhancer-linked CpGs, which researchers have tied to gene regulation during development and age-related changes, line up with recent findings about enhancer plasticity in aging tissues (Bell, et al. [3,10]). When we applied the brain model to blood using the subset of shared CpGs, we observed that most adult blood samples were classified as not_child, but a fraction of middle-aged and older adults were assigned child-like labels with high confidence. While this cross-tissue classification is based on a limited overlapping CpG panel and does not constitute a full epigenetic age estimate, it suggests that shared methylation features can highlight individuals whose blood profiles retain brain-like developmental signatures. This observation is consistent with the idea that certain age-associated CpG changes are coordinated across tissues, although the strength and direction of these changes may vary by locus.

Our blood-specific three-class model further demonstrates that tree-based methods, particularly XGBoost, are well suited for high-dimensional epigenetic age modeling when combined with appropriate class balancing and post hoc interpretability tools. The TabTransformer architecture, although attractive conceptually for tabular data, did not consistently outperform gradient-boosted trees in this setting and required more careful tuning. For many practical applications involving 450K or EPIC arrays, tree-based models with SHAP explanations may provide a robust and transparent baseline. This study has several limitations. First, we relied on two publicly available datasets and did not include independent validation cohorts, which may limit generalizability. Second, the number of shared CpGs between brain and blood was relatively small, constraining cross-tissue analyses. Third, we used coarse age categories rather than continuous age predictions, which may obscure finer-grained epigenetic age acceleration effects. Future work could extend this framework by integrating additional tissues and cohorts, using continuous age regression models, and performing systematic comparisons with published epigenetic clocks on overlapping CpG panels (Levine, et al. [9,14]).

Despite these limitations, our results highlight a practical, reproducible pipeline for cross-tissue epigenetic age modeling using public data and modern interpretable machine learning (Hu C, et al. [15]). By focusing on compact CpG panels with clear functional annotation and cross-tissue behavior, this approach may complement existing clocks and support the development of targeted assays for aging research and personalized risk assessment. Our findings illustrate how compact CpG panels, derived through interpretable machine learning, can serve as methodological innovations for mammalian genomics. By bridging tissues, these models highlight biological networks that connect developmental and aging processes to disease pathways. Such cross-tissue epigenetic classifiers may ultimately support precision medicine by identifying individuals with youthful or accelerated methylation phenotypes, informing risk stratification and therapeutic interventions.

Statements and Declarations

Competing Interests

The authors declare no competing interests.

Author Contributions

SRK conceived and designed the study; performed data preprocessing, survival modeling, and feature attribution analyses; developed the modular AI framework; prepared figures, tables, and visualizations; drafted and revised the manuscript.

HC provided supervision and guidance on study design and methodology, reviewed and refined the manuscript for scientific accuracy and clarity.

Preprint

This manuscript has been posted as a preprint on ResearchSquare (https://doi.org/10.21203/rs.3.rs-8928610/v1).

Declarations

- **Funding:** Self funded research - No external funding was received.
- **Conflicts of Interest:** The authors declare no conflicts of interest.
- **Ethics Approval:** Not applicable.
- **Data Availability:** Public dataset GSE40279 (NCBI GEO).
- **Clinical Trial Registration:** This study does not involve a clinical trial and hence trial registration details are not applicable.
- **Consent to Publish declaration:** Not applicable.
- **Consent to Participate declaration:** Not applicable.

Acknowledgement

We thank the investigators who generated and deposited the GSE41826 and GSE40279 DNA methylation datasets in public repositories, and the maintainers of the GENCODE-based 450K manifest used for CpG annotation. We also acknowledge the developers of the open-source software libraries used in this work.

This manuscript reflects the author's original research, writing, and design efforts. AI tools were used sparingly to assist with polishing language and formatting visuals, but all scientific ideas, analyses, and interpretations were developed and validated by the author.

References

1. Horvath S (2013) DNA methylation age of human tissues and cell types. Genome Biol 14(10): R115.
2. Horvath S, Raj K (2018) DNA methylation-based biomarkers and the epigenetic clock theory of ageing. Nat Rev Genet 19(6): 371-384.
3. Bell CG, Robert Lowe, Peter D Adams, Andrea A Baccarelli, Stephan Beck, et al. (2019) DNA methylation aging clocks: challenges and recommendations. Genome Biol 20(1): 249.
4. Jain N, James L Li, Lin Tong, Farzana Jasmine, Muhammad G Kibriya, et al. (2024) DNA methylation correlates of chronological age in diverse human tissue types. Epigenetics Chromatin 17(1): 25.
5. Marioni RE, Sonia Shah, Allan F McRae, Brian H Chen, Elena Colicino, et al. (2015) DNA methylation age of blood predicts all-cause mortality. Genome Biol 16(1): 25.
6. Mendonça V, Sheila Coelho Soares-Lima, Miguel Angelo Martins Moreira (2024) Exploring cross-tissue DNA methylation patterns: blood–brain CpGs as potential neurodegenerative disease biomarkers. Commun Biol 7: 6591.
7. Huang X, Ashish Khetan, Milan Cvitkovic, Zohar Karnin (2020) TabTransformer: tabular data modeling using contextual embeddings. arXiv preprint arXiv 2012: 06678.
8. Duran I, Tsurumi A (2025) Evaluating transcriptional alterations associated with ageing and developing age prediction models based on the human blood transcriptome. Biogerontology 26(2): 86.
9. Rayevskiy S, Quinn Le, Julia Nguyen, Steven Q Chen, Christina A Castellani (2023) EpigeneticAgePipeline: an R package for comprehensive assessment of epigenetic age metrics from methylation microarrays. bioRxiv.
10. Linsenfelder S, Mohamed H Elsafi Mabrouk, Jessica Iliescu, Monica Varona Baranda, Athanasia Mizi, et al. (2025) Epigenetic editing at individual age-associated CpGs affects the genome-wide epigenetic aging landscape. Nat Aging 5: 997-1009.
11. Pipek OA, Csabai I (2022) A revised multi-tissue, multi-platform epigenetic clock model for methylation array data. J Math Chem 61: 376-388.
12. Harris CJ, Brett A Davis, Jonathan A Zweig, Kimberly A Nevonen, Joseph F Quinn, et al. (2020) Age-associated DNA methylation patterns are shared between the hippocampus and peripheral blood cells. Front Genet 11: 111.
13. Kaulagi SR, Chavan H (2026) CpG traceability and pathway mapping in epigenetic aging with explainable AI. Sciety Labs (in press).
14. Levine ME, Ake T Lu, Austin Quach, Brian H Chen, Themistocles L Assimes, et al. (2018) An epigenetic biomarker of aging for lifespan and healthspan. Aging 10(4): 573-591.
15. Hu C, Yunxiao Li, Longhui Li, Naiqian Zhang, Xiaoqi Zheng (2024) BS-clock: advancing epigenetic age prediction with high-resolution DNA methylation bisulfite sequencing data. Bioinformatics 40(11): btae656.

ISSN: 2574-1241

DOI: 10.26717/BJSTR.2026.65.010270

Suresh Kaulagi. Biomed J Sci & Tech Res

This work is licensed under Creative *Commons* Attribution 4.0 License

Submission Link: https://biomedres.us/submit-manuscript.php

Assets of Publishing with us

- Global archiving of articles
- Immediate, unrestricted online access
- Rigorous Peer Review Process
- Authors Retain Copyrights
- Unique DOI for all articles

https://biomedres.us/