Age prediction using mammalian methylation profiles

Group 5 | Kunal Chaudhary, Tim Wehnes, John Collins, Matyas Iras, Gabriel Haw

Abstract

DNA methylation plays a complex role in cell physiology and has previously been shown to be able to predict age. Recently, the Mammalian Methylation Consortium released a dataset containing arrays of conserved methylation sites of over 11.000 mammals. We use linear regression, decision trees, and deep learning to re-create a pan-tissue epigenetic clock. Our results indicate the superiority of using sparse, linear models for the given task. Furthermore, we confirm the particular importance of certain genes in the aging process, as well as how some organisms seem to age much slower altogether.

1 Introduction

Generally, all cells within a multicellular organism contain near-identical DNA, but variable gene expression causes observable cell differentiation[1]. Gene activity changes not linked to alterations in DNA sequence are heritable and labeled as Epigenetics[2]. One major contributor to epigenetic variability is DNA methylation, the transfer of a methyl group from S-adenosyl methionine (SAM) to a cytosine residue in the DNA molecule by a group of enzymes called DNA Methyltransferases (Dntms)[2]. These modifications are not spread uniformly; cytosines that precede a guanine (CpG sites) are more likely to be methylated[2]. CpG sites are rare in mammalian genomes because methylated cytosines (5mC's) have an increased likelihood to mutate into the residue Thiamine[3]. Only one percent of the entire genome consists of 5mC's, making them both infrequent and influential to gene function[4, 2]. Certain conserved regions within the genome contain significantly more CpG sites relative to other bases and are called CpG islands[5]. These islands are crucial for proper cell function because they contain the majority of gene promoters and are thus usually hypomethylated[6].

As such, the role of DNA methylation in aging and age-related diseases is an active field of research. Researchers discovered that hypomethylation of regions found in non-CpG islands and hypermethylation of cytosines within CpG islands is more prevalent in older individuals[7]. This might be due to the relationship of epigenetic aging with nutrient sensing, mitochondrial activity and stem cell composition[8]. This association has led scholars to build statistical models able to predict age in humans from certain key methylation sites, even across different tissues[9, 10]. These so-called "epigenetic clocks" have been demonstrated to be superior to traditional chronological age in approximating biological age[11]. More recently, it has been proposed that epigenetic clocks might work across all mammalian species[12], as many CpG islands are conserved among them. This led to the development of a methylation array for mammals covering more than 36.000 conserved CpG sites[13].

In this paper, we use the dataset created with said array, the Mammalian Methylation Consortium[14], to investigate if machine learning models can be trained on pan-tissue methylation profiles to accurately predict age. In particular, we are interested in the difference in performance between linear, tree-based and neural network models. Additionally, we would like to investigate the most-predictive features further to check their biological significance.

In total, we extracted recorded sample information for 12.282 mammals from series matrix and SOFT files, ignoring entries where relative age could not be calculated due to missing values. For our features, we used normalised β values, representing methylation levels for 37.554 CpG sites. As our outcome variable, we calculated the relative age(1) of each organism, which combines a ratio of chronological age and maximum lifespan, while adding the species gestation period to remove any negative values due to samples collected pre-birth.

$$\text{relative age} = \frac{\text{gestation period} + \text{chronological age}}{\text{gestation period} + \text{maximum lifespan}} \tag{1}$$

We reproducibly randomized and divided the data in a $80:20$ split, giving us a test set of 2437 samples. Scikit-learn[15] library is used for splitting, evaluation and training of some models.

2 Softwares and Methods

2.1 Linear Methods

We wanted a baseline model to compare subsequent models' performances. To do so, we applied a basic linear regression model to our training set. We used it to predict the transformed age variable for all instances in our training set. This was done to run further analysis on the residuals and investigate whether the five assumptions that a linear regression model presumes were satisfied.

In particular, we used a residuals vs predicted values plot to ensure linearity of residuals and rule out any homoscedasticity. Also, we used a Q-Q plot to derive clues about normality of residues, and a Durbin-Watson test to verify their independence. While all those four assumptions were satisfied, obtaining a metric for the level of multicollinearity in the data proved to be challenging as the variance inflation factor (VIF) runs into division-by-zero errors with wide matrices. Since prior works have elaborated on the common issue with multicollinearity in high-dimensional omics data [16] , we assumed that our data was also subject to this constraint. As such, we figured there might be a moderate to high level of multicollinearity and devised strategies for dealing with varying levels of such. In case our dataset was subject to a moderate amount of multicollinearity, we figured a LASSO model might be appropriate, as prior research has shown its effectiveness in dealing with the given issue[17].

After tuning hyperparameters for LASSO (see subsection: Hyperparameter tuning), we also trained a Partial Least Squares Regression (PLSR) model in case there was such high level of multicollinearity in our data that dimensionality reduction was in order. The advantage of PLSR over a PCA + regression approach lies in its ability to capture regions in the feature space that are important to the dependent variable but show overall small variance and would hence be missed by a PCA. Based on our prior PCA analysis, we had a rough idea of how many components might be needed for a good fit and hence trained PLSR models utilizing 20, 50, 100, 160 and 223 components.

Finally, to check how each of our models performed, we utilized five-fold cross validation and compared not just validation, but also training scores for each (Figure 1a). This gave us an idea for the propensity of each model to over- or underfit the data. In fact, it was clear that the basic linear regression model and any PLSR trained on over 50 components were too likely to overfit the data. Meanwhile, PLSR 20 proved to be quite error-prone and hence was likely underfitting the data. The two best performing models were quite clearly LASSO and PLSR50.

Based on the promising results of LASSO, we figured an elastic net that combines L1 and L2 regularization might further boost performance as it can reduce some of the bias introduced by the feature elimination of L1 regularization [18]. Thus, we utilized Optuna again to run a hyperparameter search and subsequently tested the model using five-fold cross validation. As can be seen in Figure 1b. this train of thought proved successful and further improved model performance on the validation sets across all error metrics.

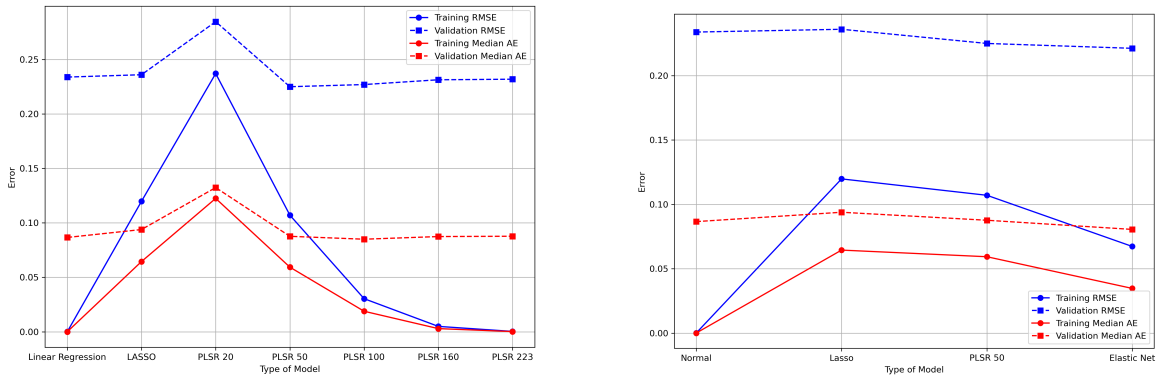
2.2 Decision Tree Methods

To overcome the limitations of linear regression models in capturing the non-linear dynamics between age and CpG site methylation, we utilised the Light Gradient Boosting Machine (LGBM) model. This gradient boosting framework enhances performance accuracy through the merging of predictions coming from multiple decision trees. We chose LGBM because it employs a histogram-based algorithm that bins continuous feature values into discrete bins to efficiently identify the optimal split. This approach considerably reduces the number of potential splits, thereby lowering the computational burden. LGBM does not assume a linear relationship between features and the target, enabling it to effectively model data where residuals do not follow a normal distribution, an important consideration in biological networks.

Here, we employed 5-fold cross-validation whereby we applied early stopping and pruning on each fold to reduce computational time along with over-fitting. In addition, we also leveraged LGBMs ability to quantify feature importance, which allowed us to interpret the models decision making process and the identify the most influential features when predicting relative age. We assessed this using two methods, namely Gain and Split (see **Figure 2**).

Gain attributed importance measures the average gain of a feature when used to construct a tree. This therefore reflects the contribution of each feature to the model when considering the total reduction of loss that is achieved by splits on this feature, summed over all trees within the model. A higher gain value for a feature means that it is more important for generating accurate predictions.

Split attributed importance, on the other hand, counts the number of times a feature is used to split a node across all the trees. This means it does not assess



(a) **(b)**

Figure 1: left Comparison of cross-validation scores for different regression models, illustrating the trade-off between complexity and performance, with LASSO and PLSR50 showing optimal balance. **right** Enhancement in predictive performance with the application of elastic net regularization, evidenced by improved validation scores across error metrics, showcasing the efficacy of combined L1 and L2 regularization.

how much a feature improves the models performance but simply how frequently that feature is used.

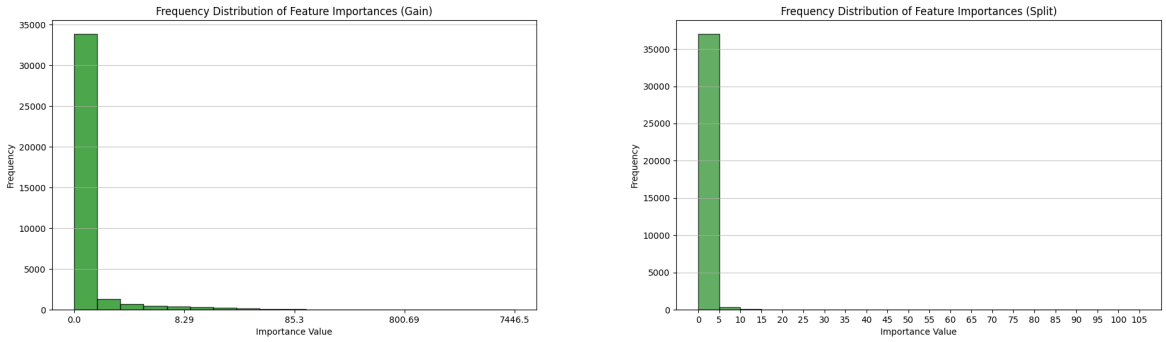
These metrics in turn allow us to assess which features contribute most to the non-linear dynamics between age and CpG site methylation and the models reliance on certain features for prediction.

2.3 Neural Network Methods

In our exploration of machine learning methodologies to predict relative age from mammalian methylation profiles, neural networks stand out due to their ability to model complex, non-linear relationships. The Keras library was used to construct sequential neural network models, employing dense (fully connected) layers due to their effectiveness in regression tasks. Utilizing the Keras library abstracts away many of the complexities associated with lower-level libraries, thereby simplifying and accelerating model construction and testing. The sequential nature of the model simplifies the process of adding layers, allowing for a consistent method for model construction. The choice of dense layers is motivated by their suitability for regression tasks, where the goal is to predict a single continuous outcome. Dense layers also allow for feature integration, where information from all neurons is propagated through the layers, and their ability for hierarchical feature learning, where lower layers can recognize basic patterns in the data while deeper layers can integrate these basic patterns into more complex representations that relate directly to biological age. Training was conducted over 100 epochs for both hyperparameter testing and the final model, ensuring consistency in model evaluation.

The ReLU (Rectified Linear Unit) activation function was utilized to introduce non-linearity into the model, a factor which we reasoned may result in performance improvements over linear- and decision tree-based methods. The ReLU is favored in many deep learning models for its ability to alleviate the vanishing gradient problem. The vanishing gradient problem occurs when the gradients of the model's loss with respect to the weights become very small, effectively halting the learning process during backpropagation because these small gradients provide little to no signal for updating the weights, and is particularly problematic for sigmoid and tanh activation functions.

Two methods from the Keras library were used to optimize the navigation of the loss landscape: the Adam optimizer and the ReduceLROnPlateau callback. Optimizers are methods used in deep learning models for iteratively adjusting the weights of the network in a way that navigates the loss landscape towards a minimum. We selected the Adam optimizer due to its ability to dynamically adjust learning rates for each parameter based on the gradients, optimizing the training process by reducing the impact of varying gradient magnitudes across different parameters. This can help the model navigate the loss landscape more smoothly and potentially avoid getting stuck in local minima early in training. On the other hand, the ReduceLROnPlateau callback modulates the learning rate in response to the model's performance, reducing it to facilitate more re-



(a) Gain **(b)** Split

Figure 2: Frequency distribution of important features determined by LGBM using the Gain and Split methods. Most features exhibit minimal importance (lower importance values) in terms of predictive accuracy and their role in decision-making. Only a select few features significantly contribute to the model's age predictions.

fined adjustments in the model's weights. The parameters used for ReduceLROnPlateau for all of the models was a starting learning rate of 0.001, a minimum learning rate of 0.00001, and a reduction factor of 0.5 if the model's MSE loss on the validation set did not improve by at least 0.005 over the span of 5 epochs. This approach helps in fine-tuning the model's convergence, navigating the loss landscape more effectively by taking smaller steps when close to a minimum, thereby enhancing the model's ability to converge on a more optimal solution and ensuring that the model does not become trapped in local minima or overshoot the global minimum.

2.4 Hyperparameter tuning

Grid search and Optuna[19] were jointly used for hyperparameter selection across our models. Grid search exhaustively searches through the model space by testing the performance of all possible combinations of a predefined set of hyperparameters for a model. Optuna is a hyperparameter selection framework which leveraged Bayesian optimization techniques and trial pruning to optimize the search of vast model spaces. Compared to grid search, Optuna can be a more resource efficient strategy for tuning hyperparameters due to the combinatorial explosion imposed by exhaustive grid search methods. However, if the search space is constrained to only a relatively small number hyperparameters in a small number of discrete steps, grid search methods remain a viable technique for hyperparameter tuning.

Linear Models: Although linear models have fewer hyperparameters compared to both tree and deep learning based methods, optimising these hyperparameters can still significantly impact model performance. Thus, we explored the effect of both L1 and L2 regularisation parameters. For the initial least absolute shrinkage and selection operator (LASSO) hyperparameter tuning, we employed both grid search and Optuna. To optimize the elastic net, we only referred to Optuna. Despite its computational efficiency, we were still limited to a relatively small search space, few iterations per trial and a small number of trials due to resource constraints.

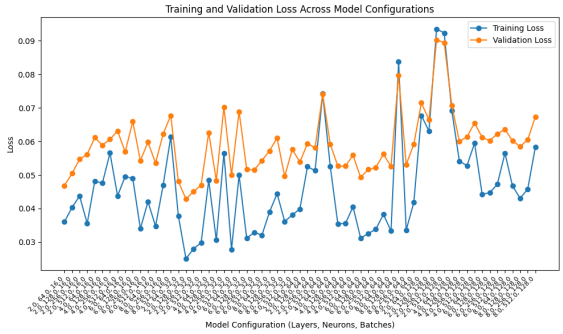


Figure 3: Loss on training and validation sets for all deep learning model configurations during hyperparameter tuning using grid search. Non-converging models (validation loss > 0.2) have been removed.

Tree based Models: Using the Light Gradient Boosting Machine (LGBM) framework, we investigated the impact of various hyperparameters including the number of leaves, learning rate, feature fraction, as well as the previously mentioned regularization parameters. Given the absence of prior published studies utilising this model in a similar setting, we opted to explore

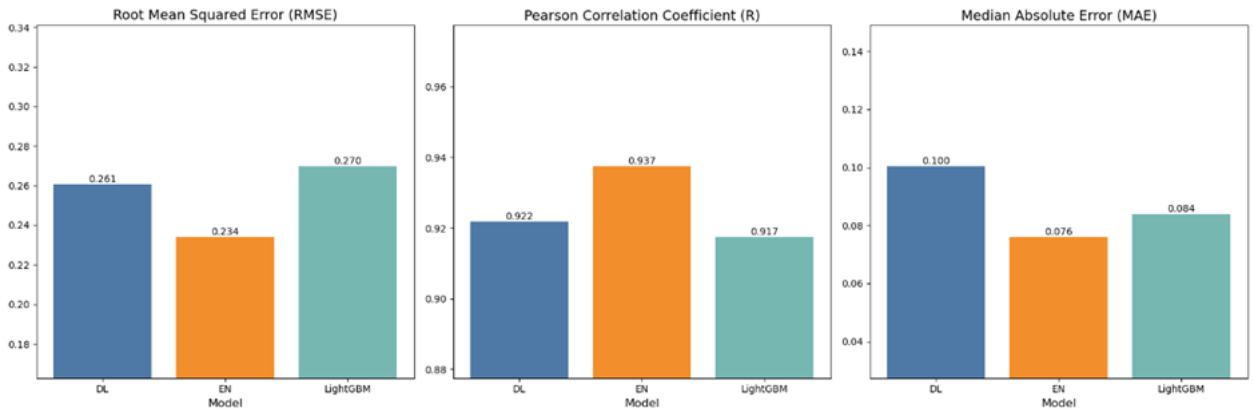


Figure 4: Performance comparison of elastic net, deep learning, and LGBM models on the test set, demonstrating the superior prediction accuracy of the elastic net model across RMSE, MAE, and Pearson's R, underscoring its robustness in age prediction from methylation profiles.

a wide range of potential values for these hyperparameters. This approach aimed to foster a comprehensive understanding of how these parameters influence the predictive accuracy of our model.

Deep Learning Model: We used a constrained grid search approach to explore the model space. We varied the number of hidden layers (**2**, 4, 6, 8), nodes per layer (64, **128**, 256, 512), and the batch size (16, **32**, 64, 128), resulting in 64 models being tested. Additionally, we experimented with configurations of the ReduceLROnPlateau callback parameters; however, the parameters used in the callback remained consistent across all models reported in figure 3. We selected the model with lowest validation loss (parameters in bold). Due to computational and time limitations, we deemed a more exhaustive hyperparameter tuning approach, such as Optuna, infeasible.

3 Results

When comparing the results of our three models on the test set, it becomes apparent that the elastic net outperforms the deep learning and tree-based models across all metrics. That is, the elastic net performs better with an outlier-sensitive metric like the RMSE, but also with more balanced error terms like the MAE. Additionally, Pearson's R coefficient also displays higher correlation between predicted vs actual values for the elastic net compared to the other two models.4

Comparison between the deep learning and LGBM model yields insights into their respective strengths and weaknesses. While the deep learning model has a slightly lower RMSE, it also seems to have a much higher median absolute error. As such, it seems safe to conclude that the LGBM is more sensitive to outliers in the data than the sequential neural network. In terms of overall correlation, both fare quite similar.

Important to note is that drawing conclusions on model superiority based on their performance might be susceptible to overfitting the test set data since three models, rather than one, have been evaluated. Hence, although the elastic net performed the best in this case, findings might differ with another train-test split.

3.1 Species Performance

In order to figure out where model accuracy was highest and lowest, we decided to investigate error terms per species per model. While we initially considered sample size per species in the overall dataset to be indicative of how well they would be classified in the test set, we surprisingly found the opposite. In fact, although the ten species for which age prediction was the least accurate all had sample sizes below the mean of 91.66 samples per species, many were still at or above the median of 7 samples per species. The disparity of mean versus median is to be attributed to the very high number of samples stemming from human and mice tissues, hence skewing the distribution. Additionally, many of the best predicted species had similarly small sample sizes, hence indicating that there might be another factor at play.

Furthermore, we observe an improved predictive performance for both the neural network and LGBM models with increased species sample sizes, indicating a significant reliance on data volume for accuracy. This is particularly evident in the case of underrepresented

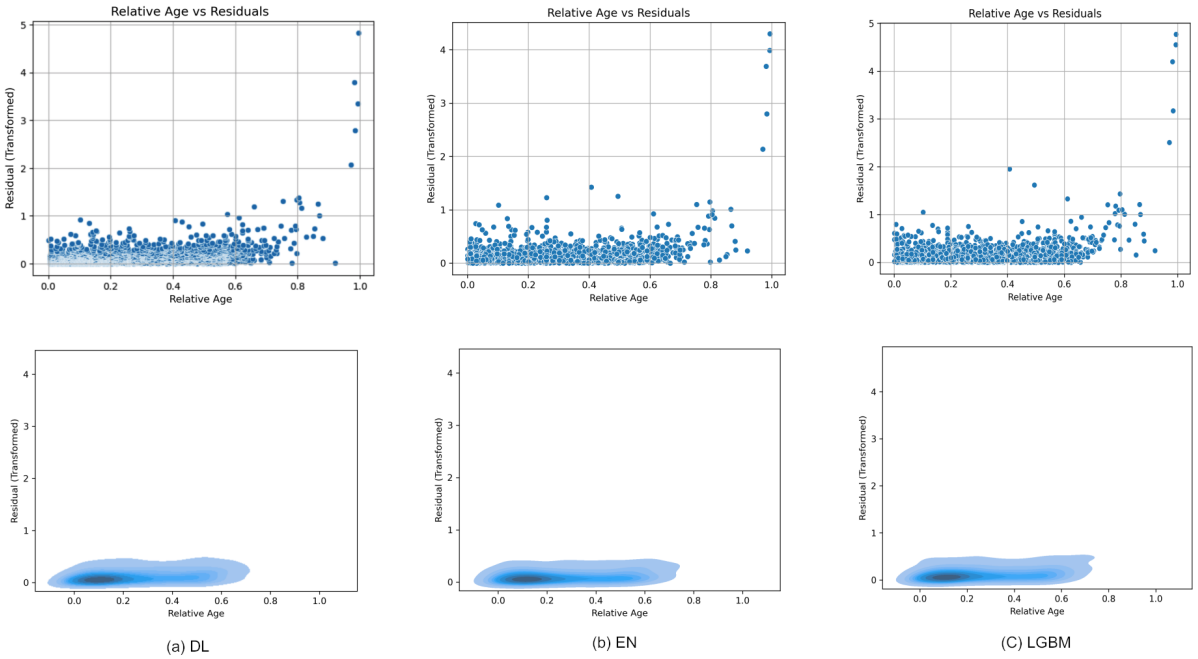


Figure 5: The **upper** figures illustrate the Kernel density estimation of the two-dimensional space of relative age and residuals. Plots show a significantly higher probability density between 0 and 0.2 of relative age at a fairly low residual for all models. In the **lower** figures the Scatter plots reveal prediction errors across age groups, highlighting the models' challenges in accurately predicting ages for older samples, with significant outliers near the maximum lifespan.

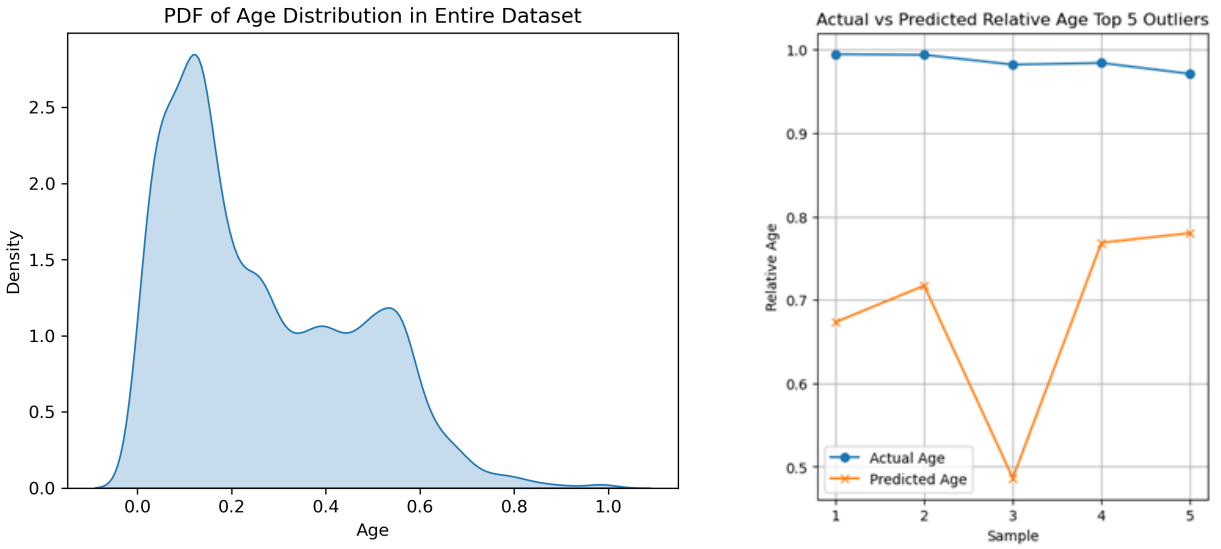
species, where performance declines due to difficulties in capturing complexity, increasing the risk of either underfitting or overfitting. Conversely, the elastic net model, which is optimised for sparse data scenarios, demonstrates enhanced effectiveness with smaller datasets by prioritising key features. This is evident as the elastic net maintains its predictive accuracy, even when dealing with the underrepresented species samples.

Additionally, it is important to note that a slightly unclear relationship between species sample size and prediction accuracy is in line with Steve Horvath's general idea of a clock that is able to predict age across species using CpG sites that are conserved across them. If this weren't the case, the entire model would most likely not work.

To further investigate where the majority of the error originates from, analysis on residuals for various age groups has been conducted. Specifically, those were aimed at finding out why the residuals of all three models were majorly skewed, displaying values of around 10, and showcasing severe leptokurtosis with scores of around 150 to 200.

As indicated in Figure 5, the kernel density function is mostly biased towards samples in the first quartile of their lives, with residuals being fairly constant across all age ranges. This result was almost identical for all three models. As outliers are not visible in this plot, additional scatter plot analysis is visible in Figure 5. This plot quickly revealed that there were a handful of drastic outliers in all three models. Interestingly, all those outliers were close in age to the maximum possible lifespan of their respective species. In addition, a more general trend of older samples being predicted worse is visible. In fact, samples in the last quartile seem to not fit the previously observed and learned methylation patterns anymore. Those samples are thus also responsible for the significant difference between RMSE and MAE in all our models. Another cause of the poorer performance on older samples is most likely the fact that there aren't many old samples in the entire dataset (see Figure 6a), thus leaving the models little data to learn from.

Investigating the outliers further also reveals that in all the five most erroneous predictions of the EN, the model assumed the samples to be much younger than they are (Figure 6b). This might indicate that the culprit is not to be found in the models but rather in those



(a) **(b)**

Figure 6: a: Probability density function of age distribution in the total dataset, showing a skewed distribution towards younger samples. **b**: The top five outliers of the EN model show a significant age underestimation of our model. All five outliers have true ages close to the maximum possible lifespan. The disparity might be due to an epigenome that is indicative of a much younger sample in our dataset, hence indicating a potential byproduct/cause of their longevity.

samples being enormous biological outliers. Since epigenetic information clearly relates to age, these organisms might just be younger biologically than they are chronologically, hence not conforming to the average methylation profile that is expected to be seen at a particular age. Interestingly, the more regularized nature of the elastic net might be what caused it to be less sensitive to the initial outliers at around 0.8 relative age. The neural network and LGBM instead learned the methylation patterns of the majority samples more closely and hence were subject to decreased performance.

3.2 Feature Importance

Both ElasticNet and LGBM were used in order to assess and identify important features within our dataset. Utilising both models enabled us to validate our results. We explore potentially important CpG sites identified in Figure 2, specifically focusing on those consistent with both our model and the findings presented in the paper[12]. This decision stems from considerations for brevity and to show the shared relevance across both studies. Overlapping sites include: cg12841266, cg10501210 and cg26844246.

4 Discussion

Three main predictive features find support in the literature. The association between the **cg12841266** CpG site within the LHFPL4 gene and aging, as highlighted in the study by Haghani, A. et al.[12], highlights the genes significant role in the central nervous system, particularly in synaptic transmission and plasticity. Research by Wu M et al. [20] further sheds light on the critical functions of LHFPL4 in inhibitory synapse formation and motor behavior, emphasising its importance in maintaining cognitive health and neurological integrity. The presence of significant CpG sites in exon 2 suggests that methylation changes could directly affect LHFPL4s protein products. Such epigenetic modifications are important in understanding the genes contribution to aging, as they can alter gene expression and impact the efficiency and health of neural circuits. This, in turn, could lead to age related declines in cognitive function, synaptic plasticity, and neural integrity.

The CpG site **cg10501210** in the C1orf132 gene (also known as MIR4435-2HG), which has been shown to code for a long non-coding RNA (lncRNA), is likely to be important in the aging process due to its role

in regulating gene expression and its involvement in aging related processes, such as stress response and DNA repair[21] . Furthermore, it has been implicated in increased risks of breast cancer [21], also pointing towards its broader implications in cellular aging mechanisms.

The CpG site **cg26844246**, located within the TLX3 gene, also known as the T-cell Leukemia homeobox 3 or HOX11L2, is a member of the orphan homeobox gene family, which makes up DNA binding nuclear transcription factors. TLX3 plays a critical role in regulating gene expression essential for the development of the nervous system. Additionally, methylation alterations of this gene have been observed in various tumor subtypes, including thyroid cancer [22], bladder cancer, and lung adenocarcinoma[23], [24]. These methylation changes in or near this gene that affect its expression can have significant implications for neuronal health, maintenance, and plasticity [25]. **Deep learning models trade performance for interpretability.** The deep learning model's hyperparameter tuning was constrained by computational resources. Unlike the linear elastic net model, where cross-validation was feasible, model complexity here was much higher, leaving potential impacts on model performance by the train-validation split unexamined. Optimizations such as GPU acceleration, early stoppage once loss reaches a plateau, and refined hyperparameter search methods can mitigate these computational burdens. Additionally, we observe a trade-off between model complexity and comprehensiveness. While deep learning models are intricate and computationally intensive, especially using advanced architectures like Transformers, linear models offer greater simplicity and interpretability, with the ability to directly highlight the importance of specific CpG sites[26]. It might be possible to enhance the accuracy of deep learning models by expanding network complexity; however, the marginal gains in performance must be weighed against increased opacity, resource demand, and potential for overfitting. Future methodologies might merge these two approaches, using linear models for initial predictions, complemented by deep learning ensembles, to balance interpretability with predictive power.

Minimal non-linear relationships observed between relative age and methylation. Deep learning models can capture linear as well as non-linear relationships, but this advantage did not lead to an increase in performance predicting relative age of mammalian species. This may indicate that the deep learning model is simply learning linear relationships in the data, leveraging little to none of its ability to capture non-linearity. Biologically, this may indicate that there are few or no non-linear relationships between methylation sites and relative age of mammals. There is some evidence that non-linear gene-environment relationships may partially explain the development of complex phenotypes, including traits related to age[27]. However, an alternative simple explanation is that models with a lower number of parameters function better on datasets with a high ratio of features to samples[28]. Non-linearity in cellular networks is an area of active research and little has been published specifically exploring the effect of non-linear methylation relationships in observed phenotypes.

Outliers indicative of longevity-associated methylation patterns. Our results show that methylation patterns change so drastically for significantly old individuals above the third quantile in terms of maximum lifespan that they don't seem to follow the same trends anymore that are observable prior to this. While this at first seems like a modeling or sample size error, it seems related to the aging process of specific organisms. The errors our models made with predicting the ages of old samples were not random - indeed, all three models continuously underestimated those samples' ages by a large margin. This might indicate an organism that is biologically younger than what chronological time suggests. In fact, very specific methylation patterns have previously been found in human centenarians, suggesting that individuals who live up to a very high age seem to be different on a cellular level than others[29]. In particular, extraordinarily old individuals seem to display age-related methylation changes only much later in life[30] - thus possibly explaining why our models considered those samples to be much younger than they were[31]. These changes might not just be associative but potentially even causal, as genes that are usually subject to increased hypermethylation with age are necessary for proper transcription regulation[32]. Similar findings have been made by the author of our dataset who reported increased methylation of CpG islands with age and a trend towards hypomethylation for CpG sites outside of such islands[12]. The observable disparity between average and individual cellular state is one of the reasons why, more recently, biological clocks have been gaining traction in the academic and medical community[33]. As opposed to their counterparts that are solely trained on chronological age, these incorporate phenotypic outcomes as well to give individuals an actual idea of their cellular state[34]. In terms of applicability, this is much more useful, as chronological age in humans can simply be measured

by using a calendar. Biological age on the other hand is much more difficult to approximate without assistance of large-scale trained clocks. Nevertheless, outliers as detected by the chronological clocks we applied are a testament to the diversity biology can create. Perhaps, the true discovery even lies in these outliers after all, as they showcase a phenomenon that has kept humans fascinated for millennia: longevity.

5 Conclusion

We observe a relationship between age and methylation profiles across mammalian species. In fact, our models are able to predict chronological age from methylation profiles with high accuracy, especially when ignoring biological outliers. However, sizeable feature numbers, as typical for biological data, lead to models having increased complexity and significant probability of overfitting. Training machine learning models on new data using only CpG sites with high feature importance might lead to improved predictions. Developments in use of semi-supervised learning approaches for biological data provide another avenue for future research.

References

[1] Ralston, A. & Shaw, K. Gene expression regulates cell differentiation. *Nat Educ* **1**, 127–131 (2008).

[2] Moore, L. D., Le, T. & Fan, G. Dna methylation and its basic function. *Neuropsychopharmacology* **38**, 23–38 (2012).

[3] Bird, A. P. Dna methylation and the frequency of cpg in animal dna. *Nucleic acids research* **8**, 1499–1504 (1980).

[4] Ehrlich, M. *et al.* Amount and distribution of 5-methylcytosine in human dna from different types of tissues or cells. *Nucleic acids research* **10**, 2709–2721 (1982).

[5] Bird, A., Taggart, M., Frommer, M., Miller, O. J. & Macleod, D. A fraction of the mouse genome that is derived from islands of nonmethylated, cpg-rich dna. *Cell* **40**, 91–99 (1985).

[6] Saxonov, S., Berg, P. & Brutlag, D. L. A genome-wide analysis of cpg dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences* **103**, 1412–1417 (2006). URL `https://www.pnas.org/doi/abs/10.1073/pnas.0510310103`. `https://www.pnas.org/doi/pdf/10.1073/pnas.0510310103`.

[7] FRAGA, M. F., AGRELO, R. & ESTELLER, M. Cross-talk between aging and cancer. *Annals of the New York Academy of Sciences* **1100**, 60–74 (2007). URL `https://nyaspubs-onlinelibrary-wiley-com.vu-nl.idm.oclc.org/doi/abs/10.1196/annals.1395.005`. `https://nyaspubs-onlinelibrary-wiley-com.vu-nl.idm.oclc.org/doi/pdf/10.1196/annals.1395.005`.

[8] Kabacik, S. *et al.* The relationship between epigenetic age and the hallmarks of aging in human cells. *Nature aging* **2**, 484–493 (2022).

[9] Bocklandt, S. *et al.* Epigenetic predictor of age. *PLOS ONE* **6**, 1–6 (2011). URL `https://doi.org/10.1371/journal.pone.0014821`.

[10] Horvath, S. Dna methylation age of human tissues and cell types. *Genome biology* **14**, 1–20 (2013).

[11] Fahy, G. M. *et al.* Reversal of epigenetic aging and immunosenescent trends in humans. *Aging cell* **18**, e13028 (2019).

[12] Lu, A. T. *et al.* Universal dna methylation age across mammalian tissues. *Nature aging* **3**, 1144–1166 (2023).

[13] Arneson, A. *et al.* A mammalian methylation array for profiling methylation levels at conserved sequences. *Nature communications* **13**, 783 (2022).

[14] Haghani, A. *et al.* Dna methylation networks underlying mammalian traits. *Science* **381**, eabq5693 (2023). URL `https://www.science.org/doi/abs/10.1126/science.abq5693`. `https://www.science.org/doi/pdf/10.1126/science.abq5693`.

[15] Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

[16] Liu, C. *et al.* High-dimensional omics data analysis using a variable screening protocol with prior knowledge integration (ski). *BMC systems biology* **10**, 457–464 (2016).

[17] Wahid, A., Khan, D. M. & Hussain, I. Robust adaptive lasso method for parameter's estimation and variable selection in high-dimensional sparse models. *PLoS one* **12**, e0183518 (2017).

[18] Liu, W. & Li, Q. An efficient elastic net with regression coefficients method for variable selection of spectrum data. *PloS one* **12**, e0171122 (2017).

[19] Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2019).

[20] Wu, M., Tian, H., Liu, X. *et al.* Impairment of inhibitory synapse formation and motor behavior in mice lacking the nl2 binding partner lhfpl4/garlh4. cell rep (2018).

[21] Shafaroudi, A. M. *et al.* Expression and function of c1orf132 long-noncoding rna in breast cancer cell lines and tissues. *International journal of molecular sciences* **22**, 6768 (2021).

[22] Kikuchi, Y., Tsuji, E., Tsuji, S., Kurebayashi, J. & Kaneda, A. Aberrantly methylated genes in human papillary thyroid cancer and their association with braf/ras mutation. *Frontiers in genetics* **4**, 69324 (2013).

[23] Pradhan, M. P., Desai, A. & Palakal, M. J. Systems biology approach to stage-wise characterization of epigenetic genes in lung adenocarcinoma. *BMC systems biology* **7**, 1–21 (2013).

[24] Zhang, Y.-H. *et al.* Distinguishing glioblastoma subtypes by methylation signatures. *Frontiers in genetics* **11**, 604336 (2020).

[25] Namihira, M., Kohyama, J., Abematsu, M. & Nakashima, K. Epigenetic mechanisms regulating fate specification of neural stem cells. *Philosophical Transactions of the Royal Society B: Biological Sciences* **363**, 2099–2109 (2008).

[26] Pudjihartono, N., Fadason, T., Kempa-Liehr, A. W. & O'Sullivan, J. M. A review of feature selection methods for machine learning-based disease risk prediction. *Front Bioinform* **2**, 927312 (2022).

[27] Hunter, M. D., McKee, K. L. & Turkheimer, E. Simulated nonlinear genetic and environmental dynamics of complex traits. *Developmental Psychopathology* **35**, 662–677 (2023). PMCID: PMC9440154.

[28] Bejani, M. M. & Ghatee, M. A systematic review on overfitting control in shallow and deep neural networks. *Artificial Intelligence Review* **54**, 6391–6438 (2021).

[29] Zeng, Q. *et al.* Methylation of the genes rod1, nlrc5, and hkr1 is associated with aging in hainan centenarians. *BMC medical genomics* **11**, 1–10 (2018).

[30] Brooks-Wilson, A. R. Genetics of healthy aging and longevity. *Human genetics* **132**, 1323–1338 (2013).

[31] Daunay, A. *et al.* Centenarians consistently present a younger epigenetic age than their chronological age with four epigenetic clocks based on a small number of cpg sites. *Aging* **14**, 7718–7733 (2022). URL `https://doi.org/10.18632/aging.204316`.

[32] Gentilini, D. *et al.* Role of epigenetics in human aging and longevity: genome-wide dna methylation profile in centenarians and centenarians' offspring. *Age* **35**, 1961–1973 (2013).

[33] Bernabeu, E. *et al.* Refining epigenetic prediction of chronological and biological age. *Genome Medicine* **15**, 12 (2023).

[34] Föhr, T. *et al.* Does the epigenetic clock grimage predict mortality independent of genetic influences: an 18 year follow-up study in older female twin pairs. *Clinical epigenetics* **13**, 1–9 (2021).

MLVU information sheet

Please include this page in your report either at the start or at the end, before the appendix. Do not change the formatting.

Group number
05

Authors

name	student number
Kunal Chaudhary	2814955
Tim Wehnes	2632146
John Collins	2802889
Gabriel Haw	2828156
Matyas Iras	2825895

Software used We used Optuna, scikit-learn, Keras libraries in python for the project.

Use of AI tools No AI tools were used for creation of project report or the models used.

Link to code Some scripts used in data preprocessing are available at `https://github.com/NomadicPython/geo_sample_parser`.

Group disagreements None