**CS 760: Machine Learning** Spring 2022

# Modeling biological age speedup factor ($bio^{age}/chron^{age}$) as random variable to predict death

AUTHORS: Mohil Patel, Sruthi Ganesh, Lipika Garg

**DO NOT POLLUTE!** AVOID PRINTING, OR PRINT 2-SIDED MULTIPAGE.

This project focuses on the analysis of Opportunistic Cardiometabolic Screening dataset [3] provided by Dr. Perry J Pickhardt. The goal of the project is to utilize the clinical and CT data provided in the dataset to predict clinical outcomes such as estimated days to death and current biological age. For solving this problem, two approaches were implemented. In the first approach, we predict age at death. Using the death prediction and assuming the biological age at death is constant, the current biological age is calculated. The data processing and model implementation was done using different ML techniques like K-Nearest Neighbours, Decision Trees, Linear Regression and Neural Networks. In the second approach, the CT data is characterized into biological age based on the assumption that people at a certain age will have specific biomarkers. Using the current biological and chronological age values, the estimated days to death is predicted. Finally, the accuracy of the model was calculated using a custom loss function.

## 1 Introduction

There has been a growing use of machine learning applications in the field of medicine. The two areas of medicine which predominantly benefit from ML are diagnostics and outcome prediction. ML techniques aid in diagnosing diseases in many ways such as medical image segmentation, neurological disease diagnosis from fMRI images, computer-aided diagnosis and detection etc. Some of the notable diagnostic applications of ML are detecting Alzheimer's and Parkinson's disease [2] using MRI images, and breast cancer detection using mammography images [1]. Machine learning has also played a very important role in the early predictions of medical conditions such as heart attacks and diabetes. Recent work has been focused on using CT images for outcome prediction . Some of the examples are: using biomarkers to predict cardiovascular events and death [6], the Framingham Heart Study [5], screening for Osteoporosis Using Abdominal CT [4] and many more.

CT scans provide a lot of valuable biometric information and are taken frequently for a variety of reasons, which makes them very useful dataset for diagnostic applications. Every abdominal CT scan contains additional data that can be objectively measured, including vascular calcification, muscle mass and density, visceral and subcutaneous fat, liver fat content, and bone mineral density information[3]. This paper leverages the predictive ability of the known CT biomarkers to predict outcomes and compares the result with well established clinical parameters such as Framingham Risk Score and Body Mass Index. The two outcomes that the paper focuses on is the prediction of current biological age and age at death.

# 2 Dataset

The dataset was made available by Dr Perry J Pickhardt, Department of Radiology, University of Wisconsin-Madison. The dataset comprises of clinical data, clinical outcomes, and computerized tomography data for 9223 asymptomatic adults between the ages of 23-95, who got abdominal CT for the purpose of colorectal cancer prevention and screening [3]. Through longitudinal follow-up, subsequent adverse events were defined and are listed as clinical outcomes.

The CT image data was processed in 2 ways: Deep-learning and image processing algorithms that were previously trained, tested and validated on a different CT database. These algorithms automatically segmented and quantified different CT parameters like aortic calcium, liver fat, bone measurements, fat measures and muscle measurements. Deep-learning models consisting of a modified 3D U-Net for segmentation of liver and muscle, and Mask RCNN algorithm for segmentation of aortic calcium were deployed [3]. For bone and fat quantification, feature based image-processing algorithms were used, starting with fully automated spine segmentation and labelling software to identify each vertebral level.

# 3 Methodology

This section explains the details regarding the mathematical modeling, assumptions, loss functions, approaches and data preprocessing steps that were taken for generating the results.

We first explain the mathematical model, which covers the intuition behind the models that we propose, followed by formalizing it mathematically. Following that, the loss function is defined for assessing the accuracy of the ML methods that were used. The Approaches subsection covers two fundamental approaches that we considered. And lastly, we cover the details regarding the data preprocessing.

## 3.1 Mathematical Model

The intuition of the mathematical model is explained by the figure 1a. As shown in the figure, ideally we expect that for a given age (in example $age = 27$), the value of $r = \frac{bio^{age}}{chron^{age}}$ is distributed around mean value of 1. That is, we have approximately equal number of unhealthy and healthy individuals, given sufficient data points. This assumption can be extended to all the age values. This leads to us saying, $E[r] = 1$.

To capture this intuition we model $r$ as a random variable. So,

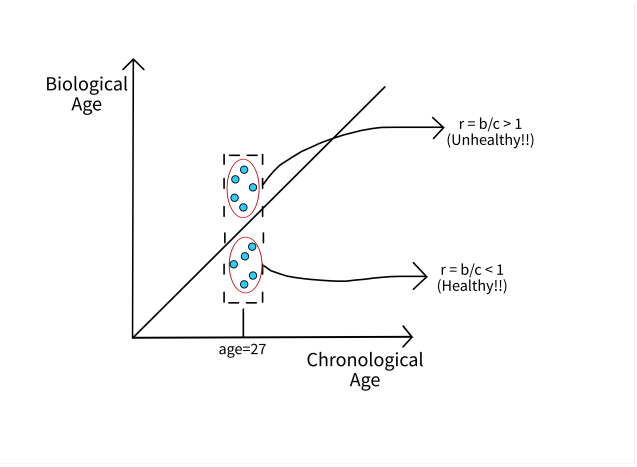
$$r = \frac{bio^{age}}{chron^{age}}$$
$$r = N(1, \sigma^2)$$

**Assumptions**: The following assumptions were made after defining the problem.

1. For every patient we only have one CT value. Which implies we can only derive $r^{CT}$, but $r^{death}$ is unknown. The value of $r^{death}$ is needed to predict the death age. So we make the following assumption:

$$r^{CT} = r^{death}$$

2. The value of biological age at death is also an unknown parameter which is needed to calculate $r^{death}$. To get the value of biological age at death, we assume that **biological age at death is same for everyone**, i.e., everyone dies at same biological age.




(a) Biological vs chronological age intuition

In the dataset, there are 549 points with death labels, which we use to get the value of $bio^{age}_{death}$. For these points we do the following:

$$E[r] = 1$$
$$\frac{\Sigma(r)}{N} = 1$$
$$bio^{age}_{death} = harmonic_{mean}(chron^{age}_{death})$$

Using the data, we calculate the $bio^{age}_{death}$ $as$ 69.35,

## 3.2 Loss Function

A loss function was modelled to evaluate the performance of the ML techniques used.

**Attempt-1**: We already know that $E[r] = 1$. We start with the simplest loss function definition as:

$$Loss = (1 - E[r])^2$$

Even though this loss function fits with the mathematical modeling, potential issues were encountered. Suppose there exist a scenario where for some age we have $r > 1$ for all the points, and for another age we have $r < 1$. Because we are taking $E[r]$ over all ages together, we can still get the value of $E[r]$ closer to 1. In this case, the loss function will determine the ML model as good, but it can be clearly seen that it does not match the base intuition. Ideally, we want $E[r]$ closer to 1 for each age individually.

**Loss Function**: Considering the above we redefine the loss function as:

$$Loss = \Sigma_{age}[(1 - E[r_{age}]^2] \tag{1}$$

Here, we calculate the $E[r]$ individually for each age. This loss is optimized to have $E[r] = 1$ for each age individually. Thus, it avoids the issue that exists in the first function.

The latter loss function was used for evaluating the ML models that were implemented.

### 3.3 Approach

After defining all the above base, we have the following equation:

$$r^{CT} = r^{death}$$

$$\frac{bio_{CT}^{age}}{chron_{CT}^{age}} = \frac{bio_{death}^{age}}{chron_{death}^{age}} \tag{2}$$

In the above equation, we already know the values of $chron_{CT}^{age}$ (given in dataset) and $bio_{death}^{age}$ (using assumption-2).

This leads to two **unknowns**. First is $bio_{CT}^{age}$ and the second is $chron_{death}^{age}$. As long as we can find one of these (using ML models), we can get the other value (using the above equation) which in turn leads two approaches. First approach is to define an ML model for calculating $chron_{death}^{age}$, and use the predictions to calculate $bio_{CT}^{age}$. And in the second model, we define $bio_{CT}^{age}$ using ML model and use it to calculate $chron_{death}^{age}$.

#### 3.3.1 Approach-1

In approach-1 we use the following methodology:

1. Define an ML model to predict $chron_{death}^{age}$. We have 549 data points with death label. We use them to train ML models to predict the value of $chron_{death}^{age}$.
2. Use equation 2 to calculate the $bio_{CT}^{age}$.
3. Assess the model using the loss function from equation 1, using $r = \frac{bio_{CT}^{age}}{chron_{CT}^{age}}$.

This approach comes with the limitation that only 549 data points available for training. We cannot achieve good accuracy for the prediction. It is also reflected in the results, as the loss score is generally higher for this approach.

#### 3.3.2 Approach-2

In approach-2, we use the following methodology:

1. Define $bio_{CT}^{age}$ using ML model.
2. Use $r = \frac{bio_{CT}^{age}}{chron_{CT}^{age}}$ and loss function 1, to assess how good the model is.
3. Lastly calculate the $chron_{death}^{age}$ using equation 2 to get the death age.

This approach does not have the same limitation as the previous approach, because here we have 9223 data points to calculate the value of $bio_{CT}^{age}$. This ensures that we have better predictions of $bio_{CT}^{age}$. This is also reflected in the results in terms of loss function.

### 3.4 Data Preprocessing

Multiple steps of preprocessing were done on the dataset including normalizing, encoding categorial variables and dealing with empty cells. Column F (BMI> 30) was dropped from the clinical data, since column E indicated BMI, thus making F redundant. Categorical data were encoded as positive integers and excluded 0 (taking linear regression into consideration).

- Column G (Sex): Female = 1, Male = 2 and Unknown = 3
- Column I (Tobacco): No = 1, Yes = 2, Unknown = 3
- Column J (Alcohol Abuse): "Some Str" - Yes = 1, "No Str" - Unknown = 2
- Column K (FRS 10-year resik(%)): "X" = NaN, "< 1%" = 0.001 (0.1%), "> 30%" = 0.5 (50%), else the default values were taken for "< 1%" and "> 30%" cases
- Column L (FRAX 10y Fx Prob): "-" = NaN, rest were converted to float
- Column M (FRAX 10y Hip Fx Prob): "-" = NaN, rest converted to float
- Column N (Met Sx): "N" = 1, "Y" = 2, Unknown = "3"

Different approaches were used to process empty cells in the dataset including:

- KNN algorithm was implemented taking only CT values into consideration for filling out the missing cells. Clinical data was excluded due to high run-time of the model.
- The empty cell was filled in with the average taken over the particular column.
- The datapoints with empty cells were dropped from the dataset.

## 4 Results

### 4.1 Approach-1

In this section we have detailed the all the results for approach-1. For each model, we display two plots depicting biological age vs chronological age. First plot shows prediction results using only CT data and the second plot combines CT and Clinical Data for its predictions. We have also reported the loss score for both the predictions.

**Linear Regression**: From Figure 2, we observe that using linear regression does not result in a good loss score. Additionally, we see that the loss score worsens after combining clinical data with CT data.

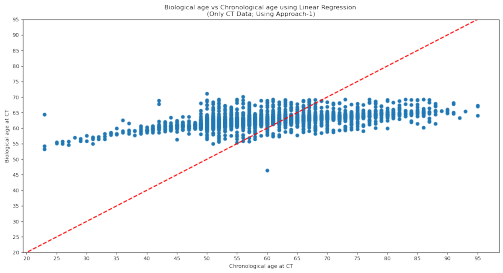
**Regression Decision Trees**: Figure 3 shows the result for decision trees. Similar to linear regression we see a degradation in loss score with addition of Clinical Data. Overall the loss score is still as bad as in the case of linear regression.

**Neural Networks**: We tried using another approach which was creating neural networks. However, even with neural network we see that there is no improvement in the results as shown in Figure 4

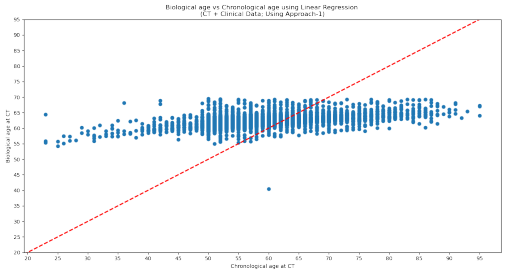
### 4.2 Approach-2

Similar to approach-1, we display two plots for biological age vs chronological for approach-2. We also calculate and report loss score for all the models.

**KNN**: Figure 5 shows the results using KNN for approach 2. There is visible improvement in the results for approach 2 as compared to approach 1. Contrary to approach 1, there is an improvement in the loss score when we add Clinical Data to the model.

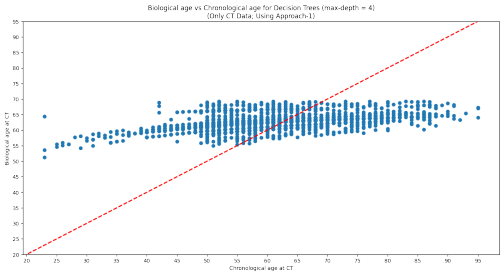



(a) Loss = 15.75 (CT Only)

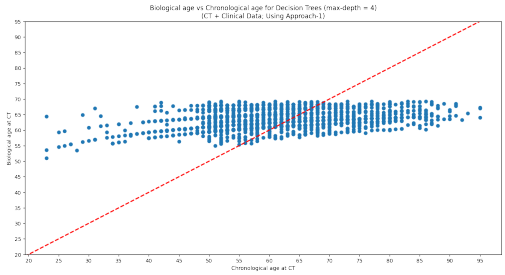



(b) Loss = 16.75 (CT + Clinical Data)

Figure 2: Linear Regression using approach-1

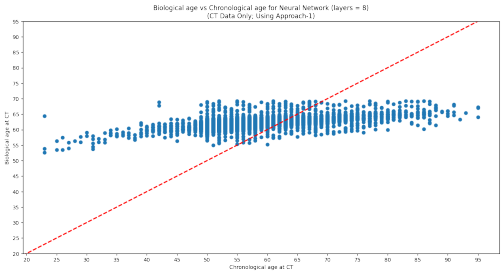



(a) Loss = 15.7 (CT Only)

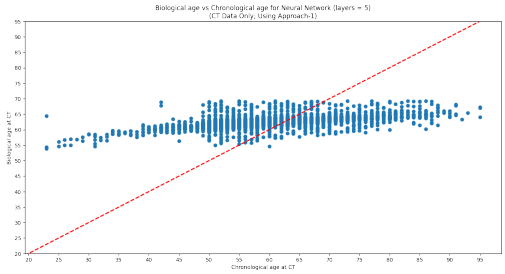



(b) Loss = 16.84 (CT + Clinical Data)

Figure 3: Decision Trees using approach-1

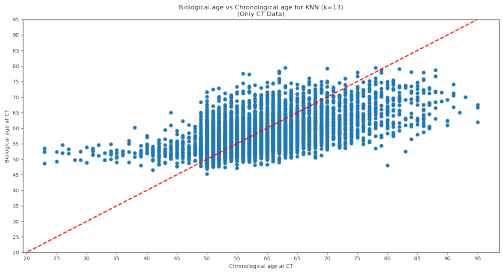



(a) Loss = 15.65 (CT Only)

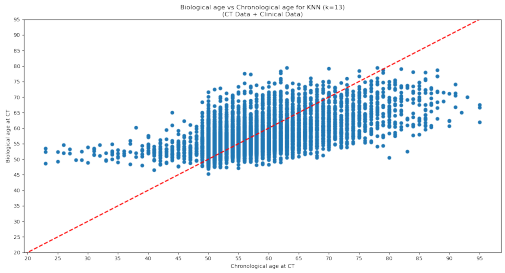



(b) Loss = 16.13 (CT + Clinical Data)

Figure 4: Neural Network using approach-1

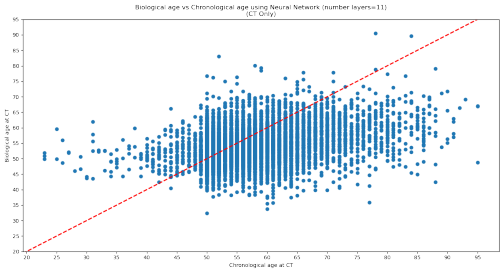



(a) Loss = 10.16 (CT Only)

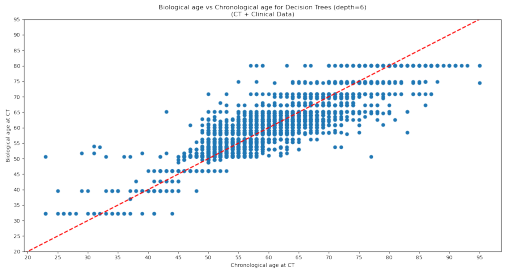



(b) Loss = 10.15 (CT + Clinical Data)

Figure 5: KNN approach-2

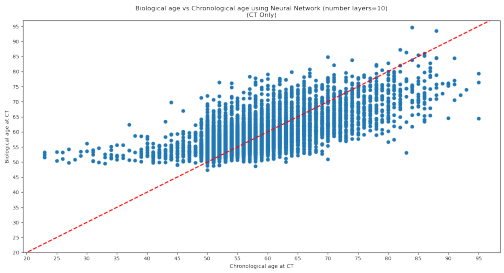



(a) Loss = 8.47 (CT Only)

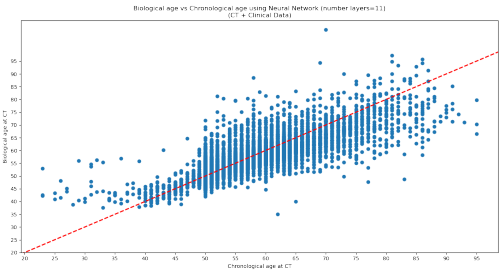



(b) Loss = 1.62 (CT + Clinical Data)

Figure 6: Regression Decision Trees approach-2

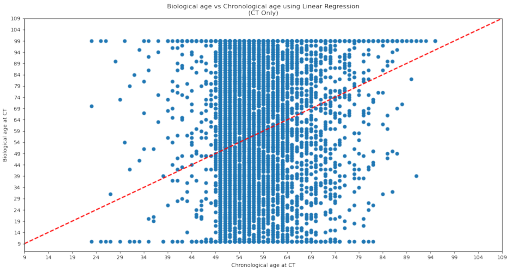



(a) Loss = 10.29 (CT Only)

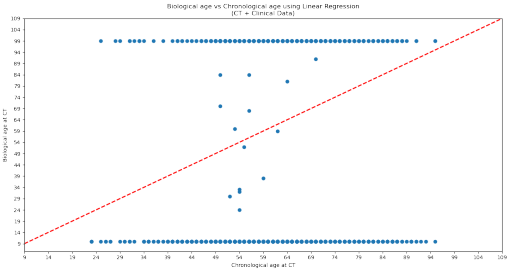



(b) Loss = 4.59 (CT + Clinical Data)

Figure 7: Neural Network approach-2




(a) Loss = 20.36 (CT Only)




(b) Loss = 34.97 (CT + Clinical Data)

Figure 8: Neural Network approach-2

1

**Regression Decision Trees**: Using decision trees gave the best loss score as shown in Figure 6. There is also a huge improvement with the addition of Clinical Data.

**Neural Networks**: The results obtained via Neural Networks using only the CT data are not that good as shown in figure 8. But with the addition of Clinical Data, results have improved significantly.

**Lookup Table using Linear Regression**: For the last approach that we tried we saw that the loss score was too high. Although idea does seems good, the results obtained don't reflect our hypothesis. Also this result shows that we need to add variance factor in our optimization goal.

# 5 Conclusion & Future Work

In this paper we have defined a mathematical model for $r = \frac{bio^{age}}{chron^{age}}$ and a loss function to assess the accuracy of the ML model. Using the mathematical modeling and the loss function, we have designed two approaches to calculate $bio^{age}_{CT}$ & $chron^{age}_{death}$. For each of the approaches we implemented multiple models and we have found that regression decision trees with approach-2 works the best. Lastly, we have also found that adding Clinical Data leads to improving accuracy for all the methods in approach-2, but same is not true for approach-1.

For future works we have few ideas, first is to incorporate variance factor in the loss score. Second is to relax the assumption 1, i.e. $r^{CT} = r^{death}$. This can be done if we have more than 1 CT value for each patient. We can model $r$ as a random process or a time series variable to predict the value of $r^{death}$. Lastly, we can develop an estimator for variance and use it to estimate the confidence interval.

# References


[1] Dromain, C., Boyer, B., Ferre, R., Canale, S., Delaloge, S., and Balleyguier, C. Computed-aided diagnosis (cad) in the detection of breast cancer. *European journal of radiology 82*, 3 (2013), 417–423.

[2] Noor, M. B. T., Zenia, N. Z., Kaiser, M. S., Mamun, S. A., and Mahmud, M. Application of deep learning in detecting neurological disorders from magnetic resonance images: a survey on the detection of alzheimer's disease, parkinson's disease and schizophrenia. *Brain informatics 7*, 1 (2020), 1–21.

[3] Pickhardt, P. J., Graffy, P. M., Zea, R., Lee, S. J., Liu, J., Sandfort, V., and Summers, R. M. Automated ct biomarkers for opportunistic prediction of future cardiovascular events and mortality in an asymptomatic screening population: a retrospective cohort study. *The Lancet Digital Health 2*, 4 (2020), e192–e200.

[4] Pickhardt, P. J., Pooler, B. D., Lauder, T., del Rio, A. M., Bruce, R. J., and Binkley, N. Opportunistic screening for osteoporosis using abdominal computed tomography scans obtained for other indications. *Annals of internal medicine 158*, 8 (2013), 588–595.

[5] Shah, R. V., Ashish, S. Y., Murthy, V. L., Massaro, J. M., D'Agostino, R., Freedman, J. E., Long, M. T., Fox, C. S., Das, S., Benjamin, E. J., et al. Association of multiorgan computed tomographic phenomap with adverse cardiovascular health outcomes: the framingham heart study. *JAMA cardiology 2*, 11 (2017), 1236–1246.

[6] Wang, T. J., Gona, P., Larson, M. G., Tofler, G. H., Levy, D., Newton-Cheh, C., Jacques, P. F., Rifai, N., Selhub, J., Robins, S. J., et al. Multiple biomarkers for the prediction of first major cardiovascular events and death. *New England Journal of Medicine 355*, 25 (2006), 2631–2639.