Comparing Temporal Change of DNA Methylation Between Type 1 Diabetes Patients and Healthy Individuals

Michael Gee, Raehash Shah, Tejpal Virdi, Rianna Campbell, Max Shushkovsky

Abstract

DNA methylation is an epigenetic mechanism that regulates gene expression. DNA methylation studies have enabled researchers to understand methylation patterns and their regulatory roles in biological processes, aging, and disease. While many comprehensive studies have identified potential methylation sites correlated with disease utilizing a range of discriminative models, little work has been done in utilizing generative models to simulate how these methylation sites change over time in diseased and healthy individuals. In this paper, we first perform feature selection on over 18,000 methylation sites to identify those correlated with aging. We then train three Markov Models to model the change of methylation sites with respect to age in individuals with type 1 diabetes and healthy individuals. We prune and structurally compare Markov Models to identify distinct patterns of methylation in healthy individuals and those with type 1 diabetes. Moreover, we show these Markov Models are generalizable by using them to predict age given unseen methylation sites and benchmark their performance against a linear regression baseline and deep neural network.

1 Introduction

1.1 DNA Methylation and Aging

DNA methylation is an epigenetic mechanism that regulates gene expression by transferring a methyl group onto the C5 position of the cytosine to form 5-methylcytosine. This process occurs at what are known as CpG sites where a cytosine nucleotide is followed by a guanine nucleotide in the linear sequence of bases along its $5' \rightarrow 3'$ direction. *CpG site* and *methylation site* are used interchangeably. DNA methylation studies have enabled researchers to understand methylation patterns and their regulatory roles in biological processes, aging, and disease [23].

1.2 Type 1 Diabetes

Type 1 diabetes is a chronic autoimmune disease characterised by insulin deficiency and resultant hyperglycaemia [10]. Although survival and patient health have improved considerably over the past 25 years, there still does not exist a cure. Furthermore, many patients with type 1 diabetes cannot access modern therapies because of the high costs of even basic care [10].

Particularly, patients with type 1 diabetes have been studied to experience a greater variation in DNA methylation compared with healthy individuals [19]. If these epigenetic variations that appear in patients with type 1 diabetes could be understood, robust treatments for the disease could be devised with new epigenetic therapies [4].

Thus, we decide to study type 1 diabetes in preference of other diseases because of its known correlation with epigenetic variability [19] and readily available data with methylation sites comparable with datasets of healthy patients.

1.3 Methylation Data

Methylation data is predominantly obtained via bisulfite sequencing but other methods such as bisulfite microarrays and enrichment-based methods exist [3]. Bisulfite sequenicng is where DNA is treated with a bisulphite chemical. The vast majority of unmethylated Cs appears as Ts among the sequencing reads, whereas methylated Cs are largely protected from bisulphite-induced conversion [3]. Reads are then aligned and a confidence score from the replicates is output. This is why values for methylation in datasets are continuous rather than discrete.

2 Definitions

Differentially methylated regions: Genomic regions with different methylation statuses among multiple samples (tissues, cells, individuals, etc.), and are regarded as possible functional regions involved in gene transcriptional regulation [24].

3 Existing Work

3.1 DNA Methylation and Age Inference

In 2013, Steve Horvath introduced a new frontier to aging and epigentics research in his proposal of the 'epigenetic clock', a multivariate chronological age predictor based on DNA methylation values of 353 individual CpG sites [17]. A key advantage of this model was that it is generalizable to all human cell types and tissues, excluding sperm, in contrast to other methylation clocks of that time, such as that of Hannum et. al [16] which does not generalize beyond single tissue types. In Horvath's approach, they utilize an elastic net regression model to select their CpG sites and make predictions. Many papers have followed proposing similar linear methods for predicting age from a fixed set of methylation sites [16, 34, 8, 22].

With the recent advent of machine learning, deep learning models have also been proposed for predicting age from CpG sites; however, do not yield significant gains in accuracy compared to linear-based models, yielding less than 0.02 difference in R^2 (predicted age vs. actual age) compared to the Horvath clock [20, 13].

3.2 DNA Methylation and Type 1 Diabetes

The wide variety of alterations to DNA methylation due to disease is widely studied [29], particularly cancer, autoimmune diseases, metabolic disorders, and neurological disorders [18].

Particularly, patients with type 1 diabetes have been studied to experience a greater variation in DNA methylation compared with healthy individuals [19]. Studies have attempted to identify disease-related methylation sites by comparing CpG sites with healthy individuals [25].

3.3 DNA Methylation and Markov Models

In epigenetics research into DNA methylation, markov models have been proposed for detecting differentially methylated regions [7, 12]. These markov models are used to compare two sets of methylation sites sequentially and simplify markov states into (0,0) both low, (0, 1) one high, (1, 0) one high, (1,1) both high. See Figure 1 in [7]. They are not used for predicting age or modelling how methylation states change over time, instead they are used for pairwise parsing of two sets of ordered methylation sites in order to detect differentially methylated regions.

4 Methods

4.1 Feature Selection, Identifying Relevant Methylation Sites

In this section we shall discuss the steps that contribute to the feature selection pipeline outlined in Figure 1.

4.1.1 ETL GEO Datasets

An ETL (extract, transform, load) process was devised to extract, parse, and clean the data from the Gene Expression Omnibus (GEO). The GEOparse Python package was utilized to read and parse the NCBI's GEO microarray SOFT files.

In order to facilitate the processing of this considerable quantity of data, each with different feature names and formats for age, a lossy matching algorithm was devised to locate the age feature from the SOFT file metadata and convert the existing format (days, months, years) into years.

Healthy patient methylation data was sampled from the following GEO datasets: GSE20067, GSE20236, GSE20242, GSE27097, GSE27317, GSE32149, GSE34257, GSE34869, GSE36064, GSE36642, GSE37008, GSE41169, GSE53128, GSE65638. The diabetic patients' methylation data was sampled from GSE20067.

4.1.2 Normalization

It is common for DNA Methylation data collection to experience batch effects [27]. In order to partially mitigate against these batch effects, we normalize each dataset by z-score, in accordance with standard practice [22].

4.1.3 Pearson Correlation

We next select all CpG sites with a magnitude of Pearson correlation coefficient with age greater than 0.5, which happens to be the top 10. Pearson correlation measures linear correation and is defined in Equation 1 where r is the correlation coefficient, x_i are the values of the x-variable in a sample, $\bar{x}$ are the values of the x-variable, y_i are the values of y-variable in a sample, and $\bar{y}$ is the mean of the values of the y-variable.

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2}} \tag{1}$$

4.1.4 Selected CpG Sites

CpG Site	Pearson Correlation Coefficient	Gene ID	Chromosome Location	References
cg16984944	-0.630	TBC1D23	chr3:100,260,992-100,325,251	[33]
cg24046474	-0.556	LRSAM1	chr9:127,451,486-127,503,501	[28]
cg15361750	-0.537	GPR77	chr19:47,331,614-47,347,329	[11]
cg17274064	-0.529	ERG	chr21:38,367,261-38,661,783	[33]
cg07979752	-0.526	CUL5	chr11:108,008,733-108,107,766	[13, 6]
cg10637955	-0.525	BZW1	chr2:200,810,594-200,827,338	[15]
cg26079320	-0.516	POGK	chr1:166,839,447-166,856,359	[32]
cg13150977	-0.510	UBE2QP1	chr15:84,526,781-84,571,216	[5, 21, 30]
cg26149738	-0.510	KCTD15	chr19:33,795,540-33,815,763	[1]
cg17142470	0.501	SORBS3	chr8:22,544,986-22,575,788	[31]

Table 1: **Selected CpG sites**. Chromosome location is referred to the Human genome reference GRCh38 version.

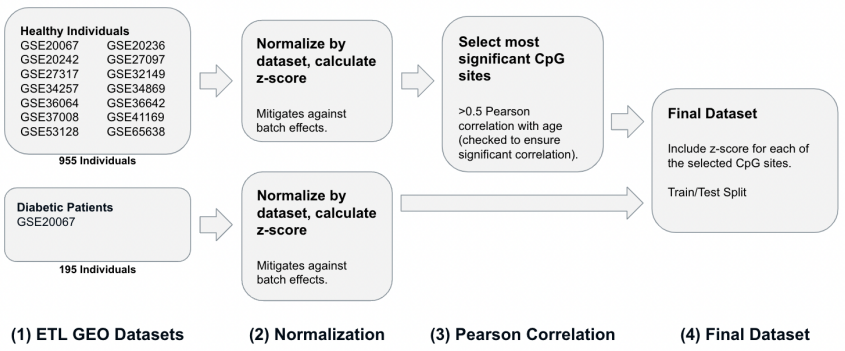


Figure 1: **Feature Selection Pipeline**.

4.1.5 Train-Test Split

The final dataset consisted of 195 unique diabetes patients and 955 unique healthy individuals. Experiments were all conducted with the same train-test split. The test holdout dataset for both healthy and individual patients consisted of a random sample of 10% of the original dataset, with an age distribution similar to the train dataset, selected with the sklearn.model_selection.train_test_split function from the Python Sci-kit Learn library [26].

4.2 Construction of Linear Regression Model

A linear regression model is a simple machine learning model for predictive analysis. It uses a linear function to predict a dependent variable based on independent variables [2]. In this case the dependent variable is patient age and the independent variables are the methylation site values.

We include linear regression models as predictors of age in healthy and diabetic patients as a baseline for comparison against our Markov Models and Neural Network. This approach is simply meant to be a benchmarking metric and is not intended to be the most accurate model for our purposes.

We used the methods from the "Epigenetic Clock" project resource [9] on GitHub to recreate and run 2 linear regression models on our methylation data for healthy and diabetic patients. Both models were built using the Python Sci-kit Learn library linear_model feature [26]. The first model was trained on the healthy patient training data (Healthy linear regression) and the second model was trained on the diabetic patient training data (Diabetic linear regression).

For the purpose of comparing the results of the linear regression age prediction to the Markov Model and Neural Network, we tested the healthy linear regression model on healthy patient test data and tested the diabetic linear regression model on diabetic patient test data. This reveals how accurate each of the models is in predicting the age of the patients. In order to compare these models with the Markov Model and the Neural Network, we collected the R-squared values and the mean absolute errors.

In addition, we tested the healthy linear regression model on diabetic patient data to observe differences in how the healthy model predicts diabetic patients' ages based on their methylation site values. Since diabetic individuals tend to have greater variation in DNA methylation compared to healthy individuals, the predictions of this model will give insight into how DNA methylation progresses and relates between these two groups of people.

4.3 Markov Model for Age Prediction

Another type of model that can provide valuable insight about the inter-relationship between methylation site expression and chronological age is a Markov Model. However, there are some fundamental assumptions that we must acknowledge. First, modeling this data as a Markov Model implies that this relationship is a stochastic process. This means that change to the methylation site expression as chronological age progresses has some level of randomness and uncertainty which means it could be described by a random probability distribution. In addition, use of a Markov Model means that it must also follow the Markov property: the future state is only dependent on the current state

(nothing before the current state). Although these assumptions simplify the actual process that occur in patients, we can still make important conclusions using this type of model.

4.3.1 Construction of Markov Model

In constructing our Markov Model we needed to scale our data to represent methylation site values for different age groups and transition probability values for methylation site values to change as a "patient gets older". Therefore we constructed 2 different matrices for each data set: a state matrix and a transition matrix for both our healthy patient data and our diabetic patient data. With this we could construct our final model which can be seen in Figure 2.

Construction of State Matrix:
Taking our data set which contains methylation site values and the patient's age when those methylation site values were taken, we came up with a distribution of m age buckets that were determined by the distribution of ages that were found within that matrix. Then by looking at the patients that fit within that bucket we distributed them into n location positions. This was because there was a range of values for each of the k methylation sites and we want to allow for the possibility for a patient to go to different positions in their next step and it is how similar their methylation site values are that determine which state they go to. So once the n location position buckets were constructed, patients were now placed into these age buckets and location position buckets. To determine the value for that node or state, we simply averaged the patients' methylation site information that were found within that contained bucket to compute one value for each of the k methylation sites found. This would then be what we label as our state matrix (dimensions: $k \times n \times m$).

Construction of Transition Matrix:
To determine what the transition matrix for our data is, we would use the assumption we made at the beginning of constructing this model: the Markov Property which is demonstrated in Equation 2 for state values π_i and π_j.

$$\pi_i = p_{ij}\pi_j \tag{2}$$

Using this equation and the constructed state matrix we computed the transitions by enacting this property and computing the transition probability, p_{ij}, for all methylation sites for each of the states that we found. Another important feature that we had to incorporate was that the sum of all of the outgoing probabilities from one of the states must equal 1, so we had to normalize all of these probabilities. This would then all we would need to create our transition matrix (dimensions: $k \times n \times m - 1$).

4.3.2 Pruning our Markov Model

Although we had developed a complete Markov Model, we had to consider some additional assumptions that we had inherently incorporated into our model and remove them to potentially provide more insight on our results. This primarily took place when we were choosing how we wanted to determine our age buckets and our location position buckets. Previously, the determination of our age buckets and our location position buckets were simply independent of one another: we would recompute the location position intervals within each age group. The issue with this approach is that, what we are subliminally claiming is that, patients have a probability to go to each of the different location positions in the next age bucket with a non-zero probability. However, this isn't necessarily true since we are treating each of the intervals unequally and skewing it based on our specific patient data which isn't an approach that can provide great insight. So instead we can determine our intervals that is standard for all of our states that we have defined in two different ways:

- A linear approach where each interval is equally spaced and spans our entire data set.
- A quartile based approach where we determine the buckets based on quartiles that fit our entire data set.

We will now have 3 different markov models (1 pre-pruned and 2 post-pruned) which will help us predict age and do downstream analysis with this constructed markov model.

4.3.3 Linear Approach, Pruning Error Transition Matrix

In order to be able to compare the transitions of our markov models, we fix our states. We consider the transition between 6 age ranges: $(0, 5], (5, 10], (10, 15], (15, 20], (20, 25], (25, 30]$. For each age

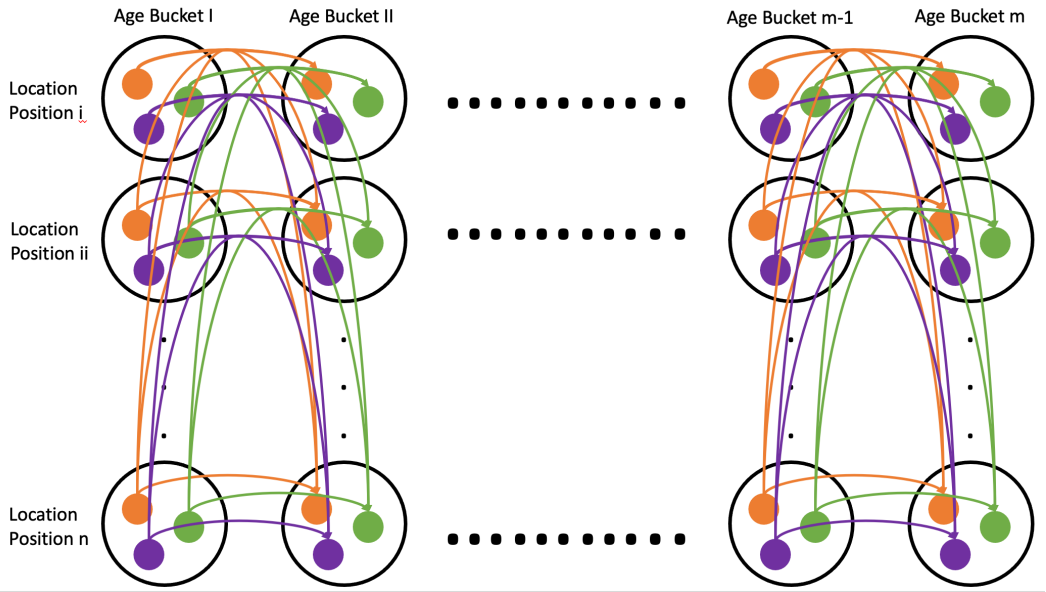


Figure 2: **Skeleton of Markov Model**. This figure demonstrates a framework of the Markov Model that would be generated. Notice in this graph how there are m age buckets that are created and n location positions. In this graph for simplicity, $k = 3$ methylation sites were shown, each labeled in a different color (orange, green and purple). The corresponding colored arrows represent a small set of the possible transitions that can be made from one state (circle in black). Note: that values of transitions and state values were removed for a clearer description of the Markov Model.

range we consider 100 equally spaced methylation ranges from -5 to 5. We do this because our data is normalized to a normal distribution, concentrated around 0. Hence, each markov model shall have 100×6 states. For ease of explanation, we construct a distinct markov model for each CpG site.

Then, considering each transition matrix constructed in this fashion, a **error transition matrix** P_{err} is first calculated in Equation 3 from the transition probability matrix of the healthy and diabetic patients, $P_{healthy}$ and $P_{diabetic}$ respectively. Thus when considering P_{err}, negative values shall indicate a stronger transition in diabetic patients and weak transition in healthy patients. Similarly, positive values indicate a stronger transition in healthy patients and weaker transition in diabetic patients.

$$P_{err} = P_{healthy} - P_{diabetic} \tag{3}$$

In order to decrease the noise of the error transition matrix, we drop all values that have a magnitude less than 0.1 yielding $P_{err,prune}$. We utilize this resultant $P_{err,prune}$ when conducting our analysis in the following section.

4.3.4 Comparing Markov Models using Error Transition Matrix

The pruned error transition matrix $P_{err,prune}$ was then visualized and examined to determine key differences in longitudinal methylation change between healthy and diabetic patients. Quantitatively, we examine the magnitude of transition. A greater magnitude indicates a greater difference in transition probability between healthy and diabetic patients. Qualitatively, we examine which methylation states are preferred between healthy and diabetic patients. Key questions that we shall seek to answer are:

1. Do diabetic patients undergo a different methylation pattern than healthy patients?
2. How does the difference in methylation between healthy and diabetic patients (if any) change as patients get older?

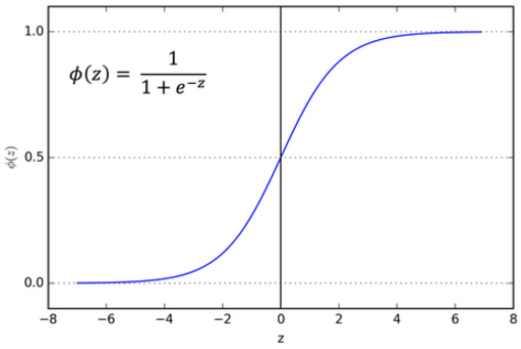


Figure 3: **Sigmoid Activation Function**. The Sigmoid function was used after each layer in order to add non-linearities to the neural network model.

4.3.5 Predicting Age with Markov Model

Once we have constructed our State Matrix and Transition Matrix, we want to predict age of a patient based on their methylation data. Thus, given our state matrix and our transition matrix, we can predict the age of the patient by starting from our smallest age bucket and proceeding through the highest probable transitions to the next state. As we proceed we compute the absolute distance between the methylation data of our patient and the state's methylation values which were computed when constructing the state matrix. The state which generates the minimum absolute distance, would be then the bucket that the patient would be placed into and the mean value of that age bucket would be the predicted age for that patient.

4.3.6 Predicting Condition with Markov Model

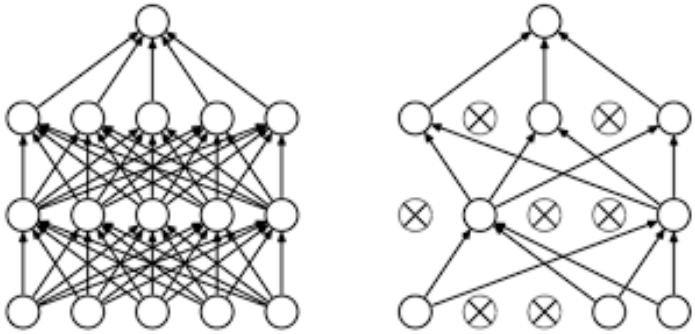
One extension we added to our project was to determine how well our Markov Model could predict an individual patient's condition (diabetic/non-diabetic). Simply from the construction of a Markov Model, we know there is a relationship between a patient's methylation site data and their chronological age. Therefore, suppose we were working in a clinic and had only access to a machine that computes the methylation site data for a patient. Then given a patient's past age and their methylation site data, when that same patient comes into our clinic, we can generate their methylation site data and run it through our Markov Model up until we reach their current age by following the most likely transitions. By simply performing an absolute distance between each of the actual methylation site values and the predicted methylation site using the diabetic/non-diabetic Markov Model, we can use the minimum absolute distance to make a possible prediction of whether the patient may be at risk for diabetes or not.

4.4 Age Prediction with Neural Networks

While regression models have been a popular method for age prediction using methylation patterns, more recent pioneering work has employed deep artificial neural networks. Clearly, age prediction is a complex problem and thus may not necessarily be governed by some linear regression model. Neural networks are particulary powerful because they are able to learn complex patterns across non-linear data in high dimensional spaces. This is achieved through the use of non-linear activation functions across connections in the network. In particular, we use the Sigmoid activation function, as shown in Figure 3.

4.4.1 Construction of Neural Network

Neural networks are composed of an input layer, hidden layers, and an output layer. It has an adaptive learning process (using the stochastic gradient descent and backprop algorithm with a mean squared error loss function) to update weights. Since the goal of a neural network is the minimize loss, it is a common pitfall for an artificial neural network to converge onto a "mean" in order to minimize the mean-squared error. However, this means that the neural network is not truly learning the underlying patterns in the distribution; one can tell if this convergence occurs if the training error rates flatten



(a) Standard Neural Net (b) After applying dropout.

Figure 4: **Dropout Architecture**. Dropout regularization technique by removing connections between neurons across layers.

to be the same, and the test error is significantly higher than the train error due to overfitting to the mean. We combat this through the use of dropout, which is randomly selected to remove a portion of neurons/connections at each layer of the network. This forces the model to not be able to overfit, and consequentially helped our model learn and generalize much easier.

4.5 Benchmarking Performance

In order to compare the ability for each of the models to predict age, we will compute the mean absolute error (MAE) which can be seen in Equation 4. In addition, the predicted age and the actual age were taken as x and y coordinates and plotted on a grid. A line of best-fit was drawn for the coordinates and the R^2 value, which determines how well the values fit a line, was compared across the models. Independently the line of best fit was compared to see how close it was to the line $y = x$ which represents that there was an exact match between the predicted age and the actual age.

$$MAE = \frac{\sum_{i \in patients} |age_{actual} - age_{predicted}|}{||patients||} \tag{4}$$

5 Results

5.1 Linear Regression Results

For each of the models (healthy vs healthy, diabetic vs diabetic, and healthy vs diabetic), we plotted the linear regression results and observed the slope of the best fit line, the R-squared values, and the mean absolute errors.

5.1.1 Healthy and Diabetic Models for Comparison Against the Markov Model and Neural Network

As shown in Figure 5, the blue line is the best fit line for the age predictions of healthy patients on the healthy linear regression model. The red $y = x$ line is for reference, because it represents the line in which the actual ages of the patients exactly match their predictions. The slope of the blue line is 0.626, the R-squared value is 0.825, and the mean absolute error is 8.735. The slope and R-squared values are fairly decent but could be improved to be closer to 1. The mean absolute error is the average difference between actual and predicted ages for each patient, so the difference of 8.375 years also indicates that the model is not ideal.

In Figure 6, the slope of the blue line is 0.041, the R-squared value is 0.171, and the mean absolute error is 5.639. The slope and R-squared values are very poor because they are close to 0 and compared to the values in Figure 5, there is a significant decrease. The mean absolute error, however, improved to 5.639 years compared to Figure 5. A possible reason for the decrease in slope and R-squared values is that the edge predictions near ages 5 and 25 are the most inaccurate, which greatly influences the

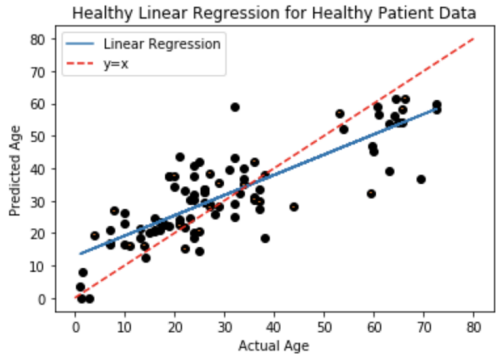


Figure 5: **Healthy Linear Regression Predictions for Healthy Patient Test Data**. The black dots represent a scatter plot of the actual vs predicted ages of the healthy patients. The blue line is the best fit line according to the linear regression model and the red line represents the function $y = x$.

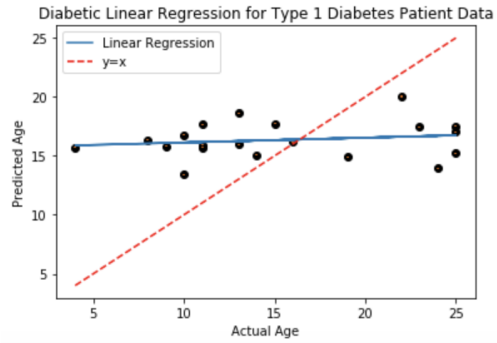


Figure 6: **Diabetic Linear Regression Predictions for Diabetic Patient Test Data**. The blue line is the best fit line for the scatter plot and the red dotted line is $y = x$ for reference.

best fit line. The differences between actual and predicted ages are better towards the center of the graph (around ages 10 to 15), which most likely contributed to the decrease in mean absolute error.

5.1.2 Healthy Model for Predicting Age in Diabetic Patients

In Figure 7, the slope of the best fit line is 0.057, the R-squared value is 0.083, and the mean absolute error is 13.547. The slope and R-squared values are extremely low and it can be observed in the

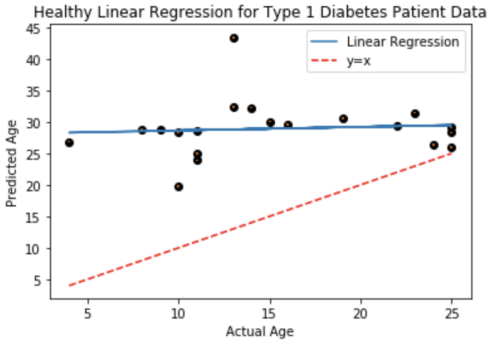


Figure 7: **Healthy Linear Regression Predictions for Diabetic Patient Test Data**. The scatter plot is fitted by the blue line according to the healthy linear regression model and the red $y = x$ line is for reference.

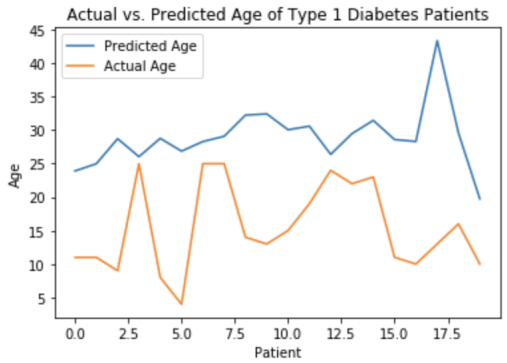


Figure 8: **Healthy Linear Regression Predictions for Diabetic Patient Test Data**. The orange line corresponds to the actual ages of patients and the blue line, which is higher, represents the predicted ages of each patient.

graph that the best fit line is not near (and does not intersect) the $y = x$ ideal line, so this model is highly inaccurate. Also, the mean absolute error is the highest out of the 3 tested models.

In Figure 8, the orange line represents the actual ages of each diabetic patient while the blue line represents the predicted ages of each patient according to the healthy linear regression model. It can be observed that the predicted ages are greater than the actual ages across all patients although the differences between them fluctuate from patient to patient. This could indicate that individuals with Type I Diabetes have methylation site values that correspond to older biological ages when compared to healthy individuals.

5.2 Markov Model Results

In this section, we will investigate our actual structure of our Markov Model and how well our Model was able to predict both age and condition.

5.2.1 Comparing Structure of the Generated Markov Models with Markov Chains

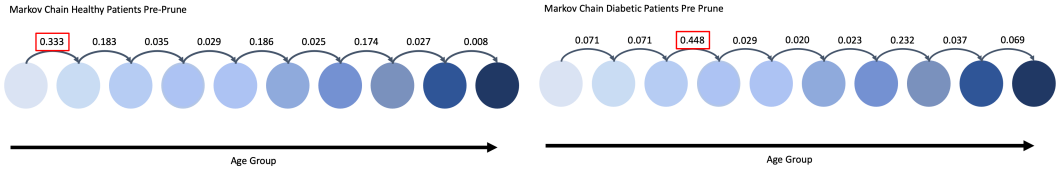
Since the Markov Models were multidimensional arrays, it is difficult to plot all of the models and make conclusions on how these models differ. Instead we aggregated our Markov Models to Markov Chains. Each of the transitions across all methylation sites and all location positions were averaged to compute one transition value from one age bucket to the next age bucket. Each of these markov chains can be seen in Figure 9. In this figure the highest probable transition was also highlighted. Notice that they occur in younger individuals and after pruning they almost occur within the exact same age bucket when comparing the Markov Chains for Diabetic and Healthy patients. This may be because our dataset is more concentrated in that region and thus overall there would be more likely transitions here for any patient that may run through this Markov Model.

5.2.2 Comparing Structure of Markov Models with Error Transition Matrix

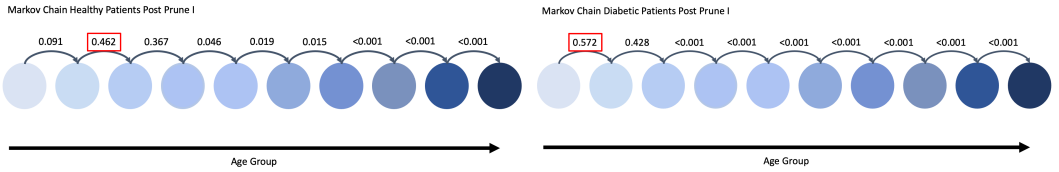
Following pruning of the error transition matrix, we investigate $P_{err,prune}$ which we present in Tables 3 through 12. Quantitatively, all remaining transitions, with the exception of cg26079320, in the pruned error transition matrix were positive. This indicates greater determinism in how methylation changes in healthy patients in comparison to diabetic patients. However, in transitions after the age of 15-20, we observe a greater number of negative transitions. This suggests greater randomness in methylation changes in young diabetic patients and greater deterministic changes as they grow older. Contrary to healthy patients that experience greater determinism in changes to methylation throughout their entire life. Qualitatively, we make distinct observations for each CpG site that are outlined in Table 2. Ultimately, we observe that diabetic individuals do experience different methylation patterns in comparison to healthy individuals, in some cases even changing inversely to healthy patients. The difference in methylation between diabetic and healthy patients remained fairly consistent throughout age. It was often the case that methylation remained fairly constant throughout life in diabetic patients whereas changed over time in healthy patients.

CpG site	Qualitative Observations
cg16984944	We observe, with high probability, a low methylation score in healthy patients across their entire lifespan, with a brief spike in methylation when transitioning to age 5-10. However, diabetic methylation scores remain uniform in early childhood, then spike at age 5-10 to a lesser extend and gradually decrease as individuals age.
cg24046474	Healthy patients appear to exhibit high methylation at early age and gradually decrease methylation with age. Diabetic patients exhibit a similar methylation pattern; however, decrease at a slower rate than healthy patients.
cg15361750	Healthy and diabetic patients appear to experience very similar methylation patterns, showing a decrease in methylation over time.
cg17274064	Diabetic patients exhibit much lower levels of methylation at age 10-20; however, at all other ages experience similar levels of methylation.
cg07979752	Methylation rapidly increase in healthy patients before age 10 and then appears to decrease for the rest of the oberserved lifespan (until age 30). Diabetic patients much lower levels of methylation in comparison to healthy patients after age 10.
cg10637955	Healthy and diabetic patients appear to experience very similar methylation patterns, showing a sudden increase in methylation at age 10-15. However, throughout entire lifespan, diabetic patients experience lower levels of methylation.
cg26079320	Healthy patients appear to experience an initial increase in methylation before age 10 and then a gradual decline in methylation over time. However, diabetic patients experience consistently low levels of methylation throughout their entire lifetime.
cg13150977	Healthy patients appear to experience an initial increase in methylation before age 10 and then a gradual decline in methylation over time. However, diabetic patients experience consistently low levels of methylation throughout their entire lifetime, with a slight increase in methylation at age 25-30.
cg26149738	Healthy and diabetic patients appear to experience very similar methylation patterns with the exception of a sharp increase of methylation in diabetic patients aged 15-20.
cg17142470	Healthy and diabetic patients appear to experience very similar methylation patterns.

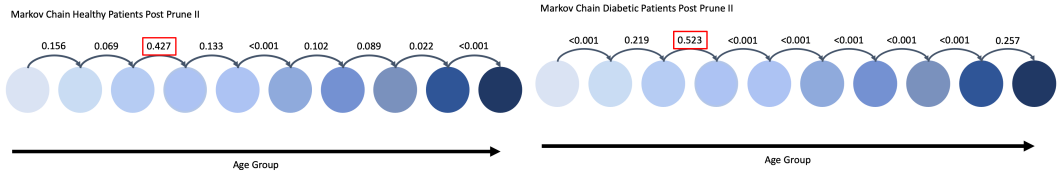
Table 2: **Qualitative Analysis of CpG site trends**. Comparison between healthy and diabetic patients.



(a) Markov Chain Healthy Patients Pre-Prune (b) Markov Chain Diabetic Patients Pre-Prune



(c) Markov Chain Healthy Patients Post-Prune (Linear) (d) Markov Chain Diabetic Patients Post-Prune (Linear)



(e) Markov Chain Healthy Patients Post-Prune (Quartile) (f) Markov Chain Diabetic Patients Post-Prune (Quartile)

Figure 9: **Markov Chains for all Markov Models**. Markov Chains that were generated by averaging the absolute value of the transition probabilities for each of the methylation site at each of the location positions in our Markov Model. Chains (a,c,e) correspond to Healthy Patient Data and Chains (b,d,f) correspond to Diabetic Patient Data. The red squares highlight the transition probability that is the greatest across the entire chain.

cg16984944		
Age Transition	CpG Range	Transition Probability
(0, 5] to (5, 10]	(-1.04,-0.96)	0.152
(0, 5] to (5, 10]	(-0.96,-0.88)	0.163
(0, 5] to (5, 10]	(-0.88,-0.8)	0.184
(0, 5] to (5, 10]	(-0.32,-0.24)	0.501
(5, 10] to (10, 15]	(0.48,0.56)	0.39
(5, 10] to (10, 15]	(1.44,1.52)	0.135
(5, 10] to (10, 15]	(0.32,0.4)	-0.543
(10, 15] to (15, 20]	(-0.4,-0.32)	0.127
(15, 20] to (20, 25]	(-0.08,0.0)	0.552
(15, 20] to (20, 25]	(0.24,0.32)	-1.0
(20, 25] to (25, 30]	(-0.64,-0.56)	0.107
(20, 25] to (25, 30]	(-0.32,-0.24)	0.236
(20, 25] to (25, 30]	(0.0,0.08)	-0.782
(20, 25] to (25, 30]	(0.24,0.32)	-0.218

Table 3: **Pruned transition matrix of cg16984944**. Pruned transition matrix of cg16984944 in pruned markov model (linear).

cg24046474		
Age Transition	CpG Range	Transition Probability
(0, 5] to (5, 10]	(0.88,0.96)	0.311
(0, 5] to (5, 10]	(1.68,1.76)	0.169
(0, 5] to (5, 10]	(1.76,1.84)	0.157
(0, 5] to (5, 10]	(2.24,2.32)	0.125
(5, 10] to (10, 15]	(0.32,0.4)	0.186
(5, 10] to (10, 15]	(0.48,0.56)	0.133
(5, 10] to (10, 15]	(0.64,0.72)	0.101
(5, 10] to (10, 15]	(0.56,0.64)	-0.528
(5, 10] to (10, 15]	(1.04,1.12)	-0.349
(10, 15] to (15, 20]	(-1.2,-1.12)	0.117
(10, 15] to (15, 20]	(-0.88,-0.8)	0.156
(10, 15] to (15, 20]	(-0.64,-0.56)	0.217
(15, 20] to (20, 25]	(-0.16,-0.08)	0.537
(15, 20] to (20, 25]	(0.8,0.88)	-1.0
(20, 25] to (25, 30]	(-0.4,-0.32)	-0.448
(20, 25] to (25, 30]	(-0.32,-0.24)	-0.257

Table 4: **Pruned transition matrix of cg24046474**. Pruned transition matrix of cg24046474 in pruned markov model (linear).

cg15361750		
Age Transition	CpG Range	Transition Probability
(5, 10] to (10, 15]	(0.24,0.32)	0.266
(5, 10] to (10, 15]	(0.72,0.8)	0.108
(5, 10] to (10, 15]	(0.32,0.4)	-0.737
(5, 10] to (10, 15]	(0.96,1.04)	-0.263
(10, 15] to (15, 20]	(0.0,0.08)	0.59
(15, 20] to (20, 25]	(0.0,0.08)	0.367
(15, 20] to (20, 25]	(0.08,0.16)	0.166
(15, 20] to (20, 25]	(0.16,0.24)	0.106
(20, 25] to (25, 30]	(0.08,0.16)	0.302
(20, 25] to (25, 30]	(-0.08,0.0)	-0.913

Table 5: **Pruned transition matrix of cg15361750**. Pruned transition matrix of cg15361750 in pruned markov model (linear).

cg17274064		
Age Transition	CpG Range	Transition Probability
(0, 5] to (5, 10]	(-0.08,0.0)	1.0
(5, 10] to (10, 15]	(0.16,0.24)	0.624
(5, 10] to (10, 15]	(0.4,0.48)	0.223
(10, 15] to (15, 20]	(-0.32,-0.24)	0.118
(10, 15] to (15, 20]	(-0.16,-0.08)	0.244
(10, 15] to (15, 20]	(-0.8,-0.72)	-0.376
(10, 15] to (15, 20]	(-0.56,-0.48)	-0.58
(15, 20] to (20, 25]	(-0.08,0.0)	0.687
(20, 25] to (25, 30]	(-0.24,-0.16)	0.131
(20, 25] to (25, 30]	(-0.08,0.0)	0.403
(20, 25] to (25, 30]	(0.0,0.08)	-0.942

Table 6: **Pruned transition matrix of cg17274064**. Pruned transition matrix of cg17274064 in pruned markov model (linear).

cg07979752		
Age Transition	CpG Range	Transition Probability
(0, 5] to (5, 10]	(-2.8,-2.72)	0.189
(0, 5] to (5, 10]	(-1.52,-1.44)	0.342
(0, 5] to (5, 10]	(-1.12,-1.04)	0.469
(5, 10] to (10, 15]	(1.04,1.12)	0.314
(5, 10] to (10, 15]	(1.28,1.36)	0.253
(5, 10] to (10, 15]	(1.44,1.52)	0.23
(5, 10] to (10, 15]	(3.2,3.28)	0.105
(5, 10] to (10, 15]	(0.08,0.16)	-1.0
(10, 15] to (15, 20]	(0.0,0.08)	0.879
(10, 15] to (15, 20]	(0.16,0.24)	-1.0
(15, 20] to (20, 25]	(-0.32,-0.24)	0.103
(15, 20] to (20, 25]	(-0.24,-0.16)	0.15
(15, 20] to (20, 25]	(-0.88,-0.8)	-1.0
(20, 25] to (25, 30]	(0.08,0.16)	0.447
(20, 25] to (25, 30]	(0.24,0.32)	0.205
(20, 25] to (25, 30]	(-0.88,-0.8)	-0.37
(20, 25] to (25, 30]	(-0.56,-0.48)	-0.63

Table 7: **Pruned transition matrix of cg07979752**. Pruned transition matrix of cg07979752 in pruned markov model (linear).

cg10637955		
Age Transition	CpG Range	Transition Probability
(0, 5] to (5, 10]	(-0.8,-0.72)	-1.0
(5, 10] to (10, 15]	(-0.08,0.0)	0.988
(5, 10] to (10, 15]	(-0.8,-0.72)	-0.154
(5, 10] to (10, 15]	(-0.16,-0.08)	-0.846
(10, 15] to (15, 20]	(0.08,0.16)	0.356
(10, 15] to (15, 20]	(0.4,0.48)	0.117
(10, 15] to (15, 20]	(0.56,0.64)	-0.917
(15, 20] to (20, 25]	(0.0,0.08)	0.335
(15, 20] to (20, 25]	(0.08,0.16)	0.121
(20, 25] to (25, 30]	(-0.16,-0.08)	0.196
(20, 25] to (25, 30]	(-0.08,0.0)	0.417
(20, 25] to (25, 30]	(-0.24,-0.16)	-0.877

Table 8: **Pruned transition matrix of cg10637955**. Pruned transition matrix of cg10637955 in pruned markov model (linear).

cg26079320		
Age Transition	CpG Range	Transition Probability
(0, 5] to (5, 10]	(0.0,0.08)	0.777
(0, 5] to (5, 10]	(0.08,0.16)	0.115
(5, 10] to (10, 15]	(-0.08,0.0)	0.924
(5, 10] to (10, 15]	(-0.4,-0.32)	-1.0
(10, 15] to (15, 20]	(0.96,1.04)	0.685
(10, 15] to (15, 20]	(2.16,2.24)	0.315
(10, 15] to (15, 20]	(-0.88,-0.8)	-0.23
(10, 15] to (15, 20]	(-0.32,-0.24)	-0.77
(15, 20] to (20, 25]	(0.08,0.16)	1.0
(15, 20] to (20, 25]	(-0.32,-0.24)	-1.0
(20, 25] to (25, 30]	(0.08,0.16)	1.0
(20, 25] to (25, 30]	(-0.48,-0.4)	-0.383
(20, 25] to (25, 30]	(-0.32,-0.24)	-0.617

Table 9: **Pruned transition matrix of cg26079320**. Pruned transition matrix of cg26079320 in pruned markov model (linear).

cg13150977		
Age Transition	CpG Range	Transition Probability
(0, 5] to (5, 10]	(-1.52,-1.44)	0.121
(0, 5] to (5, 10]	(-1.2,-1.12)	0.149
(0, 5] to (5, 10]	(-0.4,-0.32)	0.517
(5, 10] to (10, 15]	(0.72,0.8)	1.0
(5, 10] to (10, 15]	(-0.96,-0.88)	-0.486
(5, 10] to (10, 15]	(-0.88,-0.8)	-0.514
(10, 15] to (15, 20]	(0.0,0.08)	0.957
(15, 20] to (20, 25]	(0.08,0.16)	0.947
(15, 20] to (20, 25]	(-0.88,-0.8)	-1.0
(20, 25] to (25, 30]	(0.08,0.16)	1.0
(20, 25] to (25, 30]	(-0.96,-0.88)	-0.209
(20, 25] to (25, 30]	(-0.32,-0.24)	-0.791

Table 10: **Pruned transition matrix of cg13150977**. Pruned transition matrix of cg13150977 in pruned markov model (linear).

5.2.3 Age Prediction with Markov Model

Using each of the Markov Models, we plot the predicted age and the actual age in a graph and fit a line of best fit which can be seen in Figure 10. From these graphs we can see that without pruning, we are already able to predict age quite well. However, when applying this to a linear pruning method, we make a few more generalizations to create equal intervals between them which means that although it is a more accurate model, it doesn't provide great accuracy when predicting age. However, when we look at separating the intervals based on a distribution, we get a much more accurate model that can predict age quite well (low standard error and high R^2 Value).

5.2.4 Condition Prediction with Markov Model

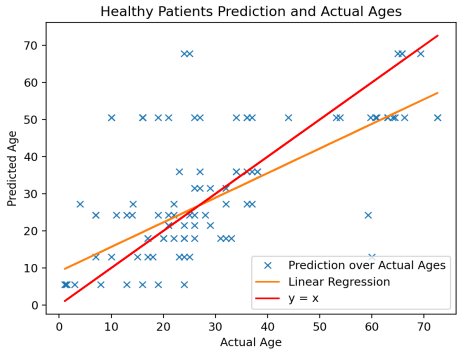
As described previously, an extension of this project is to consider if we had a clinic and whether we can predict if a patient is at risk for diabetes based on their future methylation sites. The ability for our model to predict this condition can be seen in the bar graph in Figure 11. Notice how the best performance overall was done with the pre-pruned model. This is because when we had constructed that model we had made the assumption that at each methylation site everything gets re-normalized in terms of the methylation values. This means that we are treating each state as unique and thus the model was constructed to incorporate this uniqueness and therefore prediction of the condition

cg26149738		
Age Transition	CpG Range	Transition Probability
(0, 5] to (5, 10]	(-0.16,-0.08)	0.17
(0, 5] to (5, 10]	(-0.08,0.0)	0.666
(5, 10] to (10, 15]	(-0.16,-0.08)	0.123
(5, 10] to (10, 15]	(-0.08,0.0)	0.635
(5, 10] to (10, 15]	(0.64,0.72)	-0.561
(5, 10] to (10, 15]	(0.8,0.88)	-0.439
(10, 15] to (15, 20]	(0.0,0.08)	0.726
(10, 15] to (15, 20]	(0.4,0.48)	-1.0
(15, 20] to (20, 25]	(-0.24,-0.16)	0.111
(15, 20] to (20, 25]	(-0.08,0.0)	0.472
(15, 20] to (20, 25]	(0.4,0.48)	-1.0
(20, 25] to (25, 30]	(-0.08,0.0)	0.678
(20, 25] to (25, 30]	(0.0,0.08)	-0.707
(20, 25] to (25, 30]	(0.24,0.32)	-0.188
(20, 25] to (25, 30]	(0.4,0.48)	-0.105

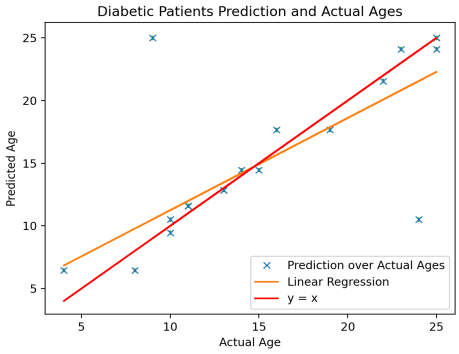
Table 11: **Pruned transition matrix of cg26149738**. Pruned transition matrix of cg26149738 in pruned markov model (linear).

cg17142470		
Age Transition	CpG Range	Transition Probability
(0, 5] to (5, 10]	(0.08,0.16)	0.459
(0, 5] to (5, 10]	(0.16,0.24)	0.239
(0, 5] to (5, 10]	(0.4,0.48)	0.14
(5, 10] to (10, 15]	(0.0,0.08)	0.597
(5, 10] to (10, 15]	(0.08,0.16)	0.164
(5, 10] to (10, 15]	(0.16,0.24)	0.115
(5, 10] to (10, 15]	(-0.16,-0.08)	-0.402
(5, 10] to (10, 15]	(-0.08,0.0)	-0.555
(10, 15] to (15, 20]	(-0.16,-0.08)	0.288
(15, 20] to (20, 25]	(-0.16,-0.08)	0.242
(20, 25] to (25, 30]	(-0.32,-0.24)	0.118
(20, 25] to (25, 30]	(-0.16,-0.08)	0.21
(20, 25] to (25, 30]	(0.0,0.08)	-0.96

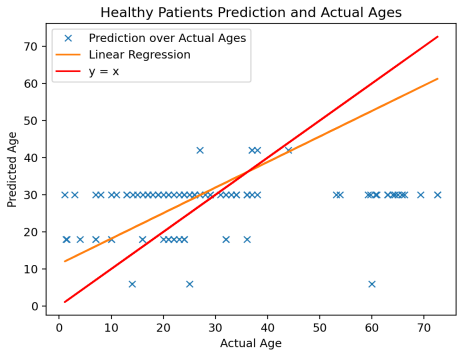
Table 12: **Pruned transition matrix of cg17142470**. Pruned transition matrix of cg17142470 in pruned markov model (linear).



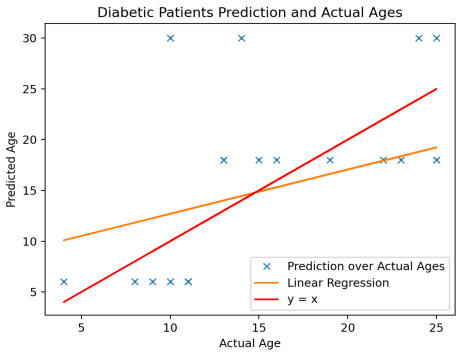
(a) Healthy Patient Pre-Prune Age Prediction
Line of Best Fit: $y = 0.663x + 9.033$
R^2-Value: 0.629
Standard Error: 0.084



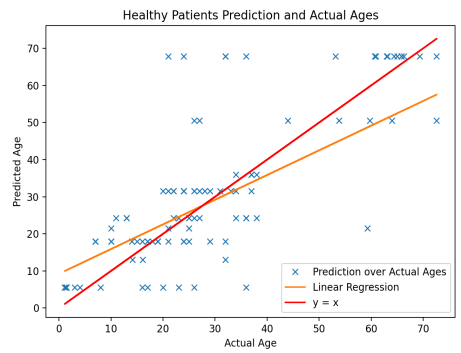
(b) Diabetic Patient Pre-Prune Age Prediction
Line of Best Fit: $y = 0.736x + 3.890$
R^2-Value: 0.716
Standard Error: 0.169



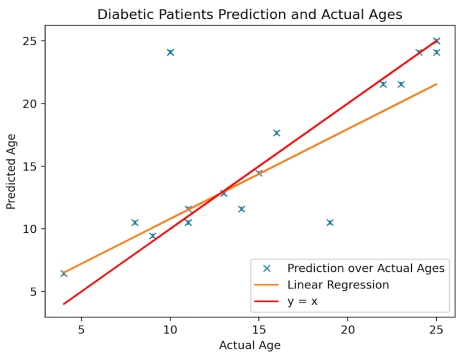
(c) Healthy Patient Post-Prune (Linear) Age Prediction
Line of Best Fit: $y = 0.688x + 11.294$
R^2-Value: 0.262
Standard Error: 0.262



(d) Diabetic Patient Post-Prune (Linear) Age Prediction
Line of Best Fit: $y = 0.436x + 8.336$
R^2-Value: 0.593
Standard Error: 0.139

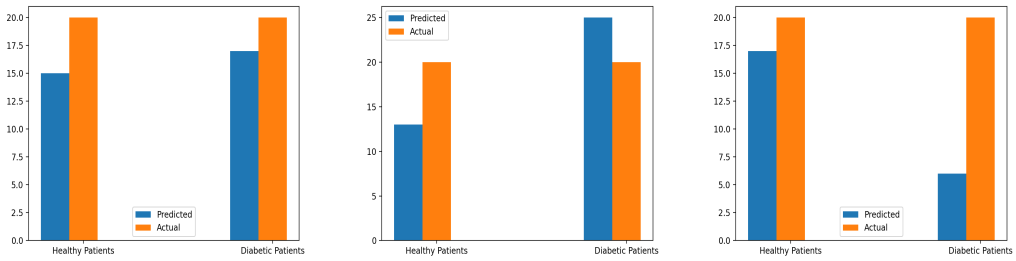


(e) Healthy Patient Post-Prune (Quartile) Age Prediction
Line of Best Fit: $y = 0.665x + 9.242$
R^2-Value: 0.730
Standard Error: 0.064



(f) Diabetic Patient Post-Prune (Quartile) Age Prediction
Line of Best Fit: $y = 0.716x + 3.640$
R^2-Value: 0.708
Standard Error: 0.169

Figure 10: **Linear Fit of Age Prediction**. The ages predicted by the three different models were graphed as blue dots in the graph. The orange line was the line of best fit to those blue dots. We can compare that line to the red line which corresponds to the line $y = x$. We can see that pruning the model in a linear format made the prediction much worse while using a Quartile pruning method improved the age prediction slightly.



(a) Pre-Pruned Model (b) Post-Pruned Model (Linear) (c) Post-Pruned Model (Quartile)

Figure 11: **Bar Graphs of the Predictions of the 3 Markov Models Comparing Healthy vs Diabetic Patients**.

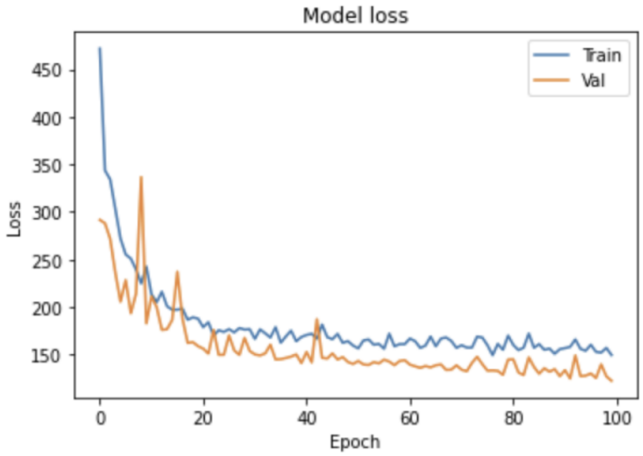


Figure 12: **Healthy Patient Data Training Loss**. Plotted is the training of the neural network on healthy patient data until clear convergence.

can be done quite well. However, once we try to generalize our model a bit more we get that the model doesnt fit quite well, because it seems to not have enough information to classify it either as a healthy patient or a diabetic patient (which is why the total number of predicted values don't add up to the total number of patients tested). The linear model falsely predicts many more diabetic patients while the quartile model does much better on the healthy data set. Either way none of these models perform extremely well in prediction and thus further research on how to improve this model should be conducted.

5.3 Neural Network Results

The neural network was trained for 2,000 epochs on both healthy and diabetic patients. Figures 12 and 13 show the progression of the loss function value across training epochs for the healthy and diabetic datasets, respectively. Convergence is indicated by the flattening of the curves. For the train set, there is clear indication of learning due to the initial volatility and further reduction of that volatility, but for the diabetic set there is potential indication of converging quickly to some mean or other numerical optimal value based on the mean-squared error loss function. The final evaluation call was done on a test set, with a loss function of a mean absolute error, which represents how far on average our model was from predicting the true age. Here are the values for each dataset:

Healthy Model Final Test Mean Absolute Error: 7.67 years

Diabetic Model Final Test Mean Absolute Error: 5.36 years

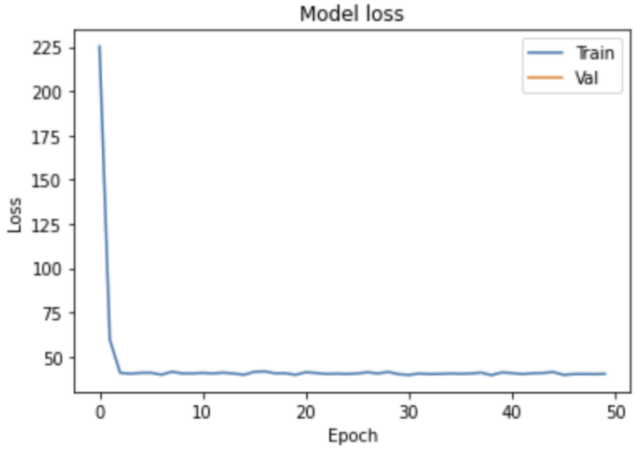


Figure 13: **Diabetic Patient Data Training Loss**. Plotted is the training of the neural network on diabetic patient data until clear convergence.

Model Type	R^2 Healthy	Mean Absolute Error Healthy	R^2 Diabetes	Mean Absolute Error Diabetes
Linear Regression Model	0.825	8.375	0.171	5.639
Markov Model Pre-Pruned	0.629	10.973	0.716	2.155
Markov Model Pruned (Linear)	0.262	13.67	0.593	5.6
Markov Model Pruned (Quartile)	0.730	9.48	0.708	2.575
Neural Network Model	0.644	7.67	0.0753	5.36

Table 13: **Comparing the Results of all of the Models**. For each of the models (linear regression, Markov models, and neural network), the R^2 values and mean absolute errors were collected and compared for both healthy and diabetic patient data.

5.4 Benchmarking Performance

A benchmark of model performance is delineated in Table 13. This table demonstrates an overall improvement between the fit of the pruned Markov model (quartile) versus the pre-pruned Markov model. The R^2 value increased from 0.629 to 0.730 for healthy patients, but the R^2 value slightly decreased for diabetic patients. However, the R^2 value was above 0.7 in both instances, which indicates an above average correlation between the model's work with predicting age. The mean absolute error follows a similar relationship between both models for healthy and diabetic patients; for healthy patients, the error drops from 10.973 to 9.48, but for diabetic patients, the error slightly raises from 2.155 to 2.575. In both healthy and diabetic patients, the pruned model (linear) performs much more poorly than the pruned model (quartile). This indicates that using a quartile method, with the effect of data aggregated more closely in certain areas, is a more appropriate way to predict age, than just splitting the age buckets in a linear fashion. The neural network model performs just slightly better than the pre-pruned Markov for healthy patients, but performed the worst out of all of the models for diabetic patients. Overall, the pruned Markov model based on a quartile system performed the best in terms of R^2 values, and possessed less mean error than other models.

6 Discussion

6.1 Selection of Data

It is important to draw attention to the small sample of data that was accessible for us to perform our study on. Methylation patterns have been shown to be divergent for some sites between demographics [14]. Diabetic patient data was all sampled from Irish patients, whereas healthy data was collected from a range of sources, mostly from the United States and United Kingdom. Moreover, batch effects

may have severely affected results especially as these datasets are from different sources. While most of the methylation data is collected via the Illumina HumanMethylation27 BeadChip, some studies utilized the Illumina HumanMethylation450 BeadChip.

6.2 Comparing Markov Models with Error Transition Matrix

While CpG sites have been used before in methylation clocks and epigenetic studies that are referenced in Figure 1, none have undergone clinical trial or rigorous study to determine if they are, in fact, indicators of aging. Thus observations from our error transition matrix should be interpreted as how changes in CpG sites with high linear correlation with age change in diabetic versus healthy individuals, rather than assuming this comparison to explain differences in aging in diabetic versus healthy individuals.

The relatively small samples size (955 healthy, 195 diabetic) is likely not representative of the population. Hence observations may be the result of sampling bias. Larger-scale studies should be conducted in the future to validate these results.

6.3 Limitations And Significance

After reviewing the results the created Markov model produced, one main limitation was noticed. In using a Markov model, the next state of methylation data was dependent only by the prior state (by the definition of a Markov process). If the next state could be dependent on multiple prior states, a more accurate step through states could be potentially produced, seeing as how methylation values increment and decrement over a lifetime. Additionally, only methylation data was used in prediction of age, rather than other characteristic features (mentioned in Extensions). Despite these limitations, the results from various applications in this paper are significant. First, the R^2 value for predicting a diabetic condition in the pre-pruned model was much greater than for the post-pruned model ($.736$ vs $.436$); the post-pruned model performed just slightly better than the pre-pruned model for healthy predictions. This is significant because a variety of models can be used as a practical application in the future. When given a large data set of methylation values for a patient, medical practitioners can create a database of pre-pruned models for their patients. These models can then be adapted through pruning, which becomes relevant when discussing the results of the models on age prediction. The R^2 value for the pruned model was greater than the R^2 value for the pre-pruned model for healthy patients (in age prediction), which indicated that our editing of the primitive model fit the data better than before. Thus, medical practitioners can also build pruned-models to store for their patients, which allow for both age prediction and diabetic prediction. With this in mind, the results between healthy and diabetic patients introduced an interesting takeaway with regards to referencing human health and age. Commonly, people are referenced by chronological age when associated with certain illnesses. However, doing so may be misleading; chronological age is purely just a number. Considering the effect epigenetic changes have on human health, along with the differences in methylation data produced between healthy and diabetic patients in our work, it may be worthwile to start using biological age, namely methylation data, to refer to people with certain illnesses.

7 Future Work

7.1 Extensions

The nature of this investigation allows for multiple new pathways to explore. For instance, more characteristic features could be used to weight the Markov model that gets produced. For example, consider the addition of gender and ethnicity. These factors may play a role in methylation data. With a comprehensive dataset that contains this information, a new Markov model can be produced that takes into account these characteristic factors; based on the relative importance of each factor, transitions can further be weighted and influenced to become more accurate. Moreover, these characteristic features can be built into the models that predict condition (healthy vs. diabetic). Doing so would allow for more accurate results. Although these models just predicted between two states of health, more states could be implemented, given similar data. Instead of just predicting between a healthy state and a diabetic state, other acquired diseases could be included. This could have powerful implications on how doctors can use quantitative measurements to prepare higher risk patients for the knowledge that they may acquire a certain disease. Additionally, the effect of shorter elapsed time

periods on methylation data could also be studied. Steve Horvath's well regarded paper concluded that "DNA methylation age measures the cumulative effect of epigenetic changes" [14]. Although the effects of methylation can be more strongly measured over longer periods of time, there is potential to look at periods of months to determine if certain methylation sites act as outliers and are affected more rapidly by changes to human health. As a result, these outliers could correspond with how biological age is perceived.

8 Contributions

Mike: Introduction, existing work, sourcing datasets, ETL from GEO, feature selection, consultation on neural network, sourcing longitudinal datasets, clean longitudinal datasets & select relevant patients, linear prune of markov model (only very small modifications to Raehash's code), analysis of linear prune markov model, & accompanying writeups in paper. Organized group meetings.

Raehash: Responsible for constructing the Markov Model from scratch and building the algorithm to predict age and condition from the Markov Model. Worked on all visualizations and investigated all possible ways of explaining the results of the Markov Model. Specifically worked on developing the framework and the Markov chains of the model and documenting all aspects of this model and analysis in the paper.

Rianna: Contributed to the brainstorming of the Markov models. Recreated and ran the linear regression models on the healthy and diabetic datasets. Responsible for all parts of the paper related to linear regressions.

Tejpal: Pitched core idea of age prediction using NN, found related work that paper + dataset was based off, and coded + wrote up the whole neural network + necessary pre-processing of training data.

Max: Worked on designing the original structure of the pre-pruned Markov model, helped with creating visualizations for Markov model structures, wrote parts of paper involved with benchmarking results, discussion, and future work.

9 References

References

[1] Differential dna methylation in experienced meditators after an intensive day of mindfulness-based practice: Implications for immune-related pathways. *Brain, Behavior, and Immunity*, 84:36–44, 2020.

[2] What is linear regression? *Statistics Solutions*, 2021.

[3] N. Ahuja, A. R. S. 1, and S. B. B. 1. Christoph bock. *Nature Reviews Genetics*, Volume 13, 705–719, 2012.

[4] N. Ahuja, A. R. Sharma, and S. B. Baylin. Epigenetic therapeutics: A new weapon in the war against cancer. *Annu Rev Med*, 67:73-89, 2016.

[5] Y. an Chen, S. Choufani, J. C. Ferreira, D. Grafodatskaya, D. T. Butcher, and R. Weksberg. Sequence overlap between autosomal and sex-linked probes on the illumina humanmethylation27 microarray. *Genomics*, 97(4):214–222, 2011.

[6] B. I. T. G. D. A. Center. Correlation between mrna expression and dna methylation. *Broad Institute of MIT and Harvard*, 2013.

[7] Y. Chen, C. K. Kwok, H. Jiang, and X. Fan. Detect differentially methylated regions using non-homogeneous hidden markov model for bisulfite sequencing data. *Methods*, 189:34–43, 2021. Machine learning for the analysis of multi-omics data.

[8] G. W. B. X. Y. S. Y. Y. Z. Z. L. H. X. F. J. Y. . L. F. Cheng Xu, Hongzhu Qu. A novel strategy for forensic age prediction by dna methylation and support vector regression model. *Sci Rep*, 5, 17788, 2015.

[9] A. Darbar. Epigenetic clock. *GitHub*, 2021.

[10] L. A. DiMeglio, C. Evans-Molina, and R. A. Oram. Type 1 diabetes. *Lancet*, 391(10138):2449-2462, 2018.

[11] X. Fang, C. Chen, J. Cai, and E. Xiang. Genome-wide methylation study of whole blood cells dna in men with congenital hypopituitarism disease. *Int J Mol Med*, 43, 1, 2019.

[12] G. B. T. P. C. M. T. G. A. L. Farhad Shokoohi, David A. Stephens. A hidden markov model for identifying differentially methylated sites in bisulfite sequencing data. *Biometrics*, Volume 75, Issue 1, 2018.

[13] K. K. D. S. Fedor Galkin, Polina Mamoshina and A. Zhavoronkov. Deepmage: A methylation aging clock developed with deep learning. *Aging Dis*, Aug; 12(5): 1252–1262, 2021.

[14] H. B. Fraser and L. L. Lam. Population-specificity of human dna methylation. *Genome Biol*, pages 9–13, Feb 2012.

[15] P. Gimenez-Xavier, E. Pros, E. Bonastre, and S. Moran. Genomic and molecular screenings identify different mechanisms for acquired resistance to met inhibitors in lung cancer cells. *Molecular Therapeautics*, 2017.

[16] G. Hannum, J. Guinney, L. Zhao, L. Zhang, G. Hughes, S. Sadda, B. Klotzle, M. Bibikova, J.-B. Fan, Y. Gao, R. Deconde, M. Chen, I. Rajapakse, S. Friend, T. Ideker, and K. Zhang. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Molecular Cell*, 49(2):359–367, 2013.

[17] S. Horvath. Dna methylation age of human tissues and cell types. *Genome Biol*, 14, 3156, 2013.

[18] Z. Jin. Dna methylation and human disease. *Genes Dis*, Mar; 5(1): 1–8, 2018.

[19] R. Johnson. Longitudinal dna methylation differences precede type 1 diabetes. *Sci Rep*, 10, 3721, 2020.

[20] C. L. P. Y. C. L. A. S. . B. C. C. Joshua J. Levy, Alexander J. Titus. Methylnet: an automated and modular deep learning approach for dna methylation analysis. *BMC Bioinformatics*, 21, 108, 2020.

[21] A. M. Kaz, C. J. Wong, S. Dzieciatkowski, Y. Luo, R. E. Schoen, and W. M. Grady. Patterns of DNA methylation in the normal colon vary by anatomical location, gender, and age. *Epigenetics*, 9(4):492–502, Apr 2014.

[22] X. Li, W. Li, and Y. Xu. Human age prediction based on dna methylation using a gradient boosting regressor. *Genes*, 9(9), 2018.

[23] L. D. Moore, T. Le, and G. Fan. Dna methylation and its basic function. *Neuropsychopharmacol*, 38, 23–38, 2013.

[24] M. Neidhart. Chapter 1 - dna methylation – introduction. In M. Neidhart, editor, *DNA Methylation and Complex Human Disease*, pages 1–8. Academic Press, Oxford, 2016.

[25] D. Paul. Increased dna methylation variability in type 1 diabetes across three immune effector cell types. *Nature Communications*, 7, 13555, 2016.

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[27] F. Perrier, A. Novoloaca, and S. Ambatipudi. Identifying and correcting epigenetics measurements for systematic sources of variation. *Clin Epigenet*, 10, 38, 2018.

[28] R. A. Philibert, B. Penaluna, T. White, S. Shires, T. Gunter, J. Liesveld, C. Erwin, N. Hollenbeck, and T. Osborn. A pilot examination of the genome-wide dna methylation signatures of subjects entering and exiting short-term alcohol dependence treatment programs. *Epigenetics*, 9(9):1212–1219, 2014. PMID: 25147915.

[29] K. Robertson. Dna methylation and human disease. *Nat Rev Genet*, 6, 597–610, 2005.

[30] L. J. Smyth, S. M. Cruise, J. Tang, I. Young, B. McGuinness, F. Kee, and A. J. McKnight. An investigation into DNA methylation patterns associated with risk preference in older individuals. *Epigenetics*, pages 1–14, Oct 2021.

[31] R. Sriraksa, C. Zeller, W. Dai, A. Siddiq, A. J. Walley, T. Limpaiboon, and R. Brown. Aberrant DNA methylation at genes associated with a stem cell-like phenotype in cholangiocarcinoma tumors. *Cancer Prev Res (Phila)*, 6(12):1348–1355, Dec 2013.

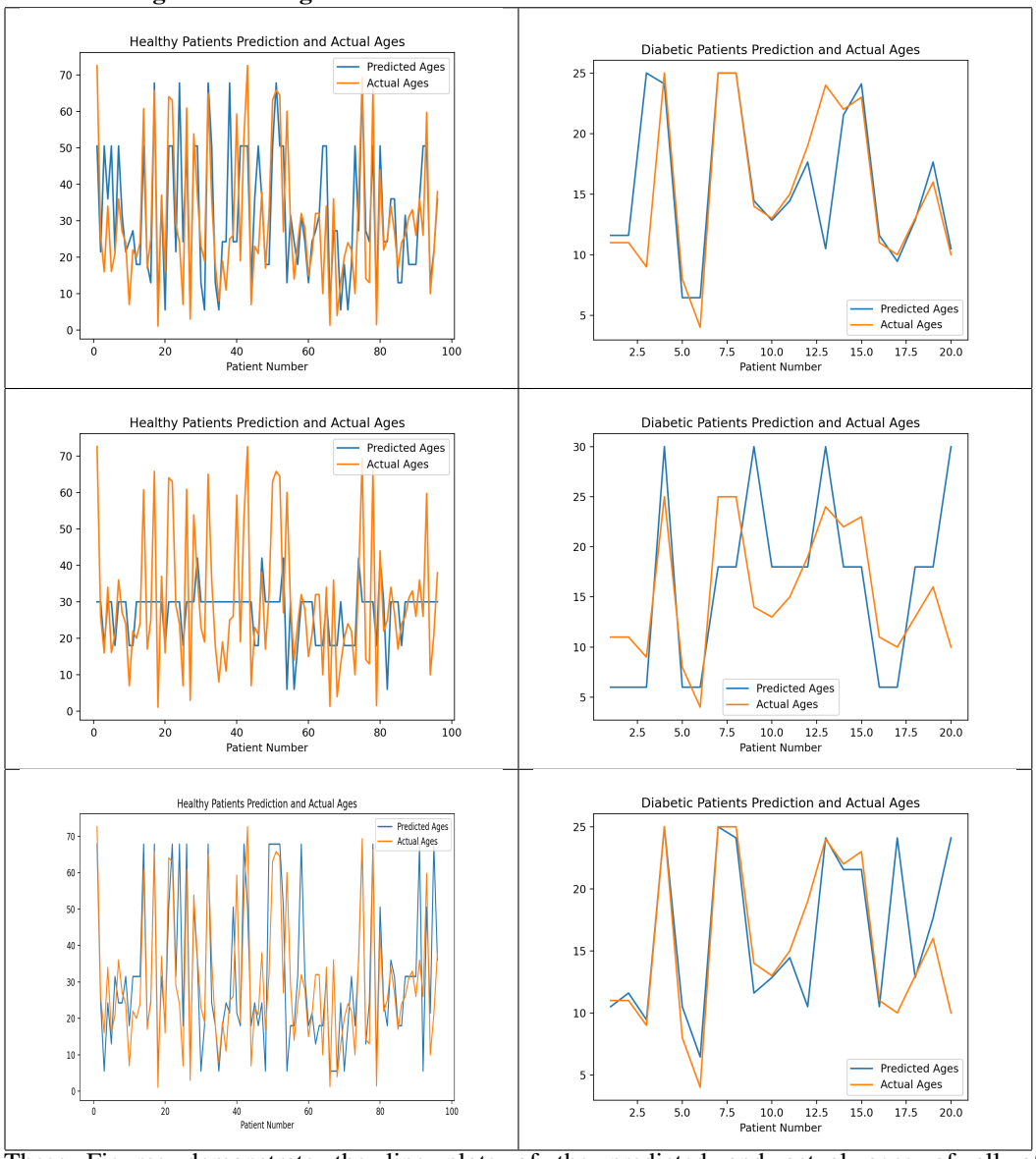
[32] Y. YANG. The role of dna methylation in white matter hyperintensity burden: An integrative approach. *UT School of Public Health Dissertations*, 114, 2020.

[33] A. S. Zannas, J. Arloth, and T. Carrillo-Roa. Lifetime stress accelerates epigenetic aging in an urban, african american cohort: relevance of glucocorticoid signaling. *Genome Biol*, 16, 266, 2015.

[34] J. Zhang, H. Fu, and Y. Xu. Age prediction of human based on dna methylation by blood tissues. *Genes*, 12(6), 2021.

A Appendix

Additional Figures from Age Prediction



These Figures demonstrate the line plots of the predicted and actual ages of all of the patients. The orange line is the actual ages of the patients while the blue is the predicted ages of the patients. The closer together the orange and blue line are, the better the prediction of age that takes place. These graphs additionally contribuhttps://www.overleaf.com/project/6192fc7cahttps://www.overleaf.com/project/6192fc7ca207970b5e375f58207970b5e375f5 to the results that we were able to see in the line of best fit graphs.