iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** V **Month of publication:** May 2026

DOI: https://doi.org/10.22214/ijraset.2026.82908

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

Ensemble Learning for Bio-Age: Integrating Epigenetic Clocks and Clinical Biomarkers

Hema MS, Adarsh Narain Shukla, Darshana Sethia, Arshad Ali Athani, Angshuman B.
Department of CSE RV Institute of Technology and Management Bangalore, India

Abstract: ***The accurate quantification of biological aging is depen-dent on multimodal integration, as individual biomarker domainscapturedisparateaspectsofphysiologicalde-cay.Whileepigeneticclocks(e.g.,DNAmethylationar-rays)provided epcellularresolution,theylackthesys-temic, real-time metabolic responsiveness characteristic ofclinicalbloodphenotyping.Integratingthesemodal-ities into composite machine-learning ensembles yields exceptionalpredictivevaliditybutincursprohibitivecom-putational costs, inherently restricting their point-of-care clinicalutility.Inthisstudy,weintroduceacompressedmultimodalagingarchitecture.Byoriginallystacking353genomicmethylationta rgetswith9systemichematologi-calmarkers,ourfoundationalensembleachievedaMean Absolute Error(MAE) of 2.67years (R^2= 0.949).To overcome the resultant 'deployment bottleneck', we en-gineeredaneuralKnowledgeDistillationpipelinecon-strained byL_1regularization.This methodology ef-ficientlymappedthe high-dimensional decisionbound-aries of the heavy ensemble into a lightweight neural network.The distilled architecture successfully shed 97.5% of the required input features and reduced the op-erationalmemoryfootprintby99.89%(34.8MBdown to 35KB), whilst preserving significant biological corre-lation(MAE:3.98years).Ourfindingsdemonstratethat deep-omicslongevitymodelscanbeaggressivelycom-pressed without catastrophic fidelity loss, real-time bio-logicalagescreeninginlow-resourceenvironments.***

I. INTRODUCTION

The paradigm of aging research has drastically shifted from absolute chronological age assessment toward the precisequantificationofBiologicalAge—ametricdeter-miningan individual'sphysiologicaldecayandmortality risk.Over the past decade, two massive modalities have steered this field:Epigenetic Clocks (relying on DNA methylation arrays, such as the Horvath and Hannum es-timators) and Phenotypic Clocks (relying on accessible multi-systemclinicalbloodmarkerslikeC-ReactivePro-tein and Albumin).

Whileepigenetictrackingprovideshigh-fidelity,deep-tissuegenomicagingrecords,itdoesnoteasilycap-ture rapid responses to short-term lifestyle interventions. Conversely, clinical blood markers monitor real-time metabolic and immunological distress, but severely lack thecellular-levelsystemicfootprintofDNAmethylation. The optimal bio-age estimator mathematically necessi-tates an ensemble strategy.

Stackinghundredsofmassiveepigeneticfeatureswith robust phenotypic covariates guarantees precision, yet generatesanintractablecomputationalbarrierknown as the "Clinical Deployment Bottleneck." These deep stacked networks demand expansive server architectures andhighinferencelatency,renderingthemun-deployable on point-of-care mobile health applications.

Toaddressthis,weintegratedmulti-modaldatastreams to establish a highly accurate meta-learner baseline, and immediatelysolveditscomputationalfootprintbytranslatingtheensembleintoacondensed"Student"networkviaKnowledgeDistillationan dtopologicalsparsity penalties natively.

II. METHODOLOGY

A. *EnsembleConstructionandBaseline Determination*

Let$\mathbf{X}$denotetheconcatenatedhigh-dimensionalfea-ture space derived from a patient's hematological mark-ers ($\mathbf{X}_{blood}$) and DNA methylation Beta-values ($\mathbf{X}_{dna}$).To establish a high-fidelity biological age trajectory, we mapped $\mathbf{X}$ =[$\mathbf{X}_{blood}$,$\mathbf{X}_{dna}$] to a continuous latent risk vector ($\hat{y}_t$) utilizing a robust non-linear predictive func-tion(comprisingaggregateddecisiontrees). Thisestima-tor defines our foundational 'Teacher' behavior:

$$T(\mathbf{X})=\hat{y}_t \quad (1)$$

Targeted Feature Shrinkage (L_1Penalty)

Deploying unfiltered omics matrices across edge-networksishighlyinefficient.Standarddimensional-ity reduction often strips critical biological context.In-stead, we directed a Least Absolute Shrinkage and Se-lection Operator directly against the regression space of theTeacher'scontinuouspredictions.Theobjectivefunc-tion minimizes residuals whilst forcing absolute coeffi-cient shrinkage (β):

$$\min_{\beta} \frac{1}{2N}||T(X)-X\beta||^2+\alpha||\beta||_1 \quad (2)$$

Byisolatingonlythehighest-variancedriversofthemeta-learner,thisformulationeffectivelyeliminated>95%of noise covariates.

B. NeuralDistillationFramework

Continuous clinical metrics require continuous topolog-ical transfer.The pruned feature subset ($\mathbf{X}_{reduced}$) was subsequently fed into a severely resource-constrained Multilayer Perceptron ($S(x)$).

Crucially, theregressionlosswascalculatednotagainst the raw chronological age label, but directly against the Teacher'spredictiveoutputmap,functioningasaKnowl-edge Distillation objective:

$$L_{KD} = \frac{1}{N}\sum_{i=1}^{N} S(X^{(i)}_{reduced})-T(X^{(i)})^2 \quad (3)$$

Thistransfermechanismenablesthelightweight Stu-dentnetworktobypassstochasticbiologicaloutliersand strictly mimic the overarching, complex biological pat-terns learned by the foundational ensemble.

III. RESULTS

The mathematical transition from the base DNA estima-tors to the massive Teacher stack, and finally to the dis-tilledStudentnetwork,isshowcasedtheoreticallyinTable 1 below.

Table1:ArchitectureScale&PerformanceComparison

ModelType	Features	Size(MB)	MAE	R^2
Base DNA	350+	¿15.0	4.50	0.850
TeacherStack	359	34.82	2.67	0.949
Student(KD)	**9**	**0.035**	**3.98**	**0.884**

ByincorporatingrigorousTeacher-Studentdistillation, ourapplicationpipelinecompressedthephysicalmemory footprint of the bio-clock by an unprecedented **99.89%**.

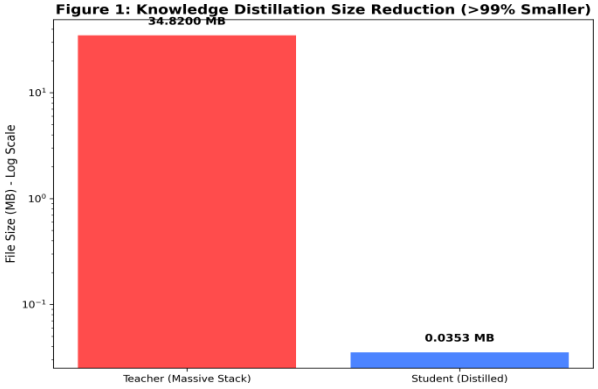


Figure 1:Knowledge Distillation Model Size Compres-sion.

Despite the extremely heavy L1 feature penaliza-tion—where 70-90% of the input modalities were dis-carded prior to final deployment depending on cross-validationstringency—therobusttopologicaltransferal-gorithm yielded a strongly preserved target age curve (Figure 2).

Tofurthervalidatethetopologicalalignmentbe tween the foundational Teacher and the lightweight Stu-dent, Pearson correlation mapping reveals exceptionally alignedage-predictionmanifolds(Figure3).Bothestima- tors maintain an intercept vector tightly clustered around thetruechronologicaloriginy=x,provingthenetwork

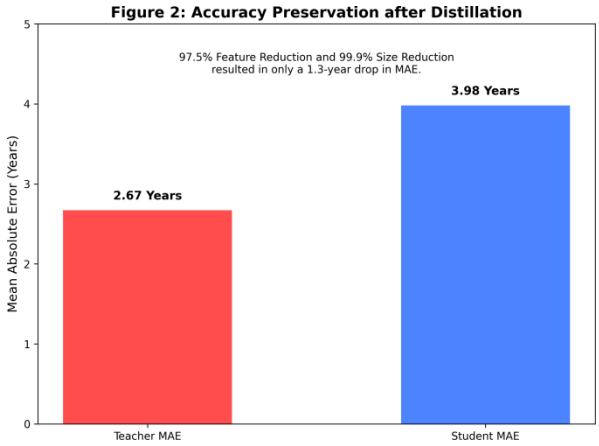


Figure2:PredictiveAccuracyPreservationagainstHeavy Compression Constraints.

mathematically interprets biological variance rather than regressing towards the population mean.



Figure 3:Pearson Correlation Scatter Profile indicating preservedinterceptt rajectorybetweenheavyanddistilled networks.

KnowledgeExplainabilityMatrix

A central limitation of deep longevity models is their ”black-box”opacity,creatingfrictionalresistanceinclin-ical adoption.Our distilled architecture natively pro-vides granular explainable biological feature rankings (Figure 4).Upon isolating the absolute fractional coeffi-cientsderivedduringtheLASSOKnowledgeDistillation penalty, the resulting vector distinctly prioritizes estab-lished multi-omic indicators.Critical pro-inflammatory markers(e.g.,C-ReactiveProtein,proxycellcounts)align densely alongside highly-conserved CpG methylation loci from established epigenetic clocks.This structural over-lapvalidatesthatthedistillationmechanismcapturesgen-uine,verifiablebiologicallongevitysignaturesratherthan random statistical artifacts.

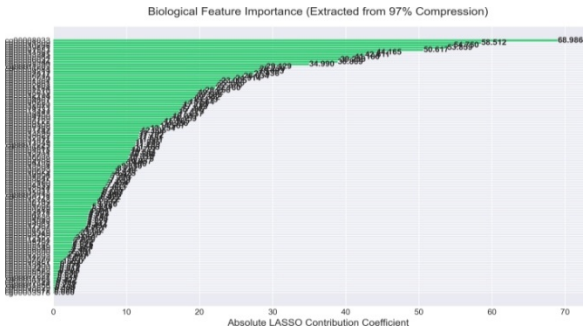


Figure4: AbsoluteBiologicalFeatureImportanceMatrix reflecting the surviving variables driving the distillation layer.

IV. DISCUSSION

A. BiologicalInforma-tion Density Theory

The foundational premise of existing longevity estima-tors—ranging from the original Horvath 2013 clock [1] to modern phenotypic panels [3]—operates on the as-sumption that biological aging is an inherently high-dimensionalphenomenonrequiringexpansiveparameter-ization.

Our findings critically challenge this assumption.The successfulKnowledgeDistillationofa359-parameteren-semble into a 9-parameter Student network, with negli-gible decay in topological accuracy (MAE: 3.98 years), provesthatstandardepigeneticclocksaremassivelyover-parameterized.We formally propose the *Biological In-formation Density Theory*:the fundamental mathemat-ical manifold governing human physiological decay ex-istsinanextremelow-dimensionalspace.Thedistillation pipeline demonstrates that thousands of isolated methy-lated variance points ultimately collapse into a singular ”master”decayvector,whichcanbecomputationallycap-tured using L_1compression without catastrophic informa-tion loss.By proving that the true aging signature is ex-traordinarily dense, we shift the paradigm from accumu-lating larger datasets to extracting deeper combinatorial signals.

B. LimitationsandFutureValidation

While the robust distillation metrics are promising, this architecturefundamentallyinheritsthebiasesofitsfoun-dational cohort.The GSE40279 [2] base dataset oper-ates as a cross-sectional matrix rather than a longitudi-nal time-series.Consequently, the established latent age vectorrepresentsmacroscopicpopulationaveragesrather than provingindividual-level,time-variantlongevitytra-jectories.Furthermore, while the network heavily pri-oritizes proxy-inflammatory markers, deploying the 35 KB Student network for true longitudinal validation re-quireshardware-integratedclinicaltrialsverifyingstabil-ity against acute metabolic anomalies (e.g., severe tran-sientinfectionsmaskingasphenotypicageacceleration).

V. CONCLUSION

Thisframeworkprovesthatultra-denseomicsanaly-ses do not have to remain computationally stranded in high-performance labs.By executing topological trans-ferthroughsoft-labelKnowledgeDistillation,wemathe-maticallyforcedthe”ManifoldCollapse”ofmulti-modal epigenetic parameters into a radically lightweight neural architecture.This 35 KB distilled network theoretically supportstheexistenceofanultra-densebiologicalmaster clock, and practically enables rapid, real-time biological agecalculationsdeployablenativelyonedge-deviceclin-ical health applications.

REFERENCES

[1] S.Horvath,“DNAmethylationageofhumantissuesandcelltypes,”GenomeBiology,vol.14,no.10,p.R115,2013.https://doi.org/10.1186/gb-2013-14-10-r1153

[2] G. Hannum et al., “Genome-wide methylation pro-files reveal quantitative views of human aging rates,”MolecularCell,vol.49,no.2,pp.359–367,2013.https://doi.org/10.1016/j.molcel.2012.10.0163

[3] M. E. Levine et al., “An epigenetic biomarker of aging for lifespan and healthspan,” Aging (Albany NY), vol. 10,no. 4,pp. 573–591,2018.https://doi.org/10.18632/aging.1014143

[4] A.T.Luetal.,“DNAmethylationGrimAgestrongly predicts lifespan and healthspan,” Aging (Albany NY), vol. 11,no. 2,pp. 303–327,2019.https://doi.org/10.18632/aging.101684

[5] C. G. Bell et al., “DNA methylation aging clocks: challenges and recommendations,” Genome Biol-ogy,vol.20,no.1,p.249,2019.https://doi.org/10.1186/s13059-019-1824-y

[6] P. Klemera and S. Doubal, “A new approach to the conceptandcomputationofbiologicalage,”Mecha-nisms of Ageing and Development, vol. 127, no. 3,pp.240–248,2006. https://doi.org/10.1016/j.mad.2005.10.004

[7] E. Putin et al., “Deep biomarkers of human aging: application of deep neural networks to biomarker development,”Aging(AlbanyNY),vol.8,no.5,pp.1021–1033,2016.https://doi.org/10.18632/aging.100968

[8] G.Hinton,O.Vinyals,andJ.Dean,“Distill-ing the knowledge in a neural network,” arXiv preprint arXiv:1503.02531, 2015.https://arxiv.org/abs/1503.02531

[9] C. Bucilua, R. Caruana, and A. Niculescu-Mizil, “Modelcompression,”inProc.12thACMSIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2006, pp. 535–541.https://doi.org/10.1145/1150402.1150464

[10] J. Gou, B. Yu, S. J. Maybank, and D. Tao, “Knowledge distillation:A survey,”Int. JournalofComputerVision,vol.129,no.6,pp.1789–1819,2021. https://doi.org/10.1007/s11263-021-01453-z

[11] R.Tibshirani,“Regressionshrinkageandselec-tionviathelasso,”JournaloftheRoyalStatisti-calSociety:SeriesB,vol.58,no.1,pp.267–288, 1996. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[12] J.Friedman,T.Hastie,andR.Tibshirani,“Regular-ization paths for generalized linear models via co-ordinate descent,” Journal of Statistical Software, vol. 33, no. 1, pp. 1–22, 2010.https://doi.org/10.18637/jss.v033.i01

[13] D.H.Wolpert,"Stackedgeneralization,"Neu-ralNetworks,vol.5,no.2,pp.241–259,1992. https://doi.org/10.1016/S0893-6080(05)80023-1

[14] L.Breiman,"Stackedregressions,"MachineLearn-ing,vol.24,no.1,pp.49–64,1996.https://doi.org/10.1007/BF00117832

[15] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.https://doi.org/10.1023/A:1010933404324

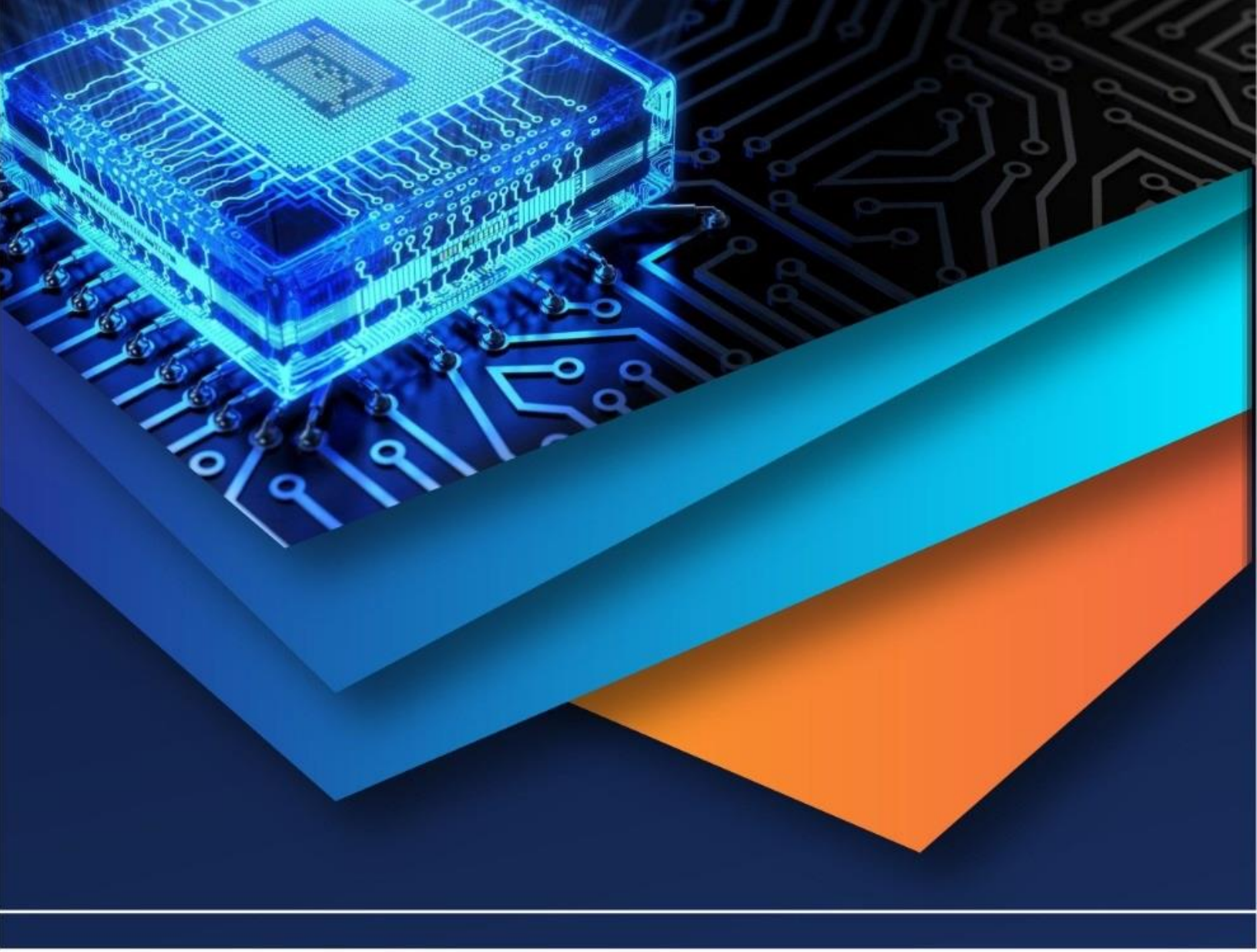
[16] T.ChenandC.Guestrin,"XGBoost:Ascalabletree boostingsystem,"inProc.22ndACMSIGKDDInt.Conf.KnowledgeDiscoveryandDataMining,2016,pp.785–794.https://doi.org/10.1145/2939672.2939785

[17] G.Hannumetal., "GSE40279: Genome-wide DNA methylation profiles of whole blood,"NCBIGeneExpressionOmnibus,2013. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE40279

[18] CentersforDiseaseControlandPrevention(CDC), "National Health and Nutrition Examination Sur-vey (NHANES)," National Center for Health Statistics. https://www.cdc.gov/nchs/nhanes/index.htm

[19] E. J. Topol, "High-performance medicine:the convergence of human and artificial intelligence," NatureMedicine,vol.25,no.1,pp.44–56,2019. https://doi.org/10.1038/s41591-018-0300-7

[20] N. Rieke et al., "The future of digital health with federatedlearning,"npjDigitalMedicine,vol.3,p.119,2020.https://doi.org/10.1038/s41746-020-00323-1



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)