University of Glasgow | School of Mathematics & Statistics

FINDING EPIGENETIC SIGNATURES FOR HUMAN AGEING

Determining an Epigenetic Signature for Biological Ageing

Mahima Mago
2748635M
School of Mathematics and Statistics

Table of Contents

Abstract

Biological ageing refers to changes that occur at the cellular level, leading to a functional decline. This ageing of human cells causes disorders like Cancer, Parkinson's disease and many other neurodegenerative diseases that are the major illnesses that plague most of the world today. Research confirms that environmental and lifestyle factors cause some epigenetic changes that directly affect the process of ageing. Some sites in the DNA called ageing-associated differentially methylated positions (aDMPs) are known to undergo these epigenetic changes. This analysis aims to use classification techniques like logistic regression and gradient boosting to prove that these aDMPs show significant differences in the epigenetic mechanisms they undergo compared to the non-aDMPs using the data on DNA methylation ratios and different histone modifications measured on various CpG sites for different cell ages. The analysis proves that this epigenetic signature of aDMPs exists and varies across proliferating (young) and senescent (old) cells.

I. Introduction

1.1. Ageing

"Ageing is an irreversible ongoing process characterized by a progressive decline in physical/mental activities that result in senescence" *(Mani, 2023)*. It is a major cause of many life-threatening diseases such as Cancer, Alzheimer's Disease (AD), Parkinson's Disease (PD) and many other neurodegenerative diseases that plague today's world. Two types of ageing occur in human beings - chronological ageing and biological ageing. Chronological ageing is simply a measurement of an individual's age in years, months, days, and hours from their time of birth. It is unaffected by any genetic or environmental changes and occurs at the same rate in every human being. Biological ageing, which is also known as intrinsic ageing refers to the changes that occur within the human body at the cellular and molecular level leading to a gradual decline in organ functions, tissue repair and the overall health of the individual. This ageing occurs at varying rates among human beings and is affected by a combination of factors that include genetic, environmental and lifestyle factors. Recent research has identified several underlying biological processes that contribute to ageing. Among these factors, cellular senescence and Epigenetic modifications are two of the most significant ones.

1.2. Cellular Senescence

There are trillions of cells present in the human body, which are the structural and functional units of life *(Wikipedia, n.d.)*. These cells are responsible for growth and tissue repair which is achieved through different phases of cellular division. An individual cell can either be active, dormant, or senescent. Cells in their active state are called proliferating cells. They undergo division leading to the production of daughter cells with identical genetic information, thus driving the growth. The proliferating cells, when subjected to any external stimuli or stress signals can enter a state of permanent cell cycle arrest and stop dividing altogether. This state is called cellular senescence. This happens in response to external and internal stress factors to maintain the survival of healthy cells and to eliminate damaged/dead cells under stress conditions *(Mani, 2023)*. Several studies have suggested that cellular senescence plays an important role in ageing and age-related diseases such as tissue degeneration and others including cancer.

1.3. Epigenetics

"Epigenetics is known to be the mechanism of how extrinsic factors modulate gene activity without causing any modifications in the DNA molecule" *(Mani, 2023).* These extrinsic factors include environmental and lifestyle factors. In simpler terms, epigenetic changes in our body are responsible for regulating gene expression and function by turning them on or off. The two major molecular mechanisms involved in Epigenetics are DNA methylation and histone modifications *(Esteller, 2007).*

"DNA methylation is the epigenetic progression of regulating gene expression by adding a methyl group to the fifth carbon position of the DNA sequence with the help of a specific enzyme called DNA methyltransferases". The highest level of DNA methylation occurs in the cytosine residues of DNA followed by guanine and these sites are called CpG sites *(Mani, 2023).* Some of these CpG sites are epigenetic markers of ageing and are known to exhibit different methylation patterns across different ages. These sites are called ageing-associated differentially methylated positions or aDMPs. Histones are major core proteins of chromatin that provide structural support to DNA around which the DNA wraps at continuous interims to form chromosomes. Histone modifications are the changes that occur by any covalent addition of a chemical group to these proteins. There are several different types of these modifications including acetylation and methylation *(Mani, 2023).*

It is known that exercise, a healthy diet, and good routines may delay the ageing process significantly *(Mani, 2023)*. Research confirms that these are the extrinsic factors that influence epigenetic changes that directly affect the process of ageing. Therefore, understanding Epigenetics is an essential factor in comprehending biological ageing and in developing new ways to delay its onset as well as mitigate ageing-related diseases.

1.4. Epigenetics and Statistical Methods

Applying appropriate statistical techniques considerably improves the capacity to reach reliable conclusions and make responsible judgements in both genetics and epigenetics. These techniques aid in the discovery of the underlying biological mechanisms and connections inside the intricate molecular networks that control epigenetic regulation and gene expression. Recent developments in the field of machine learning have further benefited researchers in getting more accurate results. Many supervised and unsupervised classification techniques have been used on the biological datasets. Supervised algorithms like neural networks and support vector machines are used when there are labelled datasets with two or more classes. Some studies have used these techniques to predict gene expressions under several different environmental and experimental conditions. Unsupervised algorithms like clustering are used in the case of unlabeled datasets. These have been used to identify genomic regions that play a biological role in many age-related disorders. Recent studies show the development of clocks that can predict the biological age of individuals which are simply based on correlation and regression methods. Epigenetic datasets usually have high dimensionality, whereas the cases of interest are very sparse. For example, specific genes that show mutations in case of certain age-related disorders like cancer usually occur in very low frequency. This makes the datasets imbalanced and to address these issues most studies use a combination of feature generation, selection, and machine learning. Thus, statistical methods including machine learning techniques can provide a promising future in the field of research for epigenetic changes that lead to ageing *(Lawrence B. Holder, 2017).*

1.5. Key Objectives

Numerous studies have shown that ageing and epigenetics are interconnected. Some epigenetic changes cause ageing and ageing results in some other epigenetic changes. This means that not all the CpG sites present in the cells and tissues of the human body are causing the body to age. They are changing due to the increase in age. These sites are the non-aDMP sites. But there are some specific CpG sites known as aDMPs (as mentioned above) that are known to exhibit a difference in epigenetic changes which leads to the onset of ageing, instead of being a result of it. This paper aims to understand the effects of epigenetic changes across the proliferating and senescent cells present in the human body and to determine how these changes are different in aDMPs and non-aDMPs, using different statistical techniques. This aim is achieved by fitting classification models on the epigenetic data and comparing the performances of different models based on their ability to segregate the aDMPs and non-aDMPs.

II. Literature Review

2.1. Epigenetics and Ageing

2.1.1. DNA Methylation

Epigenetic changes involve modifications to the DNA known as DNA methylation, without changing the DNA sequences and these changes are transferrable from parent to daughter cells. DNA is composed of two polynucleotides which contain four nitrogenous bases, Cytosine [C], Guanine [G], Adenine [A] or Thymine [T]. DNA methylation is the process of addition of a methyl group to the 5th carbon position of the C-G dinucleotide, which is known as the CpG position, where p stands for the phosphate that links the bases together. This nucleotide pair occurs with a very low frequency in the genome, and the areas that have a higher density of CpGs are known as CpG islands. Studies have shown that CpGs in these islands are often less methylated compared to the non-island CpG sites. Changes in DNA methylation begin at conception and continue throughout the lifetime of an individual. Studies assessing the level of DNA methylation at different age intervals of individuals concluded that methylation is not accurately maintained over cell division which leads to an increase in the variability of methylation levels with the growth of the individual. This phenomenon is called *epigenetic drift.* In addition to the variations and patterns in DNA methylation, studies suggest that specific regions in the genome are so highly associated with ageing, that in some cases they have been able to accurately predict the chronological age. This is a concept of the *epigenetic clock*. These age-related sites have been identified within a specific tissue and across other tissues in all individuals *(Anatomical Society, 2015).*

Both epigenetic drift and epigenetic clock are the basis on which many age prediction models have been built, using logistic regression and artificial neural networks. In a study conducted by David Ballard in 2017, a logistic regression model was used to predict the biological age of individuals using a dataset containing DNA methylation ratios of age associated CpG sites which were selected through stepwise variable selection. This model had a mean absolute age modelling error of 4.61 years and the correlation between the fitted and true ages of individuals was quite strong. Another model, fitted on the same data using an artificial neural network showed improved accuracy and an error rate of around 3.8 years. Notably, the error rate was higher in the case of individuals >60 years old and the model failed to perform well in the case of data from unhealthy individuals *(Ballard, 2017).*

The main question here is why changes in DNA methylation occur. Research has shown that these changes are highly influenced by environmental factors. The amazing property of DNA methylation is the ability for it to be altered by environmental stimuli, and in some situations, the marks produced are passable through cell divisions. A recent study on the population of smokers concluded that cigarette smoking can be linked to changes in the rate of methylation and these changes can be seen both in smokers as well as children of smokers. In a study conducted by Xiaohui in 2019, the Kruskal-Wallis test (an extension of the Mann-Whitney U test) was applied to check if there are significant differences between the biological age and chronological age of smokers. This analysis showed that smoking increased the epigenetic age of airway cells by an average of 4.9 years and lung tissue by 4.3 years (*Xiaohui Wu, 2019*). In addition to environmental factors, these changes might arise due to the incorrectly transcribed epigenetic patterns or marks during cell divisions, which is mainly a stochastic process. Thus, evidence suggests that both environmental and stochastic factors might lead to these changes in DNA methylation *(Anatomical Society, 2015)*.

2.1.2. Histone Modifications

A nucleosome polymer called chromatin is made up of DNA and different types of histone proteins. Several methods including histone modifications are used to change the structure of the chromatin. Environmental factors can alter chromatin's state, which then has an impact on the expression of genes related to ageing and long life *(Kim, 2020).* The first epigenetic modification occurs on the amino acids that make up histone proteins and these are known as post-translation modifications. These changes can change the shape of the histones. The most common modifications that occur in histones are acetylation and methylation which are generally reversible in nature.

A study was conducted in 2019 by Peggie Cheung, that used Principal Component Analysis and logistic regression models to distinguish cells based on the chromatin or histone modifications they undergo. These models were able to prove that the histone modifications increase in variability from young to older cells and these modifications are generally driven by non-inheritable changes *(Peggie Cheung, 2019).*

2.2. Importance of Understanding Epigenetics

"Epigenetics, the study of non-DNA sequence−related heredity, is at the epicenter of modern medicine because it can help to explain the relationship between an individual's genetic background, the environment, ageing, and disease" *(Andrew P. Feinberg, 2008)*. The major reason to understand epigenetics is to find new avenues that could benefit us in delaying the onset of ageing or age-related disorders. Epigenetic changes are significant for normal development, but disruptions to these changes can also lead to disturbances in normal development processes. For example, studies conducted on patients with colorectal cancer, show that there is a large loss of DNA methylation. This hypomethylation leads to abnormal activation of genes in cancer and genomic instability. Next, hypermethylation (increase in DNA methylation) of gene promoters is reported to be the cause of several tumor suppressor genes in cancer. Thus, in the case of cancer, epigenetic activation and silencing of genes turn on genes that ought to be off, and vice versa. This is followed by abnormal histone modifications *(Andrew P. Feinberg, 2008).*

Recent advancements in classical machine learning algorithms have significantly changed the field of cancer research and medical oncology. The concept discussed above has been proven by several statistical studies done on cancer research. A study conducted in 2021, compared samples of cancer patients(cases) with non-cancer patients(control) based on DNA methylation levels to find the exact regions in CpG sites that undergo changes that lead to cancer. A negative binomial model at a significance level of 0.1 was able to identify several hypermethylated CpG islands that were involved in colorectal, lung and pancreatic cancer. Hierarchical clustering and principal component analysis were also used to distinguish cases with controls. In PCA plots, the control groups were clustered tightly together, whereas the case groups were not clustered, probably because of the heterogeneity of tumors *(Jinyong Huang, 2021).*

Although this sector is expanding quickly, it is still in the early stages of development, and many of the technologies are still only utilized in animal models and have not yet received human approval (National Institute on Aging, 2021). The first step towards mitigating life-threatening age-related disorders is to try and delay the onset of ageing, because the functioning of all organ systems, both within and across individuals, deteriorates over time accounting for the onset of most diseases. There have been developments of many epigenetic techniques that use harmless viruses to introduce special genes that help transform mature or senescent cells into younger cells by reversing the epigenetic programming. These younger cells help regenerate some functionality lost due to age, illness, or injury. The study of epigenetics provides fascinating insights into how and why humans age at various rates (National Institute on Aging, 2021). Thus, it is imperative to find the exact sites in every tissue that influence the ageing process and the type of epigenetic changes they go through during the lifetime of an individual. Numerous studies have shown that aDMPs are such CpG sites that are directly associated with ageing and are mostly located within the CpG islands. Studying the DNA methylation patterns and the kinds of histone modifications these aDMPs show, both in the proliferating and senescent cells can further help in finding cures and epigenetic therapies for different age-related disorders, especially cancer. This paper aims to achieve this goal by using different classification models like logistic regression, random forest, and gradient boosting to determine the differences in methylation levels and histone modifications exhibited by these aDMPs and the rates at which their epigenetic modifications change as the cells age.

III. Data

To achieve the goals of the study, appropriate classification techniques were used to make comparisons between the two types of CpG sites- aDMPs and non-aDMPs, and the age of the human cells – Proliferate and Senescent. The analysis of the data occurred in four distinct phases. Firstly, the data was converted into a format suitable for analysis using transformation techniques. Then, some exploratory analysis was done to understand the dataset. This was followed by feature engineering and model fitting. In the final phase, the models' performances were evaluated and compared.

3.1. Dataset Description

The dataset consisted of abundance of 6 different types of histone modifications - H3.3, H4K16ac_ab1, H4K16ac_ab2, H4K20me3_ab1, H4K20me3_ab2 and H4 as well as DNA methylation ratios for proliferating and senescent cells each, measured at 285,000 CpG sites in human DNA using an Illumina 450k methylation array. H3 and H4 are the two main histones present in the chromatin. H3.3 is a variant of H3 histone, which is associated as a biomarker of certain types of tumor where the abundance of this histone variant reduces abnormally. H4K16ac_ab1, H4K16ac_ab2, H4K20me3_ab1, and H4K20me3_ab2 are formed by the methylation and acetylation of the core histone protein H4. The reduction in the levels of some of these variants is also associated with certain cancers. The total number of observations in the data was 485,096, out of which roughly 2100 CpG sites were ageing-associated differentially methylated positions (aDMPs) and the rest were non-aDMPs. The response variable was aDMP_status which was a binary variable with 1 for the aDMP site and 0 for the non-aDMP site.

This dataset was made fit to be used for valuation. The first step was to treat those instances that had no value. The data consisted of roughly 2600 missing values for the methylation ratios of both proliferating and senescent cells. These missing values were imputed using the grouped ratio means for aDMPs and non-aDMPs. The dataset was then split into training (70%) and test sets (30%). This data appeared to be skewed as it contained a lot of zero values. Thus, transforming the data was necessary to get the best results which was achieved by applying appropriate log transformations. After the transformation, the data was scaled using the min-max scaling method.

3.2. Exploratory Analysis

The tables below show a short numerical summary of 2 different types of histone modifications selected randomly and methylation ratios of proliferate and senescent cells divided into aDMPs and non-aDMPs.

Variable	Mean	SD	Min	Max	IQR
Prolif_H3.3	0.0202	0.0302	0.0000	1.0000	0.0183
Senes_H3.3	0.0016	0.0024	0.0000	0.1030	0.0015
Prolif_H4	0.0010	0.0011	0.0000	0.0591	0.0008
Senes_H4	0.0004	0.0006	0.0000	0.0556	0.0002
Prolif_Meth	0.4402	0.3875	0.0000	1.0000	0.4240
Senes_Meth	0.4647	0.3975	0.0000	1.0000	0.4344

Table 1 – Numerical summary of non-aDMP sites

Variable	Mean	SD	Min	Max	IQR
Prolif_H3.3	0.1134	0.1329	0.0000	1.0000	0.0872
Senes_H3.3	0.0020	0.0023	0.0000	0.0170	0.0016
Prolif_H4	0.0009	0.0010	0.0000	0.0141	0.0008
Senes_H4	0.0003	0.0003	0.0000	0.0027	0.0002
Prolif_Meth	0.2063	0.2678	0.0000	1.0000	0.2378
Senes_Meth	0.2657	0.3049	0.0000	1.0000	0.3426

Table 2 – Numerical summary for aDMP sites

From Table 1 for non-aDMP sites, we can see that there is a loss in histones as the cell ages. The mean abundance of H3.3 histone is decreasing from 0.0202 in proliferating cells to 0.0016 in senescent cells. The same pattern can be seen for H4 histone which has a maximum abundance of 0.0591 in proliferate cells which decreases to 0.0556 in senescent cells. In comparison, the methylation levels are increasing with the age of the cells. The aDMP sites show a similar pattern in Table 2, loss of histone abundances and increase in methylation levels as the cell ages into senescence. When compared based on the aDMP status, we can see that aDMP sites have higher histone abundances and lower methylation ratios than the non-aDMP sites. This matches the results of most of the scientific studies discussed earlier, that the aDMP sites appear to be less methylated as compared to the non-aDMP sites. Thus, initial analysis suggests that there might be a difference between the two CpG sites based on histone modifications and methylation ratios and that this signature varies across proliferating and senescent cells.

The next step in the analysis involved checking if there were any outliers present in the data. The presence of outliers was checked by plotting boxplots of different features. Boxplots provide a visual representation of the data and help in finding any outliers present in the data. Figure 1 shows boxplots of the methylation ratios of proliferate and senescent cells based on the aDMP status. From the boxplot for proliferate cells, we can see that there are some outliers present for the aDMP sites, which can be seen as individual points outside the "whiskers" of the plot. Boxplots for histone modifications also detected several outliers.

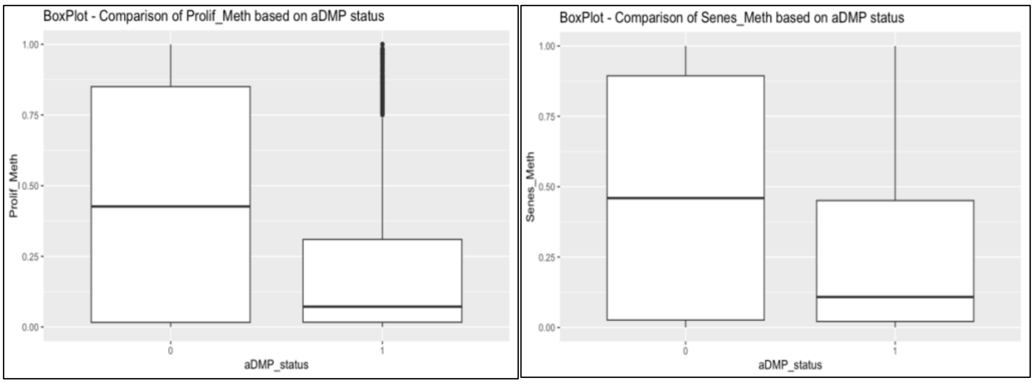

Figure 1 – Boxplots of methylation ratios

The data used to generate these boxplots was transformed through a log transformation as mentioned above. To treat the outliers, several different transformations were also used including the square root transformation and box-cox transformation. Square root transformation involves taking the square root of the data, whereas box-cox transformation, used to stabilize the variance, is conducted using the "boxcox" function from the "MASS" package in R. Despite trying different transformations there was not much effect on the outliers. Since the data is quite large, containing around 485,000 values, it would be quite difficult to eradicate all the outliers. Also, with increased data sizes there might be an increase in the sensitivity of analysis methods and diversity of data patterns. Visualizing larger datasets can also be challenging, which might result in an increased number of outliers shown. Thus, the outliers were ignored, and the analysis was continued using log-transformed data.

Before fitting the classification model, it was important to check if there is any multicollinearity present between the features in the data which might affect the results. Below are the Pairplots of all the features.

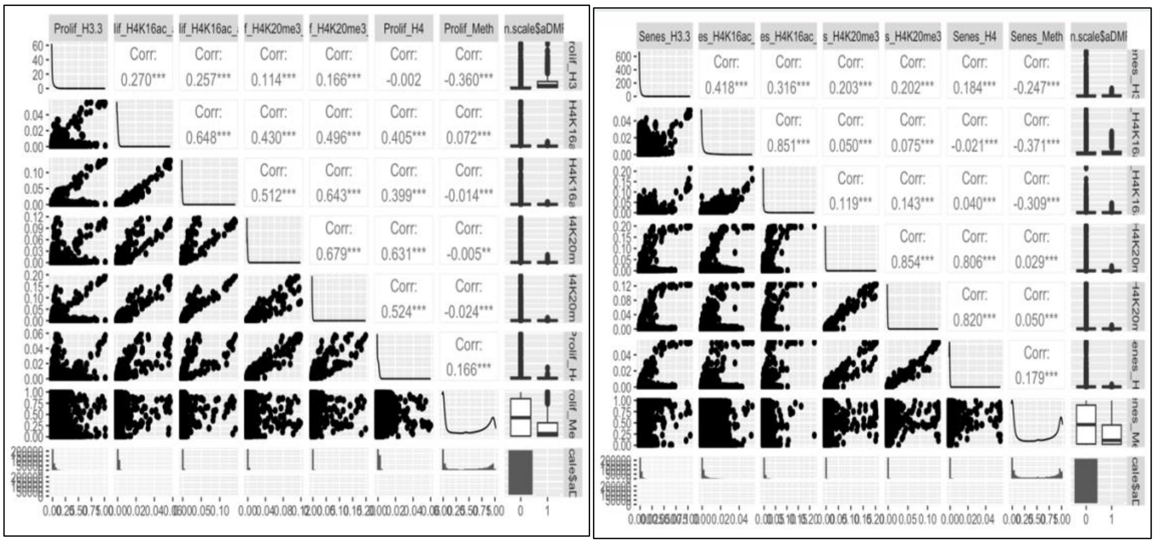

Figure 2 – Pairs plots

From Figure 2, we can conclude that there is some multicollinearity present in the data. Pairplots for features of proliferating cells show that the maximum correlation between any two variables is around 0.6, whereas plots for features of senescent cells show a little more multicollinearity in comparison. The maximum correlation here is 0.85. Hence, there was a further analysis conducted on the features using LASSO regression. LASSO (Least Absolute Shrinkage and Selection Operator) regression, also known as L1 regularization is a statistical technique commonly used for feature selection. A penalty term is added to the linear regression cost function that minimizes the sum of squared residues and shrinks some of the coefficients of features to zero, performing feature selection by removing less important features (Kumar, 2023). The model was fitted with the optimal regularization parameter value chosen through cross-validation, the value with the minimum mean squared error. The result showed that all the features in the dataset had non-zero coefficients, thus showing that all of them should be included at this point of valuation. Figure 3 displays a graph of different values of the regularization parameter against the corresponding mean squared error for the value. It is the least when all the 14 features are included.

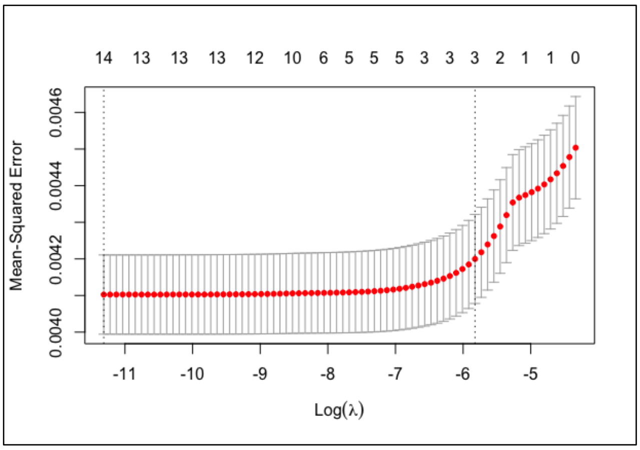

Figure 3 – LASSO Regression for feature selection

The next phase of this paper was divided into two parts to get two different results from the dataset. The first phase included fitting Logistic Regression, Random Forest, and Gradient Boosting models to the entire dataset to determine if aDMPs have distinguishing marks based on histone modifications and methylation ratios compared to the non-aDMPs judging by the model performances. The most important features were selected and their effects on the aDMP statuses were judged. The second phase included fitting two different models on the dataset after it was divided into two different sets based on the cell age – proliferate and senescent and significant features were selected. The model used for this phase was the model that performed the best in phase one.

IV. Phase 1 – Is there an epigenetic signature for aDMPs?

4.1. Methods

The response variable – aDMP_status is a binary variable, and the main aim is to compare two groups and check if the given data can be segregated into aDMPs and non-aDMPs based on epigenetic modifications, thus the classification models that have been fitted to the data are:

1. Logistic Regression
2. Random Forest
3. Gradient Boosting

All these models are supervised classification models which are used in the case of labelled datasets with two or more classes, as in this case.

4.1.1. Logistic Regression

Logistic regression is the most appropriate supervised classification model when the response variable is binary. It is used in tasks where the goal is to predict the probability of an instance belonging to a given class. It is referred to as regression as it takes the output of a linear regression function as input and uses a sigmoid function to determine the probability of a given class. Linear regression is appropriate when the response variable is continuous, and it assumes that the dependent variable is normally distributed. Thus, it does not work in the case of binary response variables as this assumption is violated.

A sigmoid/logistic function is a mathematical function which is used to map predicted values to probabilities. It maps a real value to another value which lies between 0 and 1. As these values cannot go beyond this range, it forms a curve in the shape of an S. The sigmoid function is of the form:

$$P = \frac{1}{1 - e^{-(\beta_0 + \beta_1 * x)}}$$

The equation of best line in linear regression is:

$$y = \beta_0 + \beta_1 * x$$

In the case of logistic regression, we replace y in the above equation with the probabilities. But in this case, the values may not lie within the range 0 to 1. To resolve this issue, we take the odds of the probabilities, which is the ratio of success over failure. But this ratio is always greater than 0, which makes our range restricted. Thus, log odds are taken instead to overcome this issue. This makes the equation for logistic regression as:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 * x$$

The response variable is assumed to follow the Bin (1, p) distribution. We can solve the above equation to get the required probabilities which gives us back the sigmoid/logistic function as above. The left-hand side of the equation represents the link function for the regression model. This logarithm of odds ratio is known as the logit link function, which has been used for analysis in this study. Although logit link is the most common link function, but there are other functions like probit link and complementary log-log link functions which can be used depending on the purpose and requirement.

Since the dataset was quite large, it was not possible to run the model for the entire dataset altogether. Hence, the data was divided into smaller samples, models were run separately for each sample and then the results were combined to give a final model. The samples were created through a process called Bootstrapping. It is a kind of resampling technique that is used to repeatedly draw samples with replacements from a dataset. It is based on the law of large numbers which states that a sample average from a large sample is a close approximation to the true population average and it gets closer to the true value as the size of the sample increases. Different models created based on these samples were later combined.

The current dataset appeared to be majorly imbalanced as there were exactly 2168 aDMP sites (minority class) as compared to the non-aDMP sites (majority class) which accounted for only 0.45% of the total data. Thus, there was a need to address this imbalance. This was addressed in two different ways:

1. Undersampling – In this technique, all the data is kept for the minority class and the majority class is reduced in size. A sample of roughly the same number of values as the minority class is randomly selected for the majority class.
2. Class weights – In this technique, weights are assigned to both classes, with a higher weight assigned to the minority class so that the model takes more consideration of the patterns of the minority class and reduces the bias towards the majority class (Singh, 2020).

The logistic regression model with bootstrapping was re-run using both techniques and the model performances were compared based on the AIC of the models. AIC stands for Akaike Information Criterion and is used to compare models to decide which best fits the data. It uses a model's maximum likelihood estimate as a measure of fit and adds a penalty with increasing parameter complexity. The lower the AIC value, the better the fit of the model. AIC values from all the three models are:

Simple model	**With Undersampling**	**With Class weights**
947.11	2452.9	35458

Table 3 – AIC values for different logistic regression models

As the AIC value is the lowest for the simple logistic regression model, it appears to be the best fit for the data out of the three models. This model was further used for feature selection. There was a total of 14 features in the data. The model summarizes the coefficients and p-values of all the features. The features with p-values greater than 0.05 are insignificant, whereas the features with p-values lower than 0.05 are significant. Based on this criteria, only 5 features were selected as significant with a p-value less than 0.05. These were Prolif_H3.3, Prolif_Meth, Senes_H3.3, Senes_H416ac_ab1 and Senes_Meth.

4.1.2. Interpretation of Feature effects

There were top 5 features selected from the dataset, which were considered to have the most effect on the response variable. The graph below shows the Odds ratio of all the selected features.

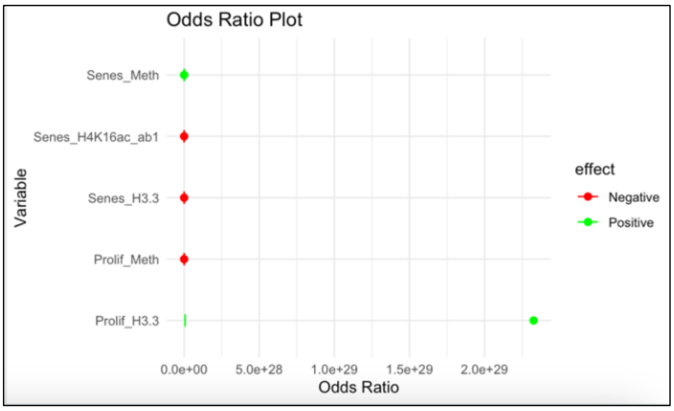

Figure 4 – Odds Ratio of selected features

In this analysis, the Odds ratio measures the change in odds of the event occurring i.e., the site being an aDMP site associated with a unit increase in the selected feature. If the odds ratio is greater than 1, then the chances of the site being an aDMP site increase with every one-unit increase in the explanatory variable and vice versa. In Figure 6, the features with an Odds ratio in green colour correspond to a ratio greater than 1, whereas the features with the ratio in red means a ratio less than 1. For example, the methylation ratios of both the senescent and proliferating cells are green i.e., greater than 1. This means that as the methylation increases, the odds of the selected site being an aDMP site increase. Whereas all the histone modifications have an odds ratio less than 1, which means that as the abundance in histone modifications increases, the odds of the selected site being a non-aDMP site increases. This interpretation thus suggests that aDMP sites become hypermethylated and lose more histones in comparison to the non-aDMP sites.

4.1.3. Random Forest

Random forest is a supervised machine learning algorithm that is constructed through a decision trees algorithm. A decision tree is a hierarchical classification model that uses a branched structure to make decisions based on the input data features. The feature space containing all the features in the data is divided into several disjoint regions. However, decision trees usually have high variance and low accuracy rates compared to other classification algorithms. Random forest is an improvement to the decision tree algorithm which consists of many decision trees. It uses the method of ensemble learning, which is a technique where results from various decision tree classifiers are combined to provide the outcome or predictions. The 'forest' of decision trees generated is trained through bagging or bootstrapping *(Mbaabu, 2020).*

The random forest method has several hyperparameters that are used to improve the predictive power of the model. Hyperparameters are the parameters of a model that are set before the training model, to control the model performance and complexity. Two major hyperparameters most used are:

1. Number of trees – This is the number of decision trees the algorithm would build before taking the average of the outcome or predictions. A higher number of trees improves the performance of the model and prevents overfitting, although it slows down the computation.

2. Number of features – There can be some features that are dominant in a dataset and most trees take that same feature for the first split leading to a high similarity among trees and hence high correlation. Random forest improves upon this by taking a random sample of features out of the feature space, every time a split is considered. This sample size or number of features to be considered is another hyperparameter.

In the analysis, the random forest method has been used with tuning of only one of these hyperparameters – the number of features. As the dataset is very large, there were some computation and memory issues. Firstly, a random forest model was built on the entire dataset. This was done to check if the features selected from the logistic regression model match the significant features selected based on the random forest model. The model was run using several different numbers of features and trees. The best hyperparameters were the ones that had the lowest error rate i.e., misclassification rate. This rate refers to the number of data points that do not belong to the class they have been assigned to.

The model was then used to check feature importance. For each feature, the average decrease in Gini Index was calculated. Gini index refers to the probability that a random instance is being misclassified. The smaller the value of this index, the lower the misclassification rate. Thus, if a feature leads to a greater decrease in the Gini index, it is more significant.

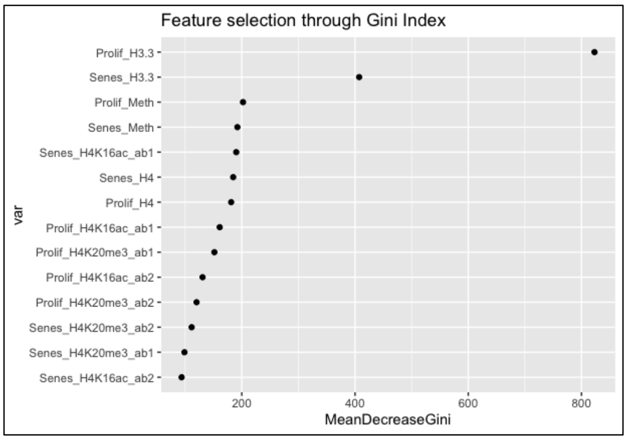

Figure 5 – Feature importance based on the Random Forest model.

Figure 4 displays the average decrease of Gini index for every feature present in the data. Prolif_H3.3 had the greatest decrease in Gini index, making it the most significant. The top 5 features selected from the logistic regression model indeed match the top 5 features with a great reduction in the Gini Index.

The random forest model was run on the dataset with selected features, same as the logistic regression model. The hyperparameter – number of features was selected through hyperparameter tuning.

4.1.4. Gradient Boosting

Boosting is another ensemble learning technique that combines predictions from several weak learning models to get a strong model with enhanced performance and improved predictions. Weak learning models are the ones with higher misclassification rates. For example, suppose there are n data points with 2 classes. Now, the first model or weak learner picks up some observation points randomly and predicts them. Weights are then assigned to the data points, with higher weights assigned to the points that have been misclassified by the first model. This increases the probability that the second weak learning model selects that misclassified data point. Then the third model works on the data points misclassified by the second model. This process continues until the misclassification errors are minimized. Boosting algorithm has two different algorithms under its umbrella – Adaptive Boosting and Gradient Boosting. For this paper, only gradient boosting algorithm has been used as it is better in handling bigger and more complex datasets *(Saini, 2021)*.

The Gradient Boosting algorithm works on the boosting algorithm with the aim of minimizing the loss function to make the model's predictions more accurate. The loss function measures how well the model labels the data points with the correct classes. For this analysis, gradient boosting has been used with a more advanced and optimized implementation of the boosting algorithm - Xgboost, as the dataset is quite large and complex. Two different models have been built using gradient boosting in two different ways.

1. Using 10-fold cross-validation only: Cross-validation is a type of resampling procedure used to enhance the model's performance. It has one parameter – k, which refers to the number of parts the dataset must be divided into (here, k is 10). Under this technique, the data is split into k parts, out of which one part is randomly kept as a test set and all the other parts are used as training sets. The model is fit on the training sets, evaluated on the test set and the score is recorded. This process is repeated until all k parts are chosen as the test set. All the recorded scores are then averaged to give the entire model's performance.

2. 10-fold cross-validation and hyperparameter tuning – Gradient boosting has several hyperparameters. The parameters used are the following:

    - Number of trees

    - Interaction depth – refers to the maximum depth of the tree i.e., the number of nodes present in the tree. Increasing the depth makes the trees more complex and helps in capturing every minute detail of the data.

    - Shrinkage – This is the learning rate of the model, that decides the rate of contribution of all the weak learners to the final model. This helps in preventing overfitting.

    - Number of observations in the node – refers to the minimum number of samples that are required to create a node in a decision tree.

4.2. Assessment

The measure that was used to compare the performances of all the models built in the analysis was AUC (Area Under the ROC Curve). An ROC (Receiver Operating Characteristic) curve is a graphical representation of the performance of a binary classification model which plots two parameters:

1. True Positive rate/Sensitivity – The proportion of positive cases correctly classified as positive:
$$TPR = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

2. False Positive rate/1-Specificity - The proportion of actual negative cases incorrectly classified as positive:
$$FPR = \frac{False\ positive}{False\ positive + True\ negative}$$

ROC curve plots TPR vs. FPR for different classification thresholds. AUC measures the entire area under this ROC curve. AUC provides an aggregate measure of the model's performance at every threshold. Its value ranges from 0 to 1. A model with 100% correct predictions has an AUC value of 1 whereas a model with only 50% correct predictions will have an AUC of 0.5. A model with an AUC greater than 0.5 performs fairly well.

In this analysis, the main focus is on evaluating the models' performances based on their ability to distinguish the aDMP and non-aDMP sites. In this context, using the AUC-ROC measure can be the best option for model comparisons, as it considers the balance between correctly identifying both positive and negative cases. Also, since our dataset is imbalanced, this measure can provide a more comprehensive evaluation than any other measure.

4.3. Results

The AUC for all the models fitted are displayed in Table 4.

Model	**AUC**
Logistic Regression	0.868
Logistic Regression after Feature selection	0.869
Random Forest	0.843
Gradient Boosting with cross validation	0.911
Gradient boosting with hyperparameter tuning	0.889

Table 4 – AUC for all models

All the models fitted on the dataset had an AUC greater than 0.8, thus indicating that the fit of these models was quite good. All models were able to segregate the aDMPs from non-aDMPs based on histone modifications and methylation ratios. The best performance out of all models was of the Gradient boosting model with 10-fold cross-validation.

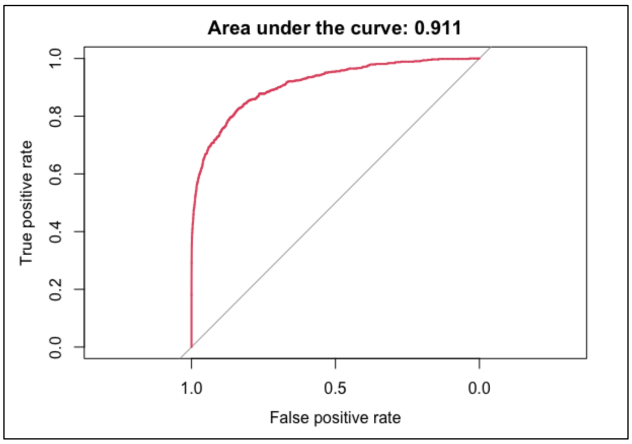

Figure 6 – ROC for Gradient Boosting model with cross-validation

V. Phase 2 – Does this epigenetic signature vary based on cell type?

5.1. Methods

For this phase of the analysis, the dataset was divided based on the cell type, into two parts. The first part contained information about proliferating cells with all the features and the second part contained all features related to the senescent cells. The analysis was done on these two datasets taking aDMP status as the response variable. The best model from Phase 1 was the gradient boosting model with 10-fold cross-validation and this was used for Phase 2 as well. The aim here was to build classification models on both the datasets and confirm based on the model performance if the aDMP sites can be segregated from the non-aDMP sites i.e., if the epigenetic signature of aDMPs exists or varies across the proliferating and senescent cells.

5.1.1. Proliferating cells

The first model built was on the dataset containing all the seven features of proliferating cells including 6 different histone modifications and the methylation ratio. After the model was run, the feature importance was checked for this data.

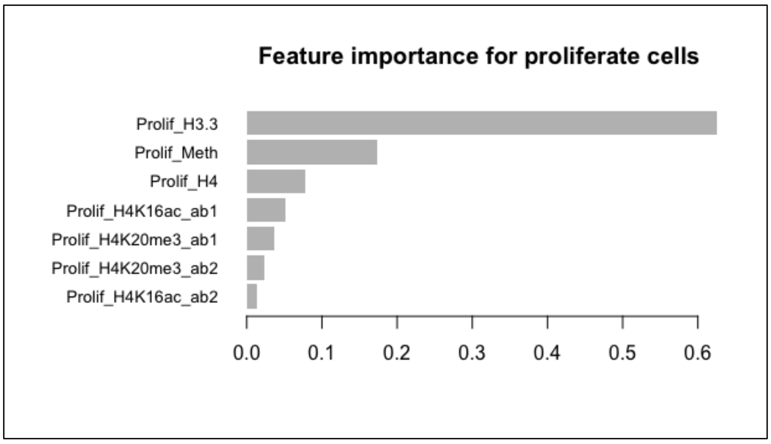

Figure 7 – Feature importance for proliferating cells.

Figure 6 displays the graph of feature importance obtained from the gradient boosting model. The most important feature is Prolif_H3.3 – the abundance of H3.3 histone, followed by Prolif_Meth – the methylation ratio. These two features were selected as significant features for proliferating cells in Phase 1 of the analysis as well. Thus, these two significant features were used to build a final model for the proliferating cells.

5.1.2. Senescent cells

For Senescent cells, the same approach was followed. The first model was built and then the feature importance was checked.

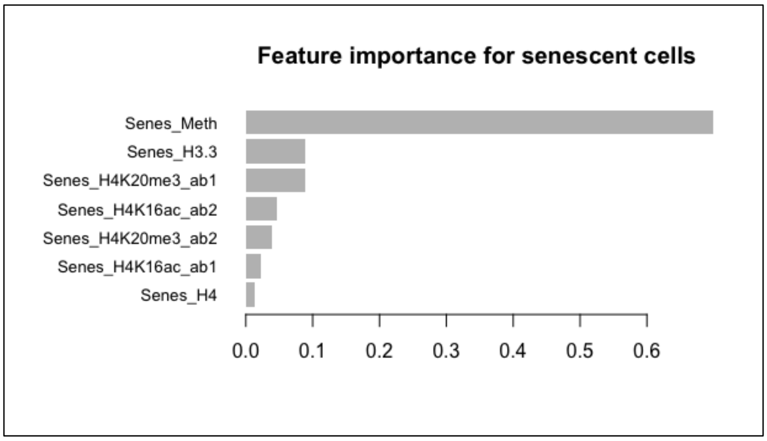

Figure 8 – Feature importance for senescent cells.

The most significant feature here was the Senes_Meth – the methylation ratio for senescent cells, followed by the abundances of the two histones H3.3 and HK16ac_ab1. These also match the significant features selected from Phase 1. The final model for senescent cells was built using these three significant features.

5.2. Results

The AUC for both the fitted models along with the AUC of the final model on the entire dataset is displayed in Table 5.

Model	**AUC**
Proliferating cells	0.846
Senescent cells	0.679
All cells	0.911

Table 5 – AUC for all models

The AUC for all models is greater than 0.5 indicating that the models performed fairly well in segregating the aDMP sites from the non-aDMP sites. The epigenetic signature of aDMPs thus exists and can be seen the most when the entire data is considered. This signature exists when proliferating and senescent cells are considered separately as well, although it is more prominent in the case of proliferating cells as indicated by the AUC for that model, which is greater than the model for senescent cells. This difference can also be seen through the ROC curves for both models.

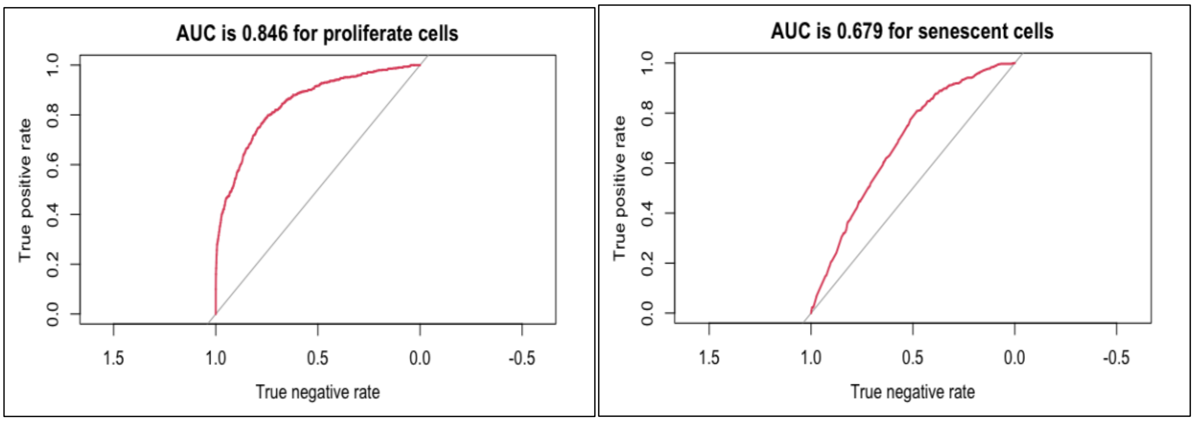

Figure 9 – ROC curves comparison based on cell age.

5.3. Feature interpretation – Does the epigenetic signature vary?

The models suggested that the aDMPs are different from non-aDMPs based on the epigenetic modifications, i.e., aDMPs have an epigenetic signature and this signature can be seen in both cell types – proliferating and senescent. Not only does the signature exist when both types of cells are considered separately, but it also varies. For proliferating cells, the aDMPs can be well separated from the non-aDMPs based on the histone modifications, whereas for senescent cells this segregation is the most significantly done by their methylation ratios as shown by the models for both cell types.

This variation can also be seen through different plots of the features. Figure 8 shows the density plots of the methylation ratios for both proliferating and senescent cells for both aDMPs and non-aDMPs. From these graphical representations, it is clear that the methylation ratios are different for the aDMPs in both proliferating and senescent cells. The aDMPs in proliferating cells are hypomethylated as compared to the senescent cells.

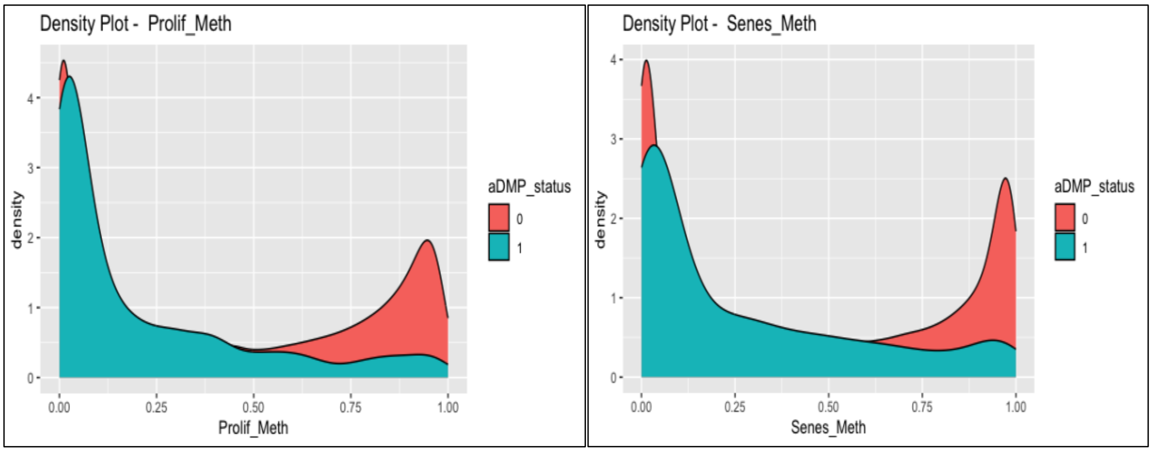

Figure 10 – DNA methylation rates comparison

VI. Conclusion

This paper thus provides evidence through graphical and numerical analysis as well as different classification models that the aDMPs are different from non-aDMPs based on epigenetic changes like DNA methylation ratios and modifications of various histones. With ageing, both the aDMPs and non-aDMPs undergo certain changes. The changes in aDMPs lead to ageing and age-related diseases, whereas the changes in non-aDMPs are considered to be a consequence of ageing. Until adulthood, the methylation ratios of non-aDMPs keep increasing, whereas the aDMP sites show very little methylation. Post-adulthood, the non-aDMPs either stop methylating or get hypomethylated and the aDMPs start getting hypermethylated. In the case of histone modifications, there is a loss at both aDMP and non-aDMP sites, but this loss is much more significant for the aDMP sites. The histone modifications that show significant differences are H3.3 and H4K16ac_ab1, which are variants of the H3 and H4 histones.

The first phase of the analysis demonstrated that an overall signature of the aDMPs exist, where both the proliferating and senescent cells were considered together. The second phase proved that this epigenetic signature of aDMPs not only exists but varies based on the different types of cells. The aDMPs in proliferating cells can be segregated more significantly from the non-aDMPs based on the differences in histone modifications whereas in senescent cells, they show a larger variation from the non-aDMPs based on the methylation ratios.

VII. Discussion

7.1. Limitations of this analysis

Epigenetic datasets are usually very large with the genes of interest occurring in very low frequencies. This leads to data imbalance that should be treated very carefully in order to avoid getting wrongful insights. Thus, limitations in this analysis arose due to the dimensionality of the dataset.

Due to the length of the data, a huge number of outliers were also seen. Significant measures were taken to treat them, however complete eradication was not possible. This might have occurred due to the software limitations as well, which show an increase in outliers with a huge increase in the size of the datasets.

The size of this dataset also created issues while applying hyperparameter tuning for classification models, especially in the case of Random Forest model. It was impossible to tune the parameter corresponding to the number of trees in the model, as discussed in Phase 1 of this paper.

7.2. Further work

This analysis aimed to use supervised machine learning techniques to segregate the data based on a binary variable. There were only three different types of techniques used. Statistics and Data sciences are fields where every problem has many different acceptable solutions. Other techniques and analysis methods can be applied here to get more meaningful insights. Support Vector machines and neural networks are some examples. The metric used for this analysis was AUC, however other metrics like Precision and accuracy can also be used to compare the performances of models.

This analysis proved that the aDMPs have an epigenetic signature. It helped in determining the exact changes that occur at these sites and how these change with the ageing of cells. The results of this analysis can be used further to detect the changes that occur in the case of age-related diseases. For example, comparisons can be made between the methylation ratios of certain diseased cells to healthy ones using logistic regression or gradient boosting, which can help find the exact changes that occur in diseased cells. This can be extended to finding ways to obliterate the factors that lead to these modifications, thus helping cure many age-related diseases.

The aDMPs usually refer to the positions in the DNA where the epigenetic changes occur. While these sites have been proven to be associated with age-related changes in the body, they have been shown to be more correlated to chronological age and have less contributions to the biological ageing process. Recent studies have discovered the occurrence of aVMPs – ageing-associated variably methylated positions which have been claimed to play a more significant role in ageing mechanisms as compared to the aDMPs *(al, 2016)*. Research can be done using similar analysis methods on the epigenetic datasets containing information about aVMPs to get more insightful results about the ageing process and find ways to treat or delay the onset of age-related diseases.

References

al, S. e., 2016. Age-related accrual of methylomic variability is linked to fundamental ageing mechanisms. *Genome Biology,* Volume 17.

Anatomical Society, 2015. DNA methylation and healthy human ageing. *Ageing cell,* 14(6), pp. 921-1127.

Andrew P. Feinberg, M. M., 2008. Epigenetics at the Epicenter of Modern Medicine. *JAMA Network.*

Anon., n.d. [Online].

Ballard, D., 2017. DNA methylation-based forensic age prediction using artificial neural networks and next generation sequencing. *Elsevier.*

Esteller, M. F. F. a. M., 2007. Epigenetics and ageing: the targets and the marks. *Trends in Genetics,* 23(8), p. 413.

Jinyong Huang, A. C. S. B. D., 2021. Cancer Detection and Classification by CpG Island Hypermethylation Signatures in Plasma Cell-Free DNA. *Cancers,* 13(22).

Kim, S.-J. Y. a. K., 2020. New Insights into the Role of Histone Changes in Ageing. *International Journal of Molecular Sciences,* 21(21).

Kumar, D., 2023. *A Complete understanding of LASSO Regression.* [Online]
Available at: https://www.mygreatlearning.com/blog/understanding-of-lasso-regression/
[Accessed 2023].

Lawrence B. Holder, M. M. H. a. M. K. S., 2017. Machine learning for epigenetics and future medical applications. *Epigenetics,* Volume 12.

Mani, V. S. a. I., 2023. *Epigenetics in Health and Disease.* 1 ed. s.l.:Elsevier Science & Technology.

Mani, V. S. a. I., 2023. *Epigenetics in Health and Disease.* 1 ed. s.l.:Elsevier Science & Technology.

Mbaabu, O., 2020. *Introduction to Random Forest in Machine Learning.* [Online]
Available at: https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/

National Institute on Ageing, 2021. *The epigenetics of ageing: What the body's hands of time tell us.* [Online]
Available at: https://www.nia.nih.gov/news/epigenetics-ageing-what-bodys-hands-time-tell-us
[Accessed 2023].

Peggie Cheung, F. V. H. C. W., 2019. SINGLE-CELL CHROMATIN MODIFICATION PROFILING REVEALS INCREASED EPIGENETIC VARIATIONS WITH AGEING. *HHS public access.*

Saini, A., 2021. *Gradient Boosting Algorithm: A Complete Guide for Beginners.* [Online]
Available at: https://www.analyticsvidhya.com/blog/2021/09/gradient-boosting-algorithm-a-complete-guide-for-beginners/#undefined
[Accessed 2023].

Singh, K., 2020. *How to Improve Class Imbalance using Class Weights in Machine Learning?.* [Online]
Available at: https://www.analyticsvidhya.com/blog/2020/10/improve-class-imbalance-class-weights/#:~:text=One%20effective%20technique%20for%20addressing,bias%20towards%20the%20majority%20class.
[Accessed 2023].

The Company of Biologists, 2021. *Stem cell quiescence: the challenging path to activation.* [Online]
Available at: https://journals.biologists.com/dev/article/148/3/dev165084/237446/Stem-cell-quiescence-the-challenging-path-to
[Accessed 2023].
Wikipedia, n.d. *DNA.* [Online]
Available at: https://en.wikipedia.org/wiki/DNA
[Accessed 2023].
Wikipedia, n.d. *Wikipedia.* [Online]
Available at: https://en.wikipedia.org/wiki/Cell_(biology)#Cellular_processes
Xiaohui Wu, Q. H. R. J. J. Z. H. G. &. H. L., 2019. Effect of tobacco smoking on the epigenetic age of human respiratory organs. *Clinical Epigenetics,* Volume 11, p. 183.