**AGE PREDICTION IN HUMANS USING A MACHINE LEARNING APPROACH**

BY

ADEGITE JUWON

MGBAFULU CHIOMA

ADEWUYI EMMANUEL

MAURICE GODWIN

AJUDUA SHARON

BEING A GROUP PROJECT REPORT SUBMITTED AS PART OF THE REQUIREMENT FOR THE COMPLETION OF THE HACKBIO BIOINFORMATICS INTERNSHIP THEMED AI \* BIO

MARCH, 2025

**TABLE OF CONTENTS**

[1 INTRODUCTION 3](#_Toc192852170)

[**1.1** **Problem Statement** 3](#_Toc192852171)

[**1.2** **Research Question** 4](#_Toc192852172)

[**1.3** **Specific Objectives** 4](#_Toc192852173)

[2 MATERIALS AND METHODS 4](#_Toc192852174)

[**2.1** **Data Source And Preprocessing** 4](#_Toc192852175)

[**2.2** **Model Building** 4](#_Toc192852176)

[**2.3** **Feature Selection** 5](#_Toc192852177)

[**2.4** **Model Evaluation** 5](#_Toc192852178)

[3 RESULTS 5](#_Toc192852179)

[4 DISCUSSION 9](#_Toc192852180)

[5 CONCLUSION 9](#_Toc192852181)

[6 REFERENCES 9](#_Toc192852182)

# INTRODUCTION

Human aging is a continuous and multifaceted biological process, marked by gradual shifts in cellular function and metabolism. These changes progressively diminish the body's ability to repair itself, maintain stability, and ultimately lead to deterioration in the structure and function of tissues and organs.( Dziechciaz and Filip, 2014)

As individuals age, they generally experience a decrease in both physical and mental capabilities, along with a heightened susceptibility to diseases (Simpson and Chandra, 2021). The human genome contains around 28 million CpG dinucleotides, where a cytosine comes before a guanine in the 5’-3’ orientation. These cytosines have the potential to be methylated, resulting in the formation of 5-methylcytosine, which is referred to as DNA methylation (DNAm). DNAm has been said to be one of the biomarker for biological aging(Li *et al*., 2022).

Aging is also affected by many different things, and a better way to predict biological age is to use machine learning, a scientific method that analyzes many aspects of a person's health and well-being to make predictions (Ni *et al*., 2023).

## **Problem Statement**

Aging is a complex biological process influenced by genetic, environmental, and epigenetic factors. DNA methylation at specific CpG sites has been shown to correlate with chronological age, leading to the development of epigenetic clocks. However, accurately determining an individual's age from biological samples remains a challenge in aging research, forensic science, and healthcare. Traditional methods often rely on physical or medical assessments, which may not always be precise or feasible. This project aims to develop a machine learning model to estimate an individual's age based on the methylation status of 1000 CpG sites. By doing so, it provides a more accurate measure of biological age, which has critical applications in understanding aging-related diseases, forensic identification, and assessing disease risk. In forensic science, such a model can aid in age estimation from biological evidence, assisting in criminal investigations and missing person cases. In healthcare, it can differentiate between biological and chronological age, helping detect premature aging and assess health risks.

## **Research Question**

* How accurately can we predict the chronological age of an individual using DNA methylation data from 1000 CpG sites?
* Which regression model performs best for age prediction in this dataset?
* Is there a trade-off between model simplicity (fewer features) and accuracy?
* How does reducing the number of features affect the model’s predictive performance (as measured by MSE)?

## **Specific Objectives**

* Develop a Regression Model for Age Prediction
* Compare Model Performance
* Investigate Feature Selection Impact
* Evaluate the Trade-off Between Simplicity and Accuracy
* Identify Key CpG Sites for Age Prediction

# MATERIALS AND METHODS

## **Data Source And Preprocessing**

The link to the dataset was provided by hackbio. We used the dataset kaggle/input/cpgmeth2age/CpGmeth2Age.csv for this project. The data preprocessing and ML model prediction was implemented in python 3.6.9 using google colab. We have a dataset containing information on 108 individuals. For each individual, we have DNA methylation levels at 1000 CpG sites, representing the methylation status at specific locations in their DNA. We also have the various ages in the dataset which is the value we want to predict.

The various libraries were imported which include; pandas, numpy, seaborn, matplotlib.pyplot, sklearn. The dataset was then cleaned by removing empty rows and columns, removing duplicate values and sorting missing values. The dataset were then splitted into test and train data.

## **Model Building**

A baseline model was built. We used Linear Regression and Random Forest model to learn the relationship between DNA methylation patterns and age.

## **Feature Selection**

Instead of using all 1000 CpG sites, we experimented using a reduced number of features (90, 80, 70, ..., 10). This is done to potentially improve model performance and reduce complexity. We use a technique called Recursive Feature Elimination (RFE) to select the most important features and

## **Model Evaluation**

To assess the performance of the models with different feature subsets, we calculated three metrics:

* **Mean Squared Error (MSE):** Measures the average squared difference between predicted and actual ages. Lower MSE indicates better accuracy.
* **Root Mean Squared Error (RMSE):** The square root of MSE, providing a more interpretable measure of error in the same units as the target variable (age).
* **R-squared (R2) score:** Represents the proportion of variance in the target variable (age) explained by the model. Higher R2 score indicates a better fit.

# RESULTS

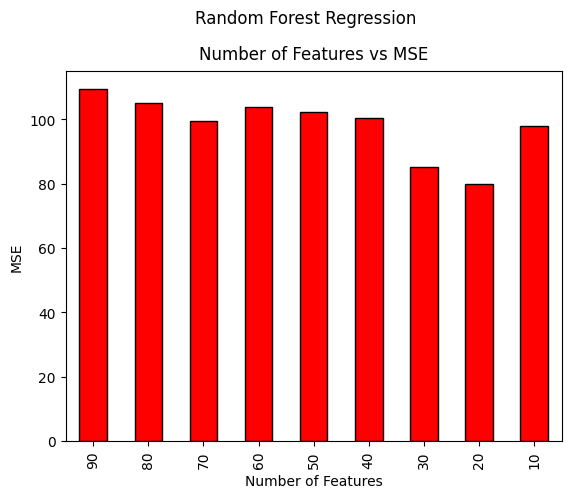
As shown in Table 1 and 2, both Linear Regression and Random Forest models achieved promising accuracy in age prediction.The Linear Regression model generally outperformed the Random Forest model exhibiting lower MSE and RMSE values and higher R2 scores across different feature subsets. Notably, reducing the number of features initially improved model performance, suggesting that a smaller set of informative CpG sites can effectively predict age

**Table 1: Model evaluation metrics for different numbers of features for a Linear Regression Model**

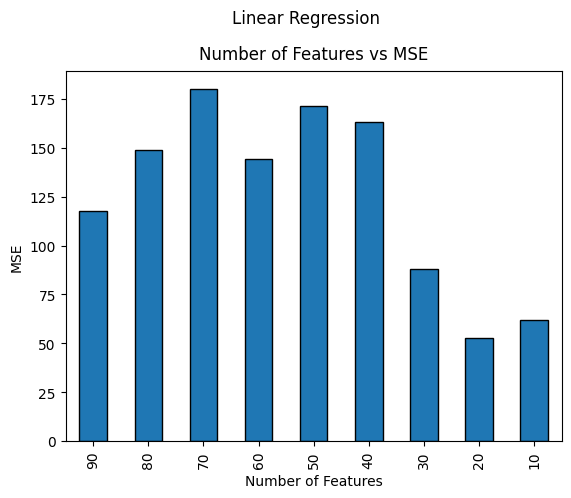
| **No. of Features** | **MSE** | **RMSE** | **R2 Score** |
| --- | --- | --- | --- |
| **7** | 20 | 52.909648 | 7.273902 | 0.931096 |
| **8** | 10 | 62.006748 | 7.874436 | 0.919248 |
| **6** | 30 | 87.891747 | 9.375060 | 0.885538 |
| **0** | 90 | 117.635060 | 10.845970 | 0.846804 |
| **3** | 60 | 144.468852 | 12.019520 | 0.811858 |
| **1** | 80 | 148.698477 | 12.194198 | 0.806350 |
| **5** | 40 | 163.303429 | 12.779023 | 0.787330 |
| **4** | 50 | 171.109621 | 13.080888 | 0.777164 |
| **2** | 70 | 179.987876 | 13.415956 | 0.765601 |

**Table 2: model evaluation metrics for different numbers of features for a Random Forest Regression Model**

| **No. of Features** | **MSE** | **RMSE** | **R2 Score** |
| --- | --- | --- | --- |
| **7** | 20 | 79.978163 | 8.943051 | 0.895844 |
| **6** | 30 | 85.214262 | 9.231157 | 0.889025 |
| **8** | 10 | 97.948725 | 9.896905 | 0.872441 |
| **2** | 70 | 99.638813 | 9.981924 | 0.870240 |
| **5** | 40 | 100.468819 | 10.023414 | 0.869159 |
| **4** | 50 | 102.241737 | 10.111466 | 0.866850 |
| **3** | 60 | 103.913737 | 10.193809 | 0.864673 |
| **1** | 80 | 105.207913 | 10.257091 | 0.862988 |
| **0** | 90 | 109.411019 | 10.459972 | 0.857514 |



**Fig 1: A Bar plot showing Random Forest Model Performance: Number of Features vs. MSE**



**Fig 2: A Bar plot showing Linear Regression Model Performance: Number of Features vs. MSE**

# DISCUSSION

This study investigated the accuracy of age prediction using DNA methylation data from 1000 CpG sites, the performance of different regression models, and the impact of feature selection on model accuracy. Our findings revealed that both Linear Regression and Random Forest models achieved promising accuracy in age prediction, with The Linear Regression model generally outperformed the Random Forest model. Notably, we observed a potential trade-off between model simplicity and accuracy, with reducing the number of features initially improving performance before reaching an optimal point beyond which further reduction led to decreased accuracy. This suggests that a smaller subset of informative CpG sites can effectively capture age-related changes in methylation patterns, but excessive feature reduction can limit the model's ability to capture the complexity of aging. Overall, our results support the feasibility of using DNA methylation data for accurate age prediction, with careful feature selection playing a crucial role in optimizing model performance.

# CONCLUSION

This study reveals key insight into using machine learning models like linear regression and random forest model for age prediction. These findings hold potential for various applications in aging research, forensic science, and personalized medicine

# REFERENCES

Dziechciaż, M., & Filip, R. (2014). Biological psychological and social determinants of old age: bio-psycho-social aspects of human aging. *Annals of agricultural and environmental medicine : AAEM*, *21*(4), 835–838. https://doi.org/10.5604/12321966.1129943

Simpson, D. J., & Chandra, T. (2021). Epigenetic age prediction. *Aging cell*, *20*(9), e13452. https://doi.org/10.1111/acel.13452

Li, A., Koch, Z., & Ideker, T. (2022). Epigenetic aging: Biological age prediction and informing a mechanistic theory of aging. *Journal of internal medicine*, *292*(5), 733–744. <https://doi.org/10.1111/joim.13533>

Ni, X., Zhao, H., Li, R., Su, H., Jiao, J., Yang, Z., Lv, Y., Pang, G., Sun, M., Hu, C., & Yuan, H. (2023). Development of a model for the prediction of biological age. *Computer methods and programs in biomedicine*, *240*, 107686. https://doi.org/10.1016/j.cmpb.2023.107686