University of Cape Town

Third Year Applied Mathematics Project

MAM3055Z

Using Neural networks to predict chronological age using DNA methylation data

Author
Qiulin Li

Supervisor
Associate Professor Jonathan Shock

Declaration of Authorship

I Qiulin Li hereby declare that this report is my original work. (except where acknowledgements indicate otherwise). I authorise the University to reproduce this for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

李秋琳

Signature:

Date: 21-10-2022

Acknowledgements

Many of the computations in this report were done on the High Performance Computing Cluster (HPC) at the University of Cape Town. My sincerest thanks go to the engineers who maintain and support the cluster.

I would also like to thank my Supervisor, Associate Professor Jonathan Shock for his time and patience with me throughout this project. Without his invaluable knowledge, this project would not have been possible.

Contents

Abstract

Genomic data has recently been made to be readily available and accessible to the general public. The data often contains complex patterns that drive fundamental processes in organisms such as aging. Understanding the molecular changes associated with aging is essential to the treatment and prevention of age-related diseases. Several epigenetic clocks have been created to quantify the aging process and its' associated molecular changes. Previous studies have shown that DNA methylation is highly correlated to age but is also influenced by environmental factors. Deep learning models are well-suited to finding complex relationships between two variables, especially when the relationship is often contorted by other factors. In this paper, a regressional neural network algorithm is trained on a variation of datasets. The models produced are compared firstly to each other and then to previous epigenetic clocks to evaluate their performance.

Chapter 1

Introduction

Deep learning is a family of machine learning methods that can be used to find complex relationships in datasets. Neural networks are deep learning algorithms that have recently been applied to a variety of fields from finance to chemistry, however one of the more interesting fields deep learning has been applied to is genomics.

Genomics is a particularly good candidate for deep learning algorithms since many topics in this field generate vasts amount of uninterperetable data. Genomic data consists of large amounts of data that have a large number of parameters. Inferring patterns in this data would be extremely difficult without deep learning algorithms.

Aging can occur at a molecular level and is often associated with changes in the DNA of an individual. DNA methylation in particular, has been shown to influence the aging process in individuals. The relationship between the process of aging and epigenetic changes is complex and often distorted by environmental factors. The relationship can therefore, not be found using analytic or numeric methods. Deep learning methods are well-suited to these types of problems, since there are large datasets available on which algorithms can learn from.

This report describes an artificial neural network algorithm that is trained using variations of the dataset GSE55763. The dataset is transformed using methods of feature selection, normalisation and removing selected outliers. The performance of the algorithm on the various datasets are evaluated and compared. The models that perform the with the greates accuracy are then compared to the epigenetic clocks produced in previous literature.

Chapter 2

Literature Review

2.1 What is DNA methylation?

Aging is a natural process of all living organisms. It is associated with both cellular and molecular changes (Xiao, F, Wang, H & Kong, Q, 2019). Molecular changes such as telomere shortening, DNA methylation and transcriptional mismatches (Jones, M et al, 2015) have been known to be heavily correlated with aging. Many of these molecular changes are referred to as epigenetic changes. Epigenetics is defined as modifications to DNA and DNA packaging that do not involve changes to the DNA sequence (Jones, M et al, 2015). A growing number of studies have shown that the aging is an epigenetic process that is highly correlated to DNA methylation. (Horvath, 2013).

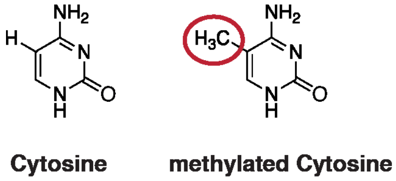


Figure 2.1: Cytosine Methylated vs unmethylated

DNA methylation (DNAm) is the process in which a methyl group (CH_3) is transferred to the 5th position of a cytosine nucleotide (Levy, J, et al, 2020) as shown in Figure 2.1. DNA methylation silences gene expression (Florath, I, et al,2013) by preventing transcription, therefore influencing gene expression without fundamentally modifying genes. DNAm occurs primarily at base pair Cytosine-phosphate-Guanine (Xiao, F, Wang, H & Kong, Q, 2019) known as CpG sites. Genome mapping shows that CpG sites can occur in 'Open-sea' areas where the base pair is not dense or in CpG islands (Florath, I, et al,2013). CpG sites in different locations behave differently due to their function in their respective locations.

DNA methylation data is often gathered by using Illumina Infinium HumanMethylation450 (450K) Bead-Chip array.(Wang, Z, Wu,X Wang, Y, 2018) The technology consists of two probes, one that reads methylated sites and the other that reads unmethylated sites. The technology targets 96 % of CpG islands. The results are commonly reported as beta values (β) between 0 and 1 (Levy, J, et al, 2020). The Beta value is the average methylation at any given site. The value takes into accounts the DNA methylation of all cells forming a body fluid sample (Vidaki, A, et al, 2017). These values follow a beta distribution with variables (Y, α) which can be seen in Figure 2.2.Beta distribution curves can vary greatly in shape based on varying Y and α values.

$$\beta = \frac{y_{methy}}{y_{methy} + y_{unmethy} + \alpha} \tag{2.1}$$

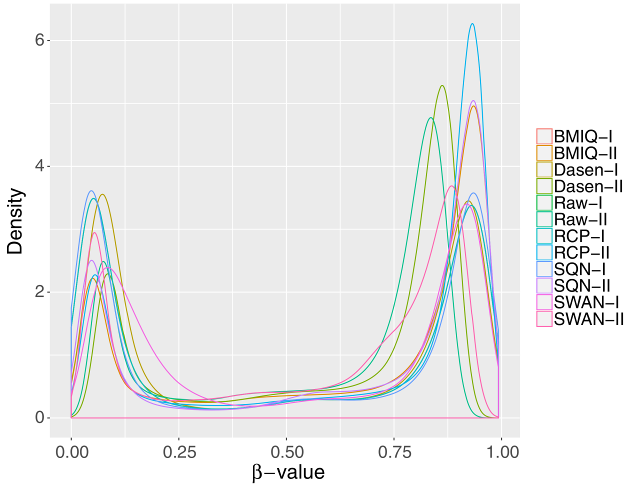


Figure 2.2: Density curves of -values of Infinium I/II probes (Du,P et al, 2010)

Equation 2.1 shows the function used to calculate the Beta values, notice that it is in the same form as the probability density function of beta-distributed data. y_{meth} is the number of cells in a sample that are methylated at that specific CpG site. $\alpha = 100$ is a constant offset used to regularise values when methylation values are low (Du,P et al, 2010). Beta values are biologically interpreted as percentage methylated (Wang, Z, Wu,X & Wang, Y, 2018), but they are highly heteroscedastic (Du, P et al, 2010), (unequal scatter) which results in weak statistical properties. M-values are another type of data that is commonly used to report DNA methylation data.

M-values are the log2 ratios of unmethylated sites compared to methylated sites. They're calculated as:

$$M_i = \frac{y_{meth} + \alpha}{y_{unmeth} + \alpha} \tag{2.2}$$

The values can range from $(-\infty, \infty)$ with α commonly set to 1. M-values are more homoscedastic, meaning the variability throughout the data is approximately constant. M-values are statistically more robust but less biologically interpretable. Utilising both Beta and M-values is ideal so that biological and statistical properties can be properly utilised. The relationship between Beta and M values can be shown by this function (Du, P et al, 2010), where $\alpha = 0$:

$$\beta = \frac{2^{M_i}}{2^{M_i} + 1} \tag{2.3}$$

Beta values are compared and normalised to prevent experimental biases (Vidaki, A, et al 2017), however cell composition is often not taken into account when done so. Higher frequencies of white-blood cells in a sample can increase methylation values at specific sites (Jones, M et al, 2015). The variation in cell composition is one source of variation in DNAm between individuals, however changes in DNA methylation are primarily governed by two phenomena, the epigenetic clock and epigenetic drift (Jones, M et al, 2015).

Epigenetic Drift

Epigenetics is also known as the intersection between environment and one's genome (Jones, M et al, 2015). Many environmental factors such as smoking, UV light and diet can influence the aging procress (Hannum, G, et al, 2013). Exposure to certain factors can accelerate the aging process by increasing DNAm. The impact of these environmental factors on DNAm is reflected in the phenomena known as Epigenetic Drift. Epigenetic drift is defined as the gradual loss and variability in DNA methylation within an individual due to inaccuracies during cell division. (Jones, M et al, 2015) Mismatches during cell division often lead to hypomethylation seen in open-sea CpG sites. (Jones, M et al, 2015) and accumulates to show global hypomethylation over time (Xiao, F, Wang, H & Kong, Q, 2019). Epigenetic drift encapsulates all changes in DNAm that are not directly related to age. Epigenetic drift can increase as a function of time (Vidaki, A, et al 2017), due to increased exposure to various stressors. Epigenetic Drift, although considered a separate factor to the Epigenetic clock, is still governed by the aging process , in addition to environmental and stochastic ones. (Jones, M et al, 2015)

Epigenetic Clock

Previous case studies and experiments have shown that DNA methylation is highly correlated to chronological age (Xiao, F, Wang, H & Kong, Q, 2019). However, it is important to distinguish chronological age and biological age. Chronological age is easily defined as the time passed since the birth of an individual, however the definition for biological age tends to be ill-defined. Biological age is an umbrella term for several age-related phenotypes such as telomere length, disease processes(6i) as well as DNA methylation itself. The ambiguous relationship between chronological age and biological age poses a problem as many age-related diseases, such as cancer (Xiao, F, Wang, H & Kong, Q, 2019), are correlated to biological age and not chronological. The use of epigenetic clocks has been seen as a potential indicator of biological age but fails due to the ambiguity of the term (Hannum, G, et al, 2013) According to (Florath, I, et al,2013), of the 155 CpG sites that are age-associated, 77% of them became hypermethylated with age. CpG sites that begin with $(\beta) < 0.35$ in newborns tend to become hypermethylated over time and sites with $(\beta) > 0.35$ become hypomethylated (Florath, I, et al,2013). Many gene promoters that are active during early life are later switched off due to methylation. This results in physical changes such as slowed elongation (height growth), reduced pigmentation and other characteristics of aging. DNA methylation increases exponentially in early life (Jones, M et al, 2015), which reflects exponential growth and change experienced by individuals during that time (Vidaki, A, et al 2017). DNA Methylation stabilizes in adulthood, where little physical change occurs. CpG islands tend to become hypermethylated with age and hypomethylation in DNA is often attributed to Epigenetic drift and Transcriptional mismatches. (Jones, M et al, 2015) Previous models using linear regression have been made to understand how DNAm relates to age by Horvath first, and then Hannum.

2.2 What is a Neural Network?

Neural networks are a type of deep learning structure used to solve a multitude of problems (Goodfellow, I, Bengio, Y & Courville, A, 2016). To understand what neural networks are, it is important to realise where they fall in the area of artificial intelligence.

Machine learning is a subfield of artificial intelligence that utilises learning algorithms to train using data (Goodfellow, I, Bengio, Y & Courville, A, 2016).Machine Learning can be classified as either supervised or unsupervised.

Supervised learning is where machines are fed both input data and the desired output, the algorithm must then find the relationship between the two. (Svozil, D, Kvasnicka, V & Pospichal, J, 1997)

Unsupervised learning means unlabelled data is fed into an algorithm and an output is produced based on patterns in the dataset (Svozil, D, Kvasnicka, V & Pospichal, J, 1997).

Machine learning may be used to solve many types of tasks, however the two most common are regression and classification.

Classification tasks require the machine to place input data into various categories. This type of task appears as object recognition problems. (Goodfellow, I, Bengio, Y & Courville, A, 2016)

Regression tasks require the machine to produce a numerical output given numerical input data.(Goodfellow, I, Bengio, Y & Courville, A, 2016) Regression algorithms often appear as mappings $f : \mathbb{R}^n \to \mathbb{R}$

In the context of this paper, we aim to utilise labelled DNAm data to predict an individual's age. The task is thus regressional in nature.

Deep learning is a subset of Machine learning that utilises multiple processing layers to understand abstract data (Setiono, R & Thong, J 2004).Since Deep Learning is a subset of Machine Learning, it can be used to solve different tasks. Deep learning has provided significant results in multiple fields (Galkin, F et al,2022) with rapid development in text and image processing. Neural networks are a type of deep learning structure, since it utilises hidden layers to discriminate between labelled data in order to find a pattern relating input and output variables (Sarker, I, 2021). Neural networks can be both supervised (Multi-Layer Perceptron) and unsupervised (Autoencoders). MLPs are commonly used to relate inputs to labels, such as classifying an animal in an image, whereas Autoencoders are dimensionality reduction algorithms (Sarker, I, 2021). Autoencoders can be used to reduce the latent space required to represent datasets of higher dimensions (Han, K et al, 2018).

Neural Networks are biologically-inspired models based on neurological structures within the brain (Svozil, D, Kvasnicka, V & Pospichal, J, 1997). The brain is a neural net consisting of billions of multi-connected neurons that process and transfer electrictrochemical impulses between one another (Hawkings, S, 2009). Neurons are building blocks of the brain; they react to external stimuli by producing an impulse that is then transferred to multiple neurons through connections known as synapses. Each neuron within the network receives multiple impulses which are combined and sent through as a single impulse to the next set of neurons (Hawkings, S, 2009). After the impulse has passed through the neural net (or brain), a response will be sent (as impulse) to tell effector tissues how to react (Moving muscles etc). One vital reason why humans can learn so efficiently is due to brain plasticity. Plasticity implies that the brain can create and modify connections between neurons (synapses) (Hawkings, S, 2009). Neural networks mimic this structure in efforts to produce machines that can learn effectively using various learning algorithms.

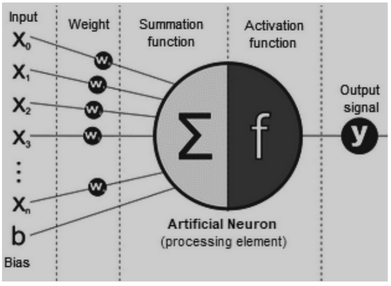


Figure 2.3: Perceptron (Sarker, I, 2021)

Data propagates through the neural network in two directions, forwards and backwards. Forward propagation moves data from the input layer to the hidden layers then to the output layer. Neural networks are made up of a series of artificial neurons (Sarker, I, 2021) known as perceptrons shown in Figure 2.3. Each perceptron is comprised of a summation function and an activation function. The summation (ϵ) is generalized as:

$$\sigma = \sum_{n=1}^{N} W_j X_j \tag{2.4}$$

where j = 1 ,2 ... N and N the number of inputs. The weights (W_j) determine the strength of the connection between an input and a perceptron. Larger weight values indicate that that specific input will have a larger influence on the perceptron. Oftentimes a bias (b_i) value is added to the summation(Panneerselvam, L, 2021). The bias value adds a degree of freedom which allows the Network to perform better. Perceptrons are nonlinear devices (Svozil, D, Kvasnicka, V & Pospichal, J, 1997), owing to non-linear activation functions. It is through nonlinearity that neural networks can recognize complex patterns (Vidaki, A, et al 2017) making activation functions a critical part of perceptrons and consequently neural networks(Dalisay Beaulieu, 2021). Fully connected neural networks such as figure 2.4 have each perceptron take in every single input variable (Hawkings, S, 2009). It passes the weighted summation of these variables into an activation function and returns a single output that travels to the next layer. The activation function defines the output of a perceptron (Hawkings, S, 2009). The ability of perceptrons to take in multiple inputs and produce a single output allows neural networks to be universal approximators (Svozil, D, Kvasnicka, V & Pospichal, J, 1997). It can map any arbitrary vector space to another. Although perceptrons are important building blocks of Neural networks, the most essential part of the network is it's back-propagation algorithm.

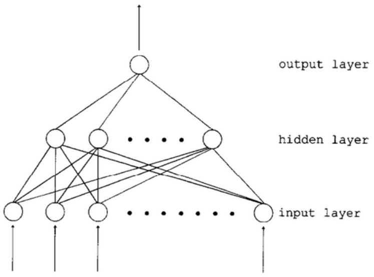


Figure 2.4: Multi-layer feedforward Neural Network

Back-propagation is the method in which Neural networks learn, or update the weights of each input. The loss function computes the difference between the given and expected output (ϵ) (Goodfellow, I, Bengio, Y & Courville, A, 2016). Back-propagation occur after this, ϵ propagated from the output layer, through all the hidden layers to the input layer, updating the weights as the data moves (Hawkings, S, 2009). Backpropagation commonly works through an algorithm known as gradient descent.

All neural networks use forward and back propagation algorithms, however the structure and exact algorithm used within each Network is unique. The neural network we have currently described is a feedforward neural network. We can classify neural networks according to their network architecture. The architecuture of supervised and unsupervised Neural Networks can appear similar but function differently.

Feed-forward neural networks feed data through multiple layers of perceptrons. The data moves through each perceptron in the forward direction only (Panneerselvam, L, 2021). The most common Feed-forward network is a multi-layer perceptron (MLP) Neural Network shown in Figure 2.4. The Network consists of a single layer of input data that is fed into multiple hidden layers of perceptrons. The data from the last hidden layer is sent to a single output layer which returns a single value as an output.

There are a range of neural networks that vary in terms of structure and problem. Each neural network is unique in the type of problem it solves, activation function, as well as structure. A common neural network that can be used to solve regression problems is a Multi-layer feed-forward neural network (also known as an MLP) (Svozil, D, Kvasnicka, V & Pospichal, J, 1997).

2.2.1 Guidelines to construct a Multi-layer perceptron (MLP) neural network

Many parameters and hyperparameters of an MLP must be chosen directly. Since there are a wide variety of algorithms to choose from, there is a lot of flexibility when building neural network. Previous literature has attempted to formulate rules and guidelines when building neural networks.

Input layer The input layer has the same number of perceptrons as there are features in the dataset. Features are independent variables that make up a single sample.

Hidden Layers These two rules are formed on the basis that a single hidden layer with some large number of perceptrons possesses the universal approximator property. More hidden layers will result in increasing computational costs with minimal improvements in the accuracy of the prediction (Svozil, D, Kvasnicka, V & Pospichal, J, 1997).

Size: The number of hidden layers should also be kept to less than two, since more layers can reduce the stability of the backpropogation algorithm (Svozil, D, Kvasnicka, V & Pospichal, J, 1997).

No. perceptrons: The number of perceptrons in each hidden layer should be more than the input layer, but less than the output layer (Svozil, D, Kvasnicka, V & Pospichal, J, 1997).

Activation Function: Common activation function for hidden layers is the ReLu, Tanh and Sigmoid function (Dalisay Beaulieu, 2021). MLP networks commonly use a ReLU activation function. It avoids vanishing gradient, but puts the network at risk for dead neurons (Panneerselvam, L, 2021). The sigmoid activation function is avoided since it often results in gradient saturation, where large outputs are approximated to 1 (Panneerselvam, L, 2021).

Output layer The output layer directly returns the prediction. It takes in several inputs, performs a weighted summation and passes it through an activation function. This layer only consists of a single perceptron (Hawkings, S, 2009).

Activation Function: The activation function depends on the task of the neural network (Dalisay Beaulieu, 2021). Regression tasks usually use an output layer with a linear activation function.

Training algorithm Back-progation through gradient descent is commonly used to train neural networks (Hawkings, S, 2009). The aim of this algorithm is to minimise the error between the predicted and expected output.

Stochastic Gradient Descent This method selects several points at random from the training data and calculates the average gradient from the randomly selected points. The algorithm moves down the gradient to decrease the loss in the training data (Srinivasan, 2019).

Loss function When building a regression model, only the loss functions Mean absolute error (MAE) and Mean square error (MSE) are considered (Hawkings, S, 2009). MSE will magnify errors in the initial predictions, resulting in a model that is training poorly.

2.2.2 Common problems in Neural Networks

Neural networks that do not make accurate predictions usually suffer from one of two problems.

Overfitting is the process where the algorithm memorises the noise in the training data and cannot be generalised to unseen data (Berrar, D, 2019).
Underfitting is the problem where algorithms fail to detect the pattern within the data (Allamy, Haider, 2014)

Combating underfitting or overfitting requires one of two things, increasing the size of the dataset or increasing the model complexity (Dalisay & Beaulieu, 2021).

2.2.3 Regularisation methods

Regularisation methods are modifications made to a deep learning algorithm that allow it to generalise better (Goodfellow, I, Bengio, Y & Courville, A, 2016). It is often used to combat both under and overfitting (Dalisay & Beaulieu, 2021).

K-fold cross validation

K-fold cross validation is a data resampling method that splits the data into different training and validation sets k times (Berrar, D, 2019). The algorithm runs the same number of epochs on each fold. Epochs are the number of times a specific dataset has passed through the neural network, forward and backwards (Hawkings, S, 2009). The subset of data that is used in training the algorithm changes with every fold, so the algorithm learns from datasets with different samples each time. This reduces the bias in the model, producing a model that can generalise better.

Dropout

Dropout is a method in which perceptrons in a layer are randomly switched off with certain probability (Goodfellow, I, Bengio, Y & Courville, A, 2016). This forces other perceptrons to contribute more to the output, improving the performance of the algorithm when all perceptrons are active. Dropout can be implemented in the input and hidden layers but not the output layer (Dalisay & Beaulieu, 2021).

Learning Rate decay

Learning rate decay is used to capture finer details in the patterns between the target and features in a neural network (Dalisay & Beaulieu, 2021). It is implemented in this algorithm to improve the accuracy of the prediction as well as reduce overfitting.

2.2.4 Data preprocessing

Since a model is only as good as the data fed into it, it is important to transform the data before feeding it into training model (Dalisay Beaulieu, 2021). The process in which data is prepared for the algorithm is known as Data preprocessing. This can be done in the form of data cleaning, normalisation etc. Data must be either standardised, normalised or both before it can be used in a deep learning algorithm. The normalisation and standardisation functions (Peck et al., 2020) are given below

$$x"_i = x_i \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{2.5}$$

$$x'_i = x_i \frac{x_i - \bar{x}}{\sigma}$$

(2.6)

x'_i is the standardised value of one sample of a feature and $x"_i$ gives the normalised value. Standardisation can only occur if the original data is a normal distribution.

Feature selection and extraction, is commonly used to reduce the number of inputs. (Pramoditha, 2021). Feature selection methods commonly fall into three categories, filter, wrapper and embedded (Han, K et al, 2018) Wrap methods depend on models for selection criteria, filter methods that filter out certain inputs based on parameters, and embedded methods. Embedded methods are often proposed as the set method due to their efficiency. (Han, K et al, 2018) In this report, a filter, embedded and wrapper method is outlined, namely r^2-values filtering, Principle componenents and Autoencoders

R^2-value filtering

R^2-values, otherwise known as coefficients of determination, are a type of statistical value used to show how well variables are correlated (Peck et al., 2020). It is the percentage variation that the target that can be explained by a feature, given a linear regression model. (Dass, 2015) This can be computed using the following formula.

$$r = \frac{\Sigma(x - \overline{x})^2}{\sqrt{\Sigma(x - \overline{x})\Sigma(y - \overline{y})}} \tag{2.7}$$

We can filter our dataset by choosing features that are highly correlated with the target. We do so by filtering through and selecting features with an $R^2 > p$ where p is some arbitrary value that indicates the signficance level we are filtering through. Low R^2 values indicate that the model (and feature) explains none of the variability in the target data (Dass, 2015), whereas high R^2 values indicate high correlation and that large portions of the variation in the data can be explained by the feature. Filtering the features with high R^2 values allows us to choose the features that have the highest correlation to a given target, allowing us to reduce the number of features and noise with minimal loss of significant data.

Principle component analysis

Principle Component Analysis (PCA) is a method of dimensionality reduction utilised when a dataset has many variables (Brems, M. 2022). PCA also similarly reduces multicollinearity, which is when the input variables are highly correlated. PCA combines the correlated variables into uncorrelated variables due to the orthogonality of Principle Components. We will produce n independent variables by finding different linear combinations of the original variables. We can then order these variables by how well they predict the dependent variable. One of the restrictions of PCA is that it produces inconsistent estimates when applied to heteroskedastic data (Zhang et al, 2021)

Let m be the number of new independent variables and n be the number of original dependent variables. The goal of PCA is to reduce the amount of variables we are working with to m where $m < n$ (Brems, M. 2022). PCA works well as it brings together several important characteristics of the data. The steps of PCA are as follows:

1. Standardize data by rescaling. The mean of the data should be shifted to 0 and the standard deviation 1.
2. A measure of how each variable is associated with one another. (Covariance matrix) Each element in the Covariant matrix is computed using equation 2.8, where N is the number of data values.
$$COV_{x_n,y_n} = \frac{\sigma(x_i - \bar{x})(y_i - \bar{y})}{N-1} \tag{2.8}$$
3. The directions in which our data are dispersed. (Eigenvectors.)
4. The relative importance of these different directions. (Eigenvalues.)

Each Eigenvector corresponds to a Principle component. Each Eigenvector, or Principle component, is a linear combination of all the variables in the data (Jolliffe, I.T & Cadima, J, 2016). In this case it is a weighted combination of the various features. The Eigenvector with the largest Eigenvalue is the first principle component and so forth. The Eigenvalues indicate how much variability in the data each Eigenvector accounts for (Jolliffe, I.T & Cadima, J, 2016). We can thus reduce the dimensionality of the data by dropping eigenvectors with small eigenvalues.

Autoencoders

Autoencoders are a type of unsupervised neural network used to learn latent representations (Sarker, I, 2021). Its' structure can be seen in Figure 2.5. One of the applications of autoencoders is in feature selection via dimensionality reduction.

It is often used to reduce the dimensionality of a dataset. Each variable of the new dataset produced by the autoencoder is made up of linear combinations of the variables from the original dataset (Han, K et al, 2018). The task of an autoencoder is to reconstruct the original input using latent attributes in the data (Han, K et al, 2018). Autoencoders are made up of two parts, the encoder which takes the datasets into a latent space, and a decoder, which recreates the dataset using it's latent representation (Sarker, I, 2021).

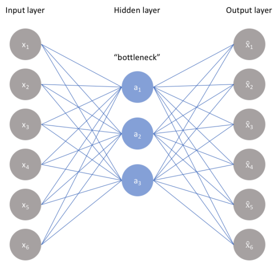


Figure 2.5: Autoencoder structure (Jordan, 2017)

2.3 The works of Horvath, Hannum and those who came before

The mapping of DNA methylation to age is a regression problem since we can define age as a function of multiple variables:

$$Age = f(CpG1, CpG2, ...) \tag{2.9}$$

Machine learning models can approximate this function using either linear or nonlinear functions. The first aging clocks based on epigenetic markers were developed in 2013 (Galkin, F et al,2022) Two of the earliest

DNAm clocks were published by (Horvath, S, 2013) and (Hannum, G, et al, 2013). Although both predict age to a similar degree of accuracy, they barely share any common input data as seen in figure 2.6, indicating that there are a multitude of CpG sites that are correlated with age. Since then multiple epigenetic clock have been developed. Clocks have been developed using DNAm of various tissues with varying degrees of accuracy. Although there are a variety of epigenetic clocks present, the two most prominent clocks are still the Horvath and Hannum clocks.

The epigenetic clock developed by Steve Horvath is a foundational paper that is still highly relevant today. It establishes DNAm as a predictor of age across multiple tissues. A generalised DNAm clock is produced using an elastic net regression model. Multiple tissue samples that are both healthy and cancerous are used to train the model. The clock utilises methylation data from 353 CpG sites as input data and predicts the age of an individual with varying accuracy depending on the tissue sample used. The average errors of it's predictions is 3.6 years. The clock is particularly accurate when predicting the age of children and adolescences. This is likely due to the fact that the effects of epigenetic drift and environmental stressors only appear later in life. Heterogenous tissue such as blood are particularly accurate at predicting age compared to hormonal tissues (breast tissue) which were the least accurate but overall, DNAm samples across multiple healthy tissues were accurate up to 3.6 years when predicting age.

Hannum et al also produced an epigentic clock in 2013 based on DNAm. Hannum also uses Elastic net model but combines it with a bootstrap approach as opposed to the Horvath clock. The model combines Lasso and ridge regression methods with bootstrap analysis. Parameter regularisation occurs using 10-fold crossvalidation. Covariates in the model included BMI, gender, ethnicity and diabetes status however only the significant covariates were tested for correlation after the model had been trained. The model selected 71 CpG sites as input data. All the sites are located near genes with age-related functions such as Alzheimers and tissue degradation. The most notable marker is cg27193080. The clock is trained using data from whole blood samples and tested using DNAm from different tissues. The average error of this clock is 3.9 years which is only 0.3 years less accurate than Horvath. Hannum computed the apparent methylomic aging rate(AMAR), which is the ratio of the predicted age to the chronological age. The AMAR values of each individual is then correlated with their gender and BMI. The study found that men age 4% faster than women. The Hannum clock is initially trained using only whole blood samples since these are the most accessible. The linear model is then adjusted and trained using multi-tissue samples. An important discovery made by both Hannum et al and Horvath is that cancer accelerates the aging process of individuals, therefore seperate aging clocks need to be made for cancer patients.

These two clocks are a form of machine learning but not deep learning. Recently a deep learning approach has been taken to produce DNAm clocks (Levy, J, et al,2020). New clocks have been developed using Neural networks produce more accurate clocks. A recent DNA methylation clock that has been developed using a deep learning is the DeepMage clock (Galkin, F et al,2022)

The DeepMage clock is a multi-layer feedforward neural network with the number of hidden layers ranging from 2 to 5. Many parameters of the neural network, such as the loss function, were varied to gain the most accurate model. Gradient-based feature selection was applied to choose the CpG sites used as input data (Dimensionality reduction). The most accurate model took in 1000 features and passed them through 4 hidden layers each consisting of 512 perceptrons. The set of CpG sites share 121 sites with Horvaths clock and 7 with Hannums as seen in figure 2.6. The activation functions was ELU and dropout at each layer was set to 0.3. The model was trained with fivefold cross-validation. The trained network predicted a healthy individuals age with an average error of 2.24 years, which is at least 1 year more accurate than linear regression clocks. One reason for the accuracy of this clock is the recognition of individuals not written as their exact age(A 23.6 year old is recorded as 23). The neural network predicted the age of unhealthy individuals with an average error of 3.29 years, which is important since previous linear models could not differentiate between healthy and unhealthy

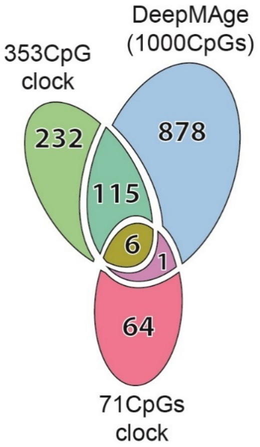


Figure 2.6: Venn diagram of CpG sets of various Epigenetic clocks (Galkin, F et al,2022)

individuals and often predicted unhealthy individuals to be much older.
When using the exact same datasets, Horvaths clock predicted an individuals age with an average error of 3.51 whereas DeepMage predicted it with an error of 2.77 years.

Artificial neural networks have been applied to various problems in multiple fields. Through machine learning, the relationship between DNA methylation has been explored and redefined by many researchers. Although Machine learning models such as Elastic net can prove correlation and are accurate to a certain degree, deep learning structures have proven to be more useful. Neural networks can predict the age of an individual regardless of health to a higher degree of accuracy which once again, shows the power of neural networks to predict and understand complex patterns.

Chapter 3

Dataset analysis and pre-processing

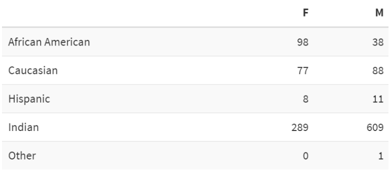
3.1 Meta-analysis

We are using publicly available dataset GSE55763. The original dataset consists of 2,664 participants. After the data has been cleaned, 386 362CpG sites from 1 219 participants are produced. The chronological age of the participants are between 1-75 years old. Some points in the data are considered outliers and will negatively impact our results. The IQR method can thus be used to find and remove outliers (Peck et al., 2020). The method determines outliers as follows with the IQR being 15.2.

$$Outlier > Q_3 + 1.5IQR$$
$$Outlier < Q_1 - 1.5IQR$$

We produce a variation of datasets, one in which all outliers are removed ($17.2 < age < 78.0$), and one in which only the upper outliers are removed (age<78.0) and one where none of the outliers are removed. Younger individuals have CpG sites that are less distorted by environmental variables hence their inclusion in one of the datasets. The dataset uses blood samples to measure DNA Methylation. This is important, as DNA methylation is known to vary between various organ tissues (Horvath), with blood tissue being the most accurate measure of biological age.

The data in Figure 3.1 shows that the demographics are not evenly distributed. The uneven distribution has an unknown influence on DNA methylation, which can potentially affect the results of our algorithm. Hannum (2013) indicated that DNA methylation is influenced by gender, with men typically aging quicker than women. DNA methylation data is not influenced by age alone, but a variety of factors, the goal of our algorithm is to find the relationship between age and DNA methylation whilst removing all noise in the data caused by other factors



	F	M
African American	98	38
Caucasian	77	88
Hispanic	8	11
Indian	289	609
Other	0	1

Figure 3.1: Demographic Data of GSE55763

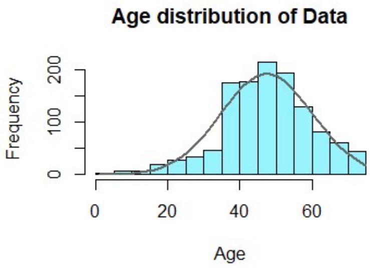


Figure 3.2: Age Distribution of GSE55763

The distribution of the age of all the participants(shown in figure 3.2) is slanted to the right since the probability of younger individuals participating is low. Many of the outliers that were removed consisted of individuals younger than 17. The average age of the participants is 47.41. The most frequent age of the participants is 41, with 17 participants being this age. Nonetheless, the age variable in this dataset has a normal distribution with Shapiro-Wilkes test giving a P-value < 0.05. The data can therefore be standardised and normalised which occurs in data preprocessing (Dalisay & Beaulieu, 2021).

3.2 CpG-site analysis

The Dna methylation data is given as beta values that show the percentage of methylation at each CpG site. Before any statistical analysis can be done, they must be transformed into M–values using a log2 scale by the transformation shown in Equation 3.1

$$M_i = log_2(\frac{Beta}{1 - Beta_i}) \tag{3.1}$$

Since the Beta values are average percentages, the log2 transformation is valid since both Beta and M- values are on the ratio scale(Peck et al., 2020). M-values are preferred since they are homoscedastic. This allows us to utilise more statistical tools such as mean square errors (MSE) and Maximum likelihood estimation (MLE) to analyse the data(42).

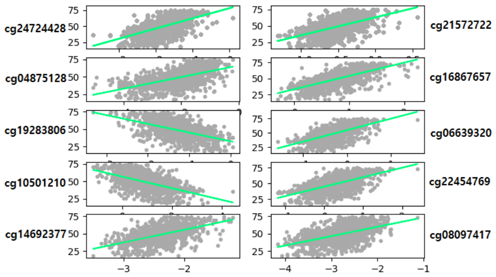


Figure 3.3: Linear regression of the 10 CpG sites with the highest R^2 -values

After transforming the dataset into m-values we took the top 10 cpg sites that had the highest r^2-values. This indicates a significant correlation with age. A linear regression of each cpg site with age is shown in Figure 3.3. The Cpg site with the highest linear correlation is cg16867657 with an R^2-value of 0.69. The CpG site with the greatest variance is cg21979724 with a range of 70.95 between individuals. The CpG site that is the most methylated in the entire dataset is cg21979724.

3.3 Feature Selection

Feature selection is used to reduce the amount of features that are fed into the algorithm (Goodfellow, I, Bengio, Y & Courville, A, 2016). The current dataset has 386 362 features (dimensions) per input, feeding this data into any algorithm as it is will be computationally expensive. Feature selection aims to reduce the computational cost of using the dataset by reducing the dimensions of the data. The reduction in dimensionality should produce datasets that capture the pattern between DNA methylation and age and have minimal noise caused by other variables. Three methods of feature selection are outlines utilised in this project, namely R^2 -value filtering, autoencoders feature selection and Principle Component analysis. Each method will select for a different subset data, although there will be overlapping features, each method will produce a unique dimensionally-reduced dataset. Features from the original dataset are first selected through the methods mentioned above. Outliers are only removed after feature selection so that the selected features represent the variability in the entire dataset, and not just the central data.

3.3.1 R^2-value filtering

A linear regression model is used to relate the CpG site and age of participants. The lowest R^2 value from all the features is 0.001 whilst the highest correlated CpG site had a R^2 value of 0.7.
The CpG sites with the highest R^2 values are chosen for the new dataset. This means that all the features selected be correlated to age by some extent. This is done by imposing a lower bound on the R^2 values of the features that are selected. The initial dataset that consisted of 200 values were selected by choosing all features with $R^2 > 0.4$. Smaller datasets can result in predictions that are inaccurate. The number of features in the dataset can be increased by choosing a lower R^2–value bound. The final dataset selected the top 1 255 sites, all of which have $R^2 > 0.295$.

3.3.2 Principle component analysis

Principle component analysis is used to compute linear combinations of meaningful features from the original dataset. These linear combinations should capture the necessary signals within the dataset so that the algorithm can find the patterns between DNA methylation and age using this Principle Components. The top 1000 Principle components are taken as our new dataset and fed into the algorithm to make predictions.

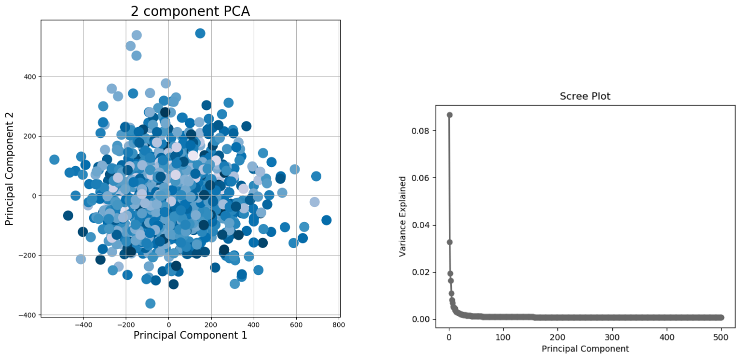


Figure 3.4: Projection of 2 principal components and a Scree plot of 500 principal components

The top two Principal components are projected on two orthogonal axis in figure 3.4. The scattered data points indicate that the data is extremely variable, so many more Principle components will be required capture all the variation within the data. The outliers that are at the bounds of the circle of scatter points will be removed since outliers can influence many algorithms and methods that assume a normal distribution within the dataset.

The scree plot in figure 3.4 indicates that the initial principle compenent explains 2% of the variation in the data. The percentage variation explained by each subsequent principle component decreases exponentially, projecting this forward means more than 1000 principle components is required to explain all the variation in the dataset. 1000 Principle components are chosen for the dataset to balance between the cost of running the algorithm and the accuracy of the prediction made by the algorithm.

3.3.3 Autoencoders

An autoencoder is built using Keras (version), tensorflow (version). We run xxx epochs until overfitting occurs and pull out the hidden layer as our feature-selected dataset. The input and output layers will be made up of 386 362 perceptrons, one for each feature. The bottleneck layer will be made of $\sim$ 1000 perceptrons. The bottlebeck layer is extracted and used as a dataset that can be fed into the Neural Network algorithm that is used to predict age.

The feature selected datasets undergo standardization and normalisation before it can be fed into the algorithm. Normalisation shifts all the values into the range [0, 1] (Dalisay & Beaulieu, 2021). The datasets are then split into test and training sets before they can be used in the algorithm. The split takes 20% of the data as the test set and the other 80% as the training set. The training dataset will consist of 956 samples which are used to both train and validate the data over each epoch, whereas the test data is only used at the end of each loop. A summary of the various datasets that the algorithm is trained on is shown in table 3.1.

	Lower Outliers included	All outliers included	No outliers included
R^2 dataset	1250 features, 1120 inputs	1250 features, 1196 inputs	1250 features, 1129 inputs
PCA dataset	1000 features, 1120 inputs	1000 features, 1196 inputs	1000 features, 1129 inputs

Table 3.1: Dimensions of various datasets

3.4 Computational Environment

The exploratory analysis of the Meta-data of the dataset is done using Stastical programming language R. The rest of the Data analysis is done in Python 3.8. Since the original GSE5573 dataset is considerably large, much of the scripts were run on the High Performance Computing Cluster (HPC) at the university of Cape Town.

Chapter 4

The Neural network architecture

The neural network algorithm is made up of four linear layers (Svozil, D, Kvasnicka, V Pospichal, J, 1997). The outputs of each layer are fed through a ReLu activation function (Panneerselvam, L, 2021), before moving onto the next layer. The input layer will take in 1000-1250 features depending on the dataset used and produces 3244 outputs. The two hidden layers have 4100 and 5210 perceptrons respectively. The output layer transforms 5 210 inputs into a single output which is the prediction generated by the algorithm. The activation function for this neural network is the ReLu function. The ReLu function kills negative outputs and keeps positive ones the same. The function is written as

$$f(x) = max(0, x) \tag{4.1}$$

The difference between the output given by the algorithm and the target output is measured using loss functions.The MAE loss function is used (Dalisay & Beaulieu, 2021), since it won't magnify errors and the output is given is in the same units as the target

$$MAE = \frac{1}{m}\Sigma|x_{target} - x_{prediction}| \tag{4.2}$$

This Neural network is trained using stochastic gradient descent to minimize loss between training and target values. The rate at which it moves down the gradients otherwise known as the learning rate $\in [0.01, 0.025]$

The algorithm underfits the data, which means it is failing to capture the patterns between the CpG sites and age and regularisation methods must be implemented (Goodfellow, I, Bengio, Y & Courville, A, 2016). K-fold cross validation is a common regularization method that is used when the dataset has a limited size (Berrar, D, 2019). 14-fold cross validation is implemented in this neural network to reduce underfitting. Dropout is also used in the algorithm. The dropout rate in the input layer is 0.08 and the dropout in the first hidden layer is 0.05. Dropout is only implemented in these two layers as increasing dropout can cause problems in the algorithm (Dalisay & Beaulieu, 2021). The learning rate in this algorithm will decay at a rate of 0.9 per steps. Steps is a parameter that will vary depending on the number of epochs that is run in each fold. Lower rates of decay are not preferred since they can result in the algorithm training with a learn rate later on.

4.0.1 Regularisation implementation

The Algorithm runs for n epochs on each fold, where n is a parameter that specifies the number of epochs. Overall the algorithm takes 3 hours to run. The neural network is trained on the same dataset with different values for the learning rate, steps, and n- epochs. The results are recorded by Weights and Biases and then averaged. The algorithm is run on the various datasets several times with different parameter values. The parameter values are shown in Table 4.1

Hyperparameter	Value Range
Learning Rate	[0.01, 0.025]
steps	[50, 100]
n epochs	[70,140]
Dropout	0.05
K-folds	14

Table 4.1: Summary of hyperparameter ranges

4.0.2 Model Analysis methods

Model performance is analysed through two metrics. The loss on test data will measure how well the model performs when making predictions from unseen data (Hawkings, S, 2009). R^2 values of the test data are calculated to evaluate the precision of the prediction (Peck et al., 2020). An acceptable threshold in this case will be 0.5.

Chapter 5

Results

The algorithm is trained on multiple datasets and several models that predict an individuals' ages are produced. A summary of the hyperparameters ranges used in the algorithm is shown in table 4.1. The different parameter values produced models that behaved differently in terms of test loss and accuracy. The dataset has a unique structure with underfitting and overfitting problems occurring due to unique splits, outliers and a non-normal distributions at many features. The time taken to run the algorithm once is 1-3 hours.

5.1 The influence of parameters

5.1.1 Learning rate

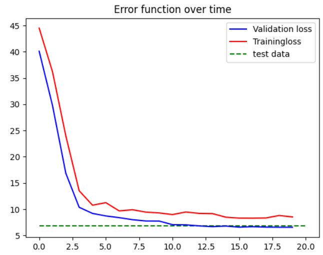
Increasing the learning rate will increase the rate at which the training loss function decreases. Learning rates that are > 0.015 are too large and the algorithm will fail to train correctly on any data. The optimal learning rate that is found heuristically is 0.15. Larger learning rate values need to decay quicker in order not to overfit the data.

5.1.2 Steps and n-epochs

The number of epochs and steps are related since n-epochs $<$ steps will indicate that the two fold will have the same learning rate at some point.If n-epochs $>$ steps, the algorithm will learn less from each subsequent fold. The grey curves in Figure 5.1 and 5.3 show that for both datasets, the difference between the loss values at the halfway point and the end of the run is negligible. The grey curve ran with n-epochs = 100. THe n-epochs can be halved so n-epochs =50 is the optimal number of epochs run. We restrict n-epochs to 50 and choose steps = 60 since this gives the best algorithm performance that does not overfit the data.

Outliers

The error function is shown in figure shows the running loss on the training and validation data when the algorithm trains a model using the R^2 including all outliers.



5.2 R^2 Dataset

The R^2 dataset provides training data that is free from noise due to other co-factors. This allows the algorithm to train better since the pattern signals will become clearer in a simplified dataset. The loss functions on the training and validation dataset of the various models is shown in figure 5.1. Both loss functions decrease exponentially as time progresses. If enough epochs are run, both training and validation loss will become zero, however the test loss value will not decrease. The average test loss over all the models is 8.25.

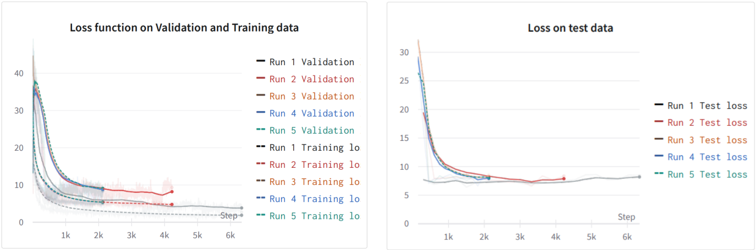


Figure 5.1: Validation and Test loss on R^2 dataset

The R^2 dataset is run several times with variations in hyperparameter value. All the models performed similarly, additional epochs shown in Run 1 indicate that additional epochs will not improve model performance on new data, nor will it cause overfitting.

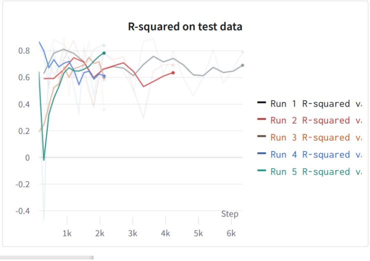


Figure 5.2: Running R^2-value of the algorithm on the R^2 dataset

The initial precision of the predictions vary, however all the R^2 values converge after a certain number of epochs. The average test loss is 0.675 which is greater than the threshold 0f 0.5 therefore the predictions are considered precise.

5.3 PCA Dataset

When the algorithm trains on the Principle Component dataset, the model have an average loss function of 9.317 years. Several runs of the algorithm show that the initial validation loss is high but decreases rapidly within the first two folds as shown in Figure 5.3. Figure 5.3 similarly shows that the loss on the test data will decrease in the beginning but converge onto a certain value. Training over additional epochs, as shown by run 4, does not produce a better or worse predictions, so less epochs can be used which will result in the same loss on the test data.

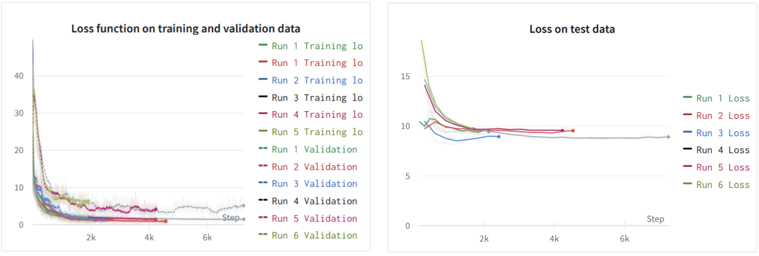


Figure 5.3: Validation and Test loss on PCA dataset

The algorithm did not show any signs of overfitting nor underfitting when training on this data. The average R^2 is 0.2575 which is lower than the threshold 0.5. This means that the predictions are not precise.

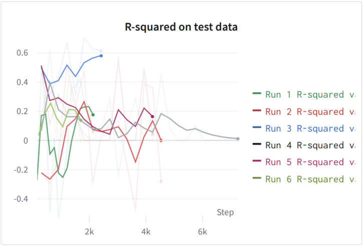


Figure 5.4: Running R^2 -value of algorithm on PCA dataset

Chapter 6

Discussion

The goal of this experiment is to predict the age of individuals based on the degree of DNA methylation at each CpG site using a Regressional neural network. Two models are produced, one that trained on the PCA dataset and the other on the R^2 datasets. Both datasets had no outliers, since outliers caused underfitting. The models are compared to the Horvath, Hannum and DeepMage epigenetic clocks in Table 6.1.

Horvath	Hannum	DeepMage	R^2 Model	PCA Model
3.44	3.09	2.77	8.25	9.317

Table 6.1: Summary of loss functions of various Epigenetic clock

Epigenetic clocks are often trained using either several CpG sites from a small amount of individuals or a few CpG sites from a few individuals. Hannum (2013) uses 450,000 CpG sites from 656 individuals. The Horvath clock uses 8 000 inputs and 21 369 features (Horvath, 2013). The DeepMage clock (Galkin, F et al,2022) uses 4,930 inputs from 17 data sources and 1,000 features. The models produced in this paper only trained on 1000 features from 956 samples. A smaller dataset is used since running deep learning algorithms through large ones is time consuming and computationally expensive.

In a direct comparison to previous epigenetic clock, the models in this report performed poorly, however if the size of the datasets are taken into account, the models in this report are performing well, particularly the model trained on the R^2 dataset.

Although computing times are not given for any of the epigenetic clocks in the literature, based on experimental runs with datasets with less features, it can be deduced that those epigenetic clocks took much longer to train than the models in this report.

6.1 The case for outliers

The datasets containing outliers performed the worst out of all the datasets.The outliers introduced so much noise that the algorithm could not deduce any relationship between the the CpG sites and age when training with data containing outlier.

The datasets containing either no outliers or only the lower outliers did not present with underfitting, however the algorithm converged to the minimum loss value on the test data quicker in the dataset with no outliers. Underfitting is therefore caused by primarily by noisy from the upper outliers.

6.2 Comparing the PCA dataset and R^2 dataset models

The model that performed the best in this report is the one trained on the R^2 -dataset with no outliers. This is to be expected since the R^2 dataset has filtered out the any variables that are poorly related to age, whereas the PCA dataset simply reduces their influence in each Principle component.

One significant difference is that the R^2 dataset produced models that made precise predictions, albeit inaccurate predictions. The predictions made by the model trained on the PCA dataset however were neither precise, nor accurate. In terms of precision and accuracy, the models trained on the R^2 dataset performed better compared to those trained on the PCA dataset.

Feature selection should focus on selecting sites that are correlated with age, dimensionality-reduction methods such as PCA fail to clean noisy data as effectively as filter methods, causing the algorithm to train on noisy data. This resulted in models that were inaccurate and unsure of their predictions.

Chapter 7

Conclusion

This experiment produced two models that can predict the age of an individual, given the degree of DNA methylation at certain CpG sites in their genes. Epigenetic clocks such as this one can be used to study the relationship between DNA methylation and aging (Galkin, F et al,2022). Changes in DNA methylation that result in age predictions older than expected can act as a diagnostic tool to diagnose the early onset of age-related diseases such as alzheimers. (Hannum S, 2013)

Training models that use less inputs and features can produce models that poorly compared to those that trained with more inputs and features. The trade-off is between accuracy and computational costs since more data requires more computational power to process it. In this report, a balance between the two is attempted, as a regressional neural network is trained for shorter periods on less data is to make accurate predictions.

The type of feature selection used can also influence the error and accuracy of models produced. Selecting features with a direct correlation to the target will produce a dataset that trains a better model, compared to dimensionality reduction methods.

7.1 Moving forward

An autoencoder has been constructed and tested on smaller datasets as a tool for feature selection, however due to time constraints, the autoencoder dataset is not produced nor does the algorithm ever trained a model with it. Given more time, feature selection should be explored to evaluate its' impact on the performance of the neural network algorithms.

Hyperparameter tuning in this report did not rely on any particular method, instead, parameters were changed heuristically to understand which parameters gave better results. Hyperparameter tuning via grid search or a hyperparameter sweep should be run to evaluate which parameters give the best models.

Saliency mapping methods have been implemented in the algorithm. However, not enough research and experimentation has been done on the topic in this paper, so it is therefore omitted.

7.2 Closing remarks

The algorithm produced in this paper can be used to produces a neural network model that predicts the age of an individual with a error of 8.25 years in a best case scenario. Previous epigenetic clocks indicate that although it is possible to get a smaller error value, the time taken to train these models is exponentially greater than the time taken in this report. Thus the models produced in this report perform relatively well given the constraints on computational power as well as time.

Bibliography

[1] Allamy, Haider. (2014). METHODS TO AVOID OVER-FITTING AND UNDER-FITTING IN SUPERVISED MACHINE LEARNING (COMPARATIVE STUDY).

[2] Berrar, D. (2019). Cross-Validation. Encyclopedia of Bioinformatics and Computational Biology. 542-545. DOI: 10.1016/b978-0-12-809633-8.20349-x.

[3] Florath, I., Butterbach, K., Muller, H., Bewerunge-Hudler, M. & Brenner, H. (2013). Cross-sectional and longitudinal changes in DNA methylation with age: an epigenome-wide analysis revealing over 60 novel age-associated CpG sites. Human Molecular Genetics. 23(5):1186-1201. DOI: 10.1093/hmg/ddt531.

[4] Galkin, F., Mamoshina, P., Kochetov, K., Sidorenko, D. & Zhavoronkov, A. 2022. DeepMAge: A Methylation Aging Clock Developed with Deep Learning. [2022 , September 04].

[5] Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sadda, S., Klotzle, B. & Bibikova, M. et al. (2013). Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. Molecular Cell. 49(2):359-367. DOI: 10.1016/j.molcel.2012.10.016.

[6] Han, K., Wang, Y., Zhang, C., Li, C. & Xu, C. 2018. Autoencoder Inspired Unsupervised Feature Selection. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). DOI: 10.1109/icassp.2018.8462261.

[7] Haykin, S. 2009. Neural Networks and Learning Machines. 3rd ed. New Jersey: Pearsons.

[8] Horvath, S. 2013. DNA methylation age of human tissues and cell types. Genome Biology. 14(10):R115. DOI: 10.1186/gb-2013-14-10-r115.

[9] Jones, M., Goodman, S. & Kobor, M. (2015). DNAmethylation and healthy human aging. Aging Cell. 14(6):924-932. DOI: 10.1111/acel.12349.

[10] Levy, J., Titus, A., Petersen, C., Chen, Y., Salas, L. & Christensen, B. 2020. MethylNet: an automated and modular deep learning approach for DNA methylation analysis. BMC Bioinformatics. 21(1). DOI: 10.1186/s12859-020-3443-8.

[11] Sarker, I. 2021. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. SN Computer Science. 2(6). DOI: 10.1007/s42979-021-00815-1.

[12] Setiono, R. & Thong, J. (2004). An approach to generate rules from neural networks for regression problems. European Journal of Operational Research. 155(1):239-250. DOI: 10.1016/s0377-2217(02)00792-0.

[13] Svozil, D., Kvasnicka, V. & Pospichal, J. 1997. Introduction to multi-layer feed-forward neural networks. Chemometrics and Intelligent Laboratory Systems. 39(1):43-62. DOI: 10.1016/s0169-7439(97)00061-0.

[14] Vidaki, A., Ballard, D., Aliferi, A., Miller, T., Barron, L. & Syndercombe Court, D. (2017). DNA methylation-based forensic age prediction using artificial neural networks and next generation sequencing. Forensic Science International: Genetics. 28:225-236. DOI: 10.1016/j.fsigen.2017.02.009.

[15] Wang, Z., Wu, X. & Wang, Y. 2018. A framework for analyzing DNA methylation data from Illumina Infinium HumanMethylation450 BeadChip. BMC Bioinformatics. 19(S5). DOI: 10.1186/s12859-018-2096-3.

[16] Wu, Xinxing Cheng, Qiang. (2020). Fractal Autoencoders for Feature Selection.

[17] Xiao, F., Wang, H. & Kong, Q. 2019. Dynamic DNA Methylation During Aging: A “Prophet” of Age-Related Outcomes. Frontiers in Genetics. 10. DOI: 10.3389/fgene.2019.00107.

[18] Goodfellow, I., Bengio, Y. & Courville, A. 2016. Deep learning. 1st ed. Cambridge (EE. UU.): MIT Press.

[19] Panneerselvam, L. (2021) Activation functions: What are activation functions, Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/04/activation-functions-and-their-derivatives-a-quick-complete-guide/ (Accessed: October 21, 2022).

[20] Srinivasan, A.V. (2019) Stochastic gradient descent clearly explained !!, Medium. Towards Data Science. Available at: https://towardsdatascience.com/stochastic-gradient-descent-clearly-explained-53d239905d31 (Accessed: October 21, 2022).

[21] Dalisay, D. Beaulieu, K. 2021. Machine learning mastery. Available: https://machinelearningmastery.com/ [2022, October 21].

[22] Brems, M. 2022. A one-stop shop for Principal Component analysisMedium. Available: https://towardsdatascience.com/principal-component-analysis-pca-explained-visually-with-zero-math-1cbf392b9e7d [2022, October 21].

[23] Jolliffe, I.T. Cadima, J. 2016. Principal component analysis: A review and recent developmentsPhilosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences374(2065):20150202. DOI: 10.1098/rsta.2015.0202.

[24] Dass, G. 2018. Regression analysis: How do I interpret R-squared and assess the goodness-of-fit? Available: https://www.linkedin.com/pulse/regression-analysis-how-do-i-interpret-r-squared-assess-gaurhari-dass/ [2022, October 21].

[25] Peck, R., Short, T. amp; Olsen, C. 2020. Introduction to statistics and data analysis3rd ed. Belmont, California: Cengage.

[26] R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R- project.org/.

[27] G. van Rossum, Python tutorial, Technical Report CS-R9526, Centrum voor Wiskunde en Informatica (CWI), Amsterdam, May 1995.”