CS 760: Machine Learning Spring 2022

Prediction of Clinical Outcomes and Biological Age

AUTHORS: Hokeun Cha, Woojin Kim, YoonChae Na

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

In this project, we predict clinical outcomes using CT data and clinical data and derive our biological age. We use 3 modeling approaches to predict clinical outcomes, using regressions, using regressions with balanced training data, and using classification and regression with balanced training data. Although regression achieves high accuracy for alive test data, it fails to correctly predict the dead due to high data skew. We resolve the skew problem by balancing the ratio of the two, improving the accuracy for dead test data by up to 26.67%. We classify data into groups by considering data discrepancy among different ages, which further improves the accuracy for alive test data by up to 20.37%. By additionally utilizing clinical data for training, our best approach achieves 72.4% of accuracy. We define our biological age using an individual's chronological age, average life expectancy, and predicted life expectancy. We comprehensively verify our biological age based on the definition and model prediction, and conclude that it is valid.

1 Introduction

Machine learning has brought much convenience to our life by making useful predictions and observations by learning data. It is one of the fastest growing area in artificial intelligence research. Thus, it is gaining a lot of attention and interests from various fields such as biomedics, statistics, and mathematics due to its broad utilizability.

Such advancement in machine learning technologies has gathered people's attention to their health conditions. There are many methods to see the health of a person, and opportunistic cardiometabolic screening is one of them. That is, people can measure their CT data such as muscle area and clinical data such as BMI. Among various clinical outcomes, the most important one is an individual's life expectancy, which is usually derived by other outcomes such as developing cancers or diabetes.

Biological age is another useful indicator to predict one's life expectancy. It represents how old an individual's body is, while chronological age simply means how old an individual is. Biological age is strongly related to the life expectancy as it can summarize general health condition.

In this project, we predict clinical outcomes using CT data and clinical data by using three models, and derive our biological age. We used opportunistic screening dataset made by Perry Pickhardt, a professor of department of radiology at University of Wisconsin-Madison. We focus on three goals throughout the paper:

- Make prediction of clinical outcomes using CT data.
- Measure how our prediction is improved when utilizing clinical data in addition to CT data.

- Derive and verify our biological age definition.

The paper is organized as follows. Section 2 reviews related work. We discuss our strategies for data preprocessing in Section 3. We present our comprehensive modeling approaches and prediction evaluation results for clinical outcomes in Section 4. Section 5 describes and verifies our definition of biological age, and presents its prediction evaluation. Section 6 concludes our paper.

2 Related Work

In this section, we cover relevant studies in machine learning models and data analyses.

Models. A plethora of machine learning models has been studied to improve prediction results as well as data analyses. Linear regression [9] is a linear approach to model the relationship between a response and a set of variables. Logistic regression [2] is a process of modeling the probability of a discrete outcome given an input variable. Decision tree [10] is a tree-like model that supports making decisions by classifying attributes. K-nearest neighbors (KNN) algorithm [12] is a non-parametric learning method that predicts a response based on the aggregation of the characteristics of nearest neighbors. Bayesian network [3] is a probabilistic graphical model that represents a set of variables and their conditional dependencies via a directed acyclic graph (DAG). Support vector machine (SVM) algorithm [7] is a model to find a hyperplane in a multi-dimensional space that distinctly classifies the data points. Neural network [4] is a series of algorithms that recognizes underlying relationships in a set of data through a process that mimics the way the human brain operates. K-means clustering [5] aims to partition observations into K clusters in which each observation belongs to the cluster with the nearest mean.

Analysis studies. The models mentioned above are actively utilized for predicting responses in various fields such as health care, medical diagnosis, weather condition, and image recognition. Tian et al. [11] presents a recurrent neural network model, developed for health condition prediction of gearboxes based on the vibration data collected from a gearbox experimental system. Hoppner et al. [6] discusses the methods for classification, and image and pattern recognition with fuzzy data. Kononenko et al. [8] provides an overview of the development of intelligent data analysis in medical diagnosis, and presents a comparison of representative systems.

3 Data Preprocessing

In this section, we discuss our strategies for data preprocessing. First, we analyze the columns that contain missing or ambiguous values, and fill out or reinterpret those values with various approaches such as estimating them with regressions and smoothing them out with mean values to minimize unnecessary data drops. We break down the data into three categories; clinical data, clinical outcome, and CT data.

3.1 Clinical Data

We run regressions to estimate the missing values in those columns that have real values, i.e., FRS 10-year risk (%), FRAX 10y Fx Prob (Orange-w/ DXA), and FRAX 10y Hip Fx Prob (Orange-w/ DXA). For the fairness of prediction, the missing value of the clinical data is predicted by only other existing clinical data. i.e. FRS 10-year risk = 5.038(Sex)+1.949(tobaco)+0.29(age)

We put additional categories for those columns that have categorical values, i.e., BMI, Tobacco, and Met

Sx. For Alcohol abuse, we reinterpret the values into a binary variable as missing values dominate, and the occasionally observed abuse behavior is excessively fragmented.

Finally, we exclude highly correlated columns to minimize regression analysis errors. We use both clinical and CT data to analyze correlation between columns. Among the candidates, BMI and Total Body form high correlation with value of 0.88. Therefore, we remove Total Body in our dataset.

3.2 Clinical Outcome

The majority of values in clinical outcomes are missing. As the ratio of the missing data and the given ones is highly skewed to the former, we cannot estimate missing values based on the latter. Therefore, we narrow down our focus to specific columns, i.e., Death [d from CT], Heart Failure DX, and Type 2 Diabetes.

First, we install an additional binary column that classifies death based on the existence of the values in column Death [d from CT]. Then, we fill in the missing values in the column with the average life expectancy of each person based on one's age and sex [1].

We also create binary columns for columns Heart Failure DX and Type 2 Diabetes, and further create their categorical columns for which value maps to each disease behavior.

3.3 CT Data

We also exclude specific columns in CT data which are highly correlated with another columns as done in clinical data. We remove VAT in our dataset as it is highly correlated with Total Body (0.88), and Muscle are as it is highly correlated with L3 SMI (0.89).

4 Prediction of Clinical Outcomes

In this section, we describe our modeling approaches and evaluation results to predict clinical outcomes using CT data, and using CT and clinical data.

4.1 Modeling Approaches (CT data)

We develop our approach in three steps, using a simple regression, using regressions with balanced training data, and using regressions after classifying the data into groups with balanced training data. For each approach, we incrementally add features.

Table 1: Prediction accuracy for clinical outcomes

Models	**Test with dead**	**Test with alive**	**50% mixed**
Regression	5.33%	46.69%	26.01%
Balanced training data	32.0%	20.14%	26.07%
Classification and regression	37.5%	40.51%	39.1%

Table 1 summarizes the prediction accuracy for clinical outcomes for our three approaches with different test sets. We use 75% of data for training and the rest for validation. For the validation, we consider a prediction

is correct if the difference of prediction and given response is within 3 years. We note that we only show the results with the best learning models, i.e., neural network, due to limited space [1].

Table 2: Prediction accuracy of regressions

Models	Test with dead	Test with alive	50% mixed
Linear regression	5.33%	44.53%	24.93%
KNN regression	1.33%	45.94%	23.64%
Neural network	5.33%	46.69%	26.01%

4.1.1 Regression

We first use three regression models to predict clinical outcome, LinearRegression, KNeighborsRegressor (KNN), and MLPRegressor (Neural Network) from scikit-learn. Table 2 shows the accuracy for each regression. While all the regressions show relatively high accuracy, they report poor performance when testing with dead data, i.e., a testset that only consists of values given in column Death d from CT. This is because only 5% of our total dataset (494/8,887) are the dead data. As the opposite case dominates, the model makes highly skewed predictions to the alive data.

4.1.2 Balanced training data

To resolve the skew problem, we balance the ratio of dead and alive data in training data. Table 1 shows the performance improvement from the previous to current approach. A balanced mix of the dead and alive data in training data improves the accuracy with dead test data, efficiently resolving the data skew, while that with alive test data decreases.

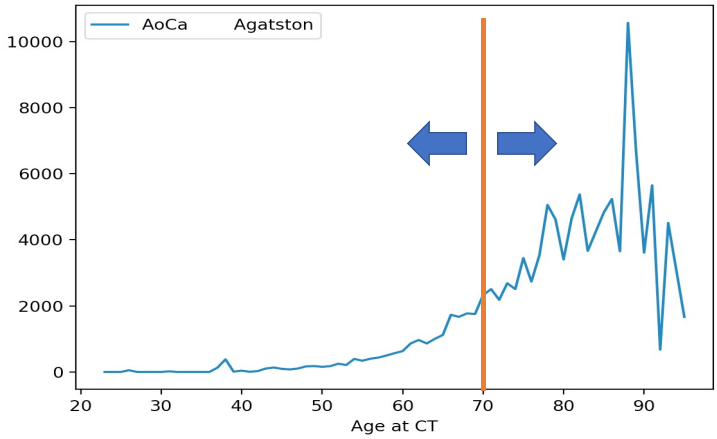


Figure 1: Data discrepancy in age distribution

[1] All the implemented models covering various classifiers and regressors are provided in our source codes

4.1.3 Classification and Regressions

We further analyze the data based on the distribution of age, and found out there is discrepancy between different ages, as shown in Figure 1. That is, the value of AoCA Agaston increases in proportion to Age at CT within 70, but the observed values after then fluctuate, creating an irregular pattern.

To minimize the effect of the discrepancy, we classify the data into two groups, young and old, and run regression for each group. We study our classification results in Section 4.1.4. Table 1 reports the performance improvements between the previous and current increment. Not only the accuracy with dead test data increases, but also that with alive test data improves by a large margin.

Table 3: Prediction accuracy of classifications

Models	**Test with dead**	**Test with alive**	**50% mixed**
Classification	16.22%	93.69%	54.96%
Balanced training data	64.83%	84.23%	74.53%
Augmented data	65.48%	85.81%	75.65%

4.1.4 Classification

We apply the same approach for classification as regressions stated above. Table 3 shows the performance of stand-alone classification, classification with balanced training data, and augmented data (synthetically replicating dead data). It follows the same performance trend as observed in the Section 4.1. Balanced ratio of the dead and alive data improves the accuracy for dead test data, and data augmentation further increases the performance of both tests. We note that we use KNeighborsClassifier (KNN), GaussianNB (Naive Bayes), SGDClassifier (SVM), and MLPClassifier (Neural Network) in scikit-learn as our classifier models, but we only provide the results with MLPClassifier as using different classifiers only has small impact on the accuracy.

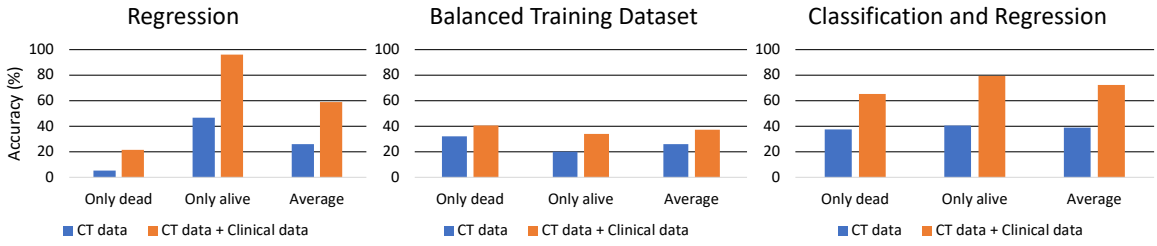
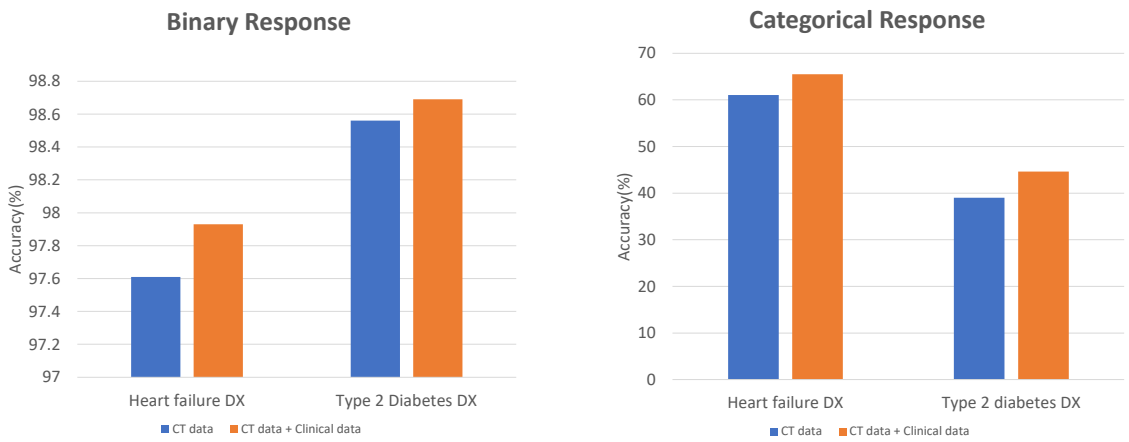


Figure 2: Accuracy comparisons with models using CT data, and models using both CT and clinical data

4.2 Model Extension: using both Clinical and CT data

We now apply our modeling approaches to further analyze the performance impact when clinical data are given. Figure 2 shows the accuracy comparisons between the models that use CT data, and the models that use both CT and clinical data for each approach. The performance of those models using both CT and clinical data dominate that using CT data in every test case.



((a)) Binary response

((b)) Categorical response

Figure 3: Accuracy comparisons with models using CT data, and models using both CT and clinical data for Heart Failure DX and Type 2 Diabetes DX prediction

4.3 Model Extension: additional clinical outcomes

Now we further dive into predicting additional clinical outcomes other than Death [d from CT]. We use KNeighborsRegressor to predict a binary response, and a categorical response for Heart Failure DX and Type 2 Diabetes, as shown in Figure 3(a) and (b), respectively. We note that the accuracy in each Figure is in a relative values, not fixing the range from 0 to 100%. From the results, we can infer that both the two models can efficiently predict Heart Failure DX and Type 2 Diabetes DX. It is noteworthy that the accuracy improvement from the model using CT data to the model using both CT and clinical data is not as large as that observed in Section 4.2.

5 Biological Age

In this section, we define our biological age, introduce its observation, and verify the biological age using the definition and observation. We define AGE as the random variable for age and $BAGE$ as the random variable for our biological age. We also define EXP as the random variable for life expectancy, which can be obtained by dividing predicted Death [d from CT] by 365.

5.1 Definition and Observation

We define $BAGE$ as follows:

Definition 1. $BAGE = AGE + \mathbb{E}(EXP|AGE) - EXP$.

A person's $BAGE$ is same as the person's age under the assumption that the person will die when the person's age is $AGE + \mathbb{E}(EXP|AGE)$. As we will prove, $\mathbb{E}(BAGE|AGE) = AGE$. From this fact, we observe that $BAGE$ of a person with average CT data of healthy people with age a is a.

5.2 Verification Using Definition

We define X as $BAGE - AGE$, σ as the standard deviation of X, which is unknown, N as the number of our data, which is 8877, $\bar{X}$ as the sample mean of X for N data, and $\hat{\bar{X}}$ as the mean of X for our data. By the definition of $BAGE$, $X = \mathbb{E}(EXP|AGE) - EXP$. Taking mean on people with age AGE, we can get the following equality:

$$\begin{aligned}\mathbb{E}(X|AGE) &= \mathbb{E}(EXP|AGE) - \mathbb{E}(EXP|AGE)\\ &= 0.\end{aligned}$$

By central limit theorem, we can get the following information on the distribution of $\bar{X}$:

$$\bar{X} \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma^2}{N}\right).$$

To determine whether our biological age is valid, we define the following two hypotheses and select one hypothesis between the two:

$$\begin{aligned}H_0 &: \hat{\bar{X}} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{N}\right), \mu = 0,\\ H_1 &: \hat{\bar{X}} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{N}\right), \mu \neq 0.\end{aligned}$$

We use the following generalized likelihood ratio test where Φ_χ is the tail function of χ^2 distribution:

$$\frac{\max\left\{\mathbb{P}\left(\hat{\bar{X}}|\mu\right) : \mu \neq 0\right\}}{\mathbb{P}\left(\hat{\bar{X}}|\mu = 0\right)} \underset{H_0}{\overset{H_1}{\gtrless}} e^{\frac{1}{2}\Phi_\chi^{-1}(0.05)}.$$

Computing the left hand side of the above inequality, we can rewrite the test as follows:

$$\left(\frac{\hat{\bar{X}}}{\frac{\sigma}{\sqrt{N}}}\right)^2 \underset{H_0}{\overset{H_1}{\gtrless}} \Phi_\chi^{-1}(0.05).$$

$\hat{\bar{X}} = 57.41 - 56.91 = 0.5$, and $N = 8877$. Hence, $\left(\frac{\hat{\bar{X}}}{\frac{\sigma}{\sqrt{N}}}\right)^2 \leq \Phi_\chi^{-1}(0.05)$ if and only if $\sigma \geq 24.04$. Considering the definition for X, 24.04 is relatively high value as σ, but it is reasonable that $\sigma = 24.04$, so we conclude that our biological age is fairly valid using the definition for our biological age.

5.3 Verification Using Observation

As we create models predicting clinical outcome using CT data, we can also create models predicting biological age using CT data. Based on the observation for our biological age, we guess that predicted biological age of a person with average CT data of healthy people with age a is near to a. For this reason, we also verify our biological age using the models predicting biological age using CT data.

We create 2 models predicting biological age using CT data using the methodology of the models predicting clinical outcome using CT data. The first model is the model predicting biological age using regression. The second model is the model classifying data as data in old group and data in young group and predicting biological age using regression for each group. We use KNeighborsRegressor, MLPRegressor, and GaussianNB in scikit-learn as our regressors. Biological age can be considered as discrete variable, so we use GaussianNB

as one of regressors. We use KNeighborsClassifier, MLPClassifier, and GaussianNB in scikit-learn as our classifiers. We only provide the results for KNeighborsClassifier because classifiers have almost no effect on the accuracy of the models.

We use all the tuples in CT data for training our models. The test data are the average CT data of healthy people with age a for all a. We define a healthy person as a person who is predicted to live after 3 years, so we exclude data who are predicted to die in 3 years when we make test data. We consider prediction is correct if the difference of predicted biological age and age is not bigger than 3 years. We believe that predicting precise biological age is as hard as predicting precise death dates, and if the models are correct, and our biological age is valid, predicted biological age is near from age for test data, so we define correct prediction in this way.

Table 4 shows the prediction accuracy of the models for the case that regressor is KNeighborsRegressor, MLPRegressor, and GaussianNB, and classifier is KNeighborsClassifier. Accuracy of the second model is higher than that of the first model, accuracy of the model with GaussianNB regressor is higher than that of the models with the other regressors, and the highest accuracy is attained as 63.64%. This accuracy is fairly high, so we conclude that our biological age is fairly valid using the observation for our biological age.

Table 4: Accuracy of models predicting biological age

Models	**KNN regression**	**Neural network**	**Gaussian naive Bayes**
Regression	27.27%	25.45%	50.91%
Classification and regression	30.91%	29.09%	63.64%

6 Conclusion

In this work, we predict clinical outcomes using CT data and clinical data. We utilize different machine learning models from classifiers and regressors. Based on our extensive data analysis, we present three modeling approaches, regression, regression with balanced training data, and classification and regression with balanced training data. Our evaluation study shows each incremental brings huge performance gain. We further extend our models to predict additional clinical outcomes such as diabetes and heart failure. We define our biological age based on an individual's chronological age, average life expectancy and predicted life expectancy. Our verification with definition and modeling observation concludes that the biological age is valid.

In the future, we want to improve models using more training data. If we have more training data of young people, we can classify data into more than 2 groups, so regression for each group will be improved. Also, if we have more training data of dead people, more real data will be used as training data, so models become more realistic.

References

[1] S. S. Administration. URL: https://www.ssa.gov/oact/STATS/table4c6.html.

[2] S. Dreiseitl and L. Ohno-Machado. Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics*, 35(5-6):352–359, 2002.

[3] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine learning*, 29(2):131–163, 1997.

[4] M. T. Hagan, H. B. Demuth, and M. Beale. *Neural network design*. PWS Publishing Co., 1997.

[5] J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108, 1979.

[6] F. Höppner, F. Klawonn, R. Kruse, and T. Runkler. *Fuzzy cluster analysis: methods for classification, data analysis and image recognition*. John Wiley & Sons, 1999.

[7] T. Joachims. 11 making large-scale support vector machine learning practical. *Advances in kernel methods: support vector learning*, page 169, 1999.

[8] I. Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*, 23(1):89–109, 2001.

[9] D. C. Montgomery, E. A. Peck, and G. G. Vining. *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

[10] S. R. Safavian and D. Landgrebe. A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, 21(3):660–674, 1991.

[11] Z. Tian and M. J. Zuo. Health condition prediction of gears using a recurrent neural network approach. *IEEE transactions on reliability*, 59(4):700–705, 2010.

[12] M.-L. Zhang and Z.-H. Zhou. Ml-knn: A lazy learning approach to multi-label learning. *Pattern recognition*, 40(7):2038–2048, 2007.