Ensemble Techniques for Prediction of Adverse Clinical Outcomes using CT Biomarkers

Harshith Gundappa, Reetuparna Mukherjee, Sanyam Kakkar

Abstract

A computerized tomography (CT) scan provides vital information about the underlying health condition of a patient. Using CT and clinical data of close to nine thousand patients, we build machine learning models to predict (1) Death, (2) Diabetes, and (3) Biological Age. We propose an ensemble model to accomplish the task of death prediction. In order to determine the death value of patients who are alive, we utilize a KNN-based approach. We compare the alive patient's clinical data (BMI, Tobacco, Diabetes, etc) with dead patients who had similar health indicators and use this result to fill in the death column. Next, we learn to predict this death value from CT data. Six different machine learning models are utilized for the regression task. A Linear Regression layer sits on top of all the models and the ensemble layer learns to assign weights to each of the underlying models' predictions. The ensemble model produces a mean absolute percentage error(MAPE) of 9%, mean absolute error(MAE) of 0.47 years and RMSE score of 1.01 years. The biological age of a person depends chiefly on genetic markers[1]. However, in the absence of such genetic indicators in the dataset, we propose a machine learning-based approach to compute the biological age. By analyzing the dead patients' health records, we train a Polynomial Regression model to learn how many years of life were lost/gained due to the patient's underlying health condition. We make use of the dead patients' clinical data such as BMI, Alcohol Abuse, Diabetes, etc for this task. We then use this model to predict the biological age of patients who are alive. After determining the biological age from clinical data, we deploy Neural Network and Random Forest to learn the relationship between CT data and biological age. These models produce an RMSE error of 8.8 and 7.5 years respectively. We also classify the diabetes condition of the patient using CT data. We train six different machine learning models to accomplish this task and KNN produces the best results with an accuracy score of 67%. Our code can be found at `https://github.com/harshith7823/CS760-Project`

1 Introduction

The search for efficient predictors of lifespan and mortality rate has been an area of interest for the medical research community for a few decades. While chronological age has been the standard, the quest to find factors that influence and hold the ability to better estimate mortality and the rate of aging is an interesting but difficult problem to solve. Advances in various fields of medical science have expanded the list of potential candidate "biomarkers" that can be exploited in this direction. However, the difficulty lies if the fact that the health records available for study usually belong to people with underlying health conditions which might lead to inaccurate assumptions and bias. Given the interdependence in physiological and biological attributes, it is difficult to extract a particular marker for study, in its untainted form.

Most of the literature available is concentrated around blood and genetic markers. There has been little exploration on using radiological attributes to predict mortality. We focus on "Opportunistic Cardiometabolic Screening" dataset from UW-Madison's Department of Radiology and try to exploit data for a purpose different from what it was collected for.For example, using a thoracic X-ray that was collected to treat a broken rib can be used to predict an osteoporotic fracture later on. We also explore the correlation between the various clinical attributes and tomography parameters on the rate of aging, and thereafter try to come up with a novel estimate of biological age and mortality. We aim to use machine learning approaches to derive insights from the data and train a model to effectively predict the biological age and expected lifespan of an individual, given his/her tomography records. We identify that there is a scope for using the same data to predict if a patient has a type of diabetes. Although we tried to study other clinical outcomes that the data could be suggestive of, we do not think the confidence of such models is high enough and therefore omit that from the discussion below.

2 Related work

Baker and Sprott [1] define a biomarker of aging as a biological parameter of an organism, which in the absence of disease, is capable of predicting functional capability at some late age better than chronological age. A biomarker can exhibit this capability either alone or in the form of a multivariate composite.

Later on, the American Federation for Aging Research proposed the following as criteria for a Biomarker of Aging(BoA):
1. It must be able to predict the life span or rate of aging better than chronological age.
2. It must monitor a basic process underlying the aging process, and not the effects of a disease.
3. It must be able to be tested repeatedly without harming the individual involved, for example, an imaging technique, or a blood test.
4. It must be something that works in humans and in laboratory animals so that it can be tested in lab animals before being validated in humans.

The search for a BoA has been an area of continuous research in the biological and medical community owing to its tremendous relevance to the current push to identify drugs that ameliorate the aging processes. However, with the vast diversity of potential candidate biomarkers, and lack of concrete consensus on how accurate a biomarker should be, it is still a difficult problem to solve. Moreover, the fact that biomarkers, by definition, must be able to predict life span in the absence of an underlying disease, makes it all the more difficult given that such data is mostly available from people suffering from one or the other form of ailment.

Predictors for one age group may fail with other groups. Further, the biomarkers may exhibit nonlinearity, as shown by Sagers et al [11]. They used a supervised machine learning approach to identify 356 biomarkers and predicted age in held-out data of 4.76 years with an R2 value of 0.92. While the age prediction with random forests and piecewise linear regression was highly accurate for the pediatric cohort (MAE = 0.87, R2 = 0.94), but inaccurate for ages 65+ (MAE = 4.30, R2 = 0.25).

In a study from 2013, ten biomarkers significantly associated with chronological age (CRP, glycated hemoglobin, serum creatinine, serum albumin, total cholesterol, systolic blood pressure, cytomegalovirus optical density, forced expiratory volume, serum alkaline phosphatase, serum urea nitrogen) were combined into a composite biomarker in an NHANES III study [10]. Using Cox proportional hazards, the composite biomarker predictor was associated with a mortality factor that is independent of chronological age. This model was later validated in the Dunedin study, for a younger birth cohort [2].

Biomarkers related to the cardiovascular system and metabolic processes have also been the focus of study in this field. Markers of the sympathetic nervous system, functioning of organs like kidneys, lungs, and heart, and immunological responses like inflammation activity in the hypothalamic pituitary axis are slowly becoming a part of population measurement. [7] lists an extensive range of biomarkers and hails epigenetic clock and telomere length as the most plausible candidates. They, however, highlight that epigenetic clock is associated with a wide spectrum of aging outcomes and mortality, and not just specific to biological lifespan prediction. Both DNA methylation age estimates were highly correlated with chronological age (r=0.90). Karlamangla et al. [8] examined the relationship between change in allostatic load - a composite risk score derived from ten biological markers, and mortality in a 4.5-year span in a cohort of 171, 70-year-old adults, as a part of the MacArthur studies of successful aging,

In radiology, there is a well-established relationship between 12-lead electrocardiogram (ECG) and age and mortality. Furthermore, there is increasing evidence that ECG can be used to predict biological age. [6] used 438 features from 12-lead ECG in patients without structural heart disease, and reduced it to 26 features with PCA. The models for biological age were developed by principal component analysis (BA) and the Klemera and Doubal's method [9].

While deep neural networks are becoming increasingly popular due to their capability to fit overly complex data, there are concerns when it comes to usage for the prediction of biological age. This is primarily due to the fact that for deep learning to succeed in radiology, well-annotated large data sets are needed since deep networks are complex, computer software and hardware are evolving constantly, and subtle differences in disease states are more difficult to perceive than differences in everyday objects [4].

3 Dataset

The dataset we explore is "Opportunistic Cardiometabolic Screening" made available by the Department of Radiology, UW Madison. This comprises of clinical data and computerized tomography (CT) data from 9223 individuals with diverse clinical records.

4 Approach

We started with data analysis and preprocessing which includes clean-up of the data.
Three clinical outcomes were predicted from the dataset

- the estimated age at death, type of diabetes, and biological age of an individual. We consider the estimation of age at death as a regression problem, while the diabetes prediction problem is considered a multi-class classification problem. The estimation of biological age is also modelled as a regression problem. Figure [1] gives an overview of our methodology for predicting the age at death.

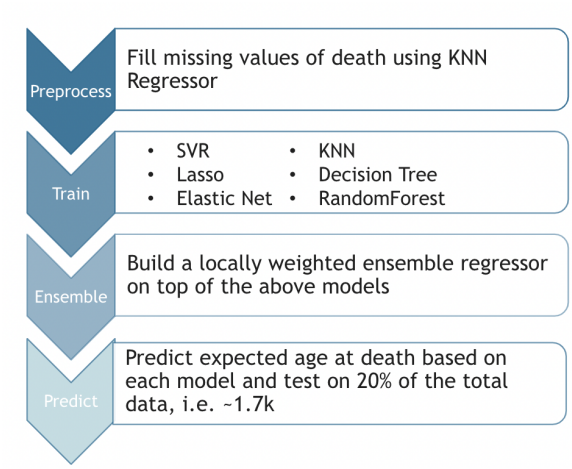

Figure 1: Methodology

4.1 Data Analysis

In order to analyze the data, we started by plotting various distribution plots. We reckon that the data is centered around the age group of 50-60. Exploring the correlation and covariance metrics revealed a few attributes which are quite closely related to mortality.

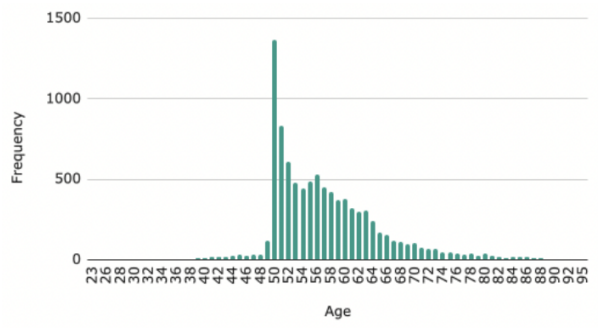

Figure 2: Distribution of age

4.2 Data Preprocessing

The first two steps in data preprocessing were:
1. Clean up: We replaced all the null values in CT data with mean. All the categorical features were converted to numerical features, e.g. BMI greater than 30 was hot encoded as 1 and BMI less than 30 as -1. We treat blanks in "Tobacco" column as no tobacco usage.
2. Dimensionality Analysis: A Principal Component Analysis on the data revealed that 90 percent variance can be captured using 6 components. However, since the number of features are not too many, we proceeded with the entire dataset.

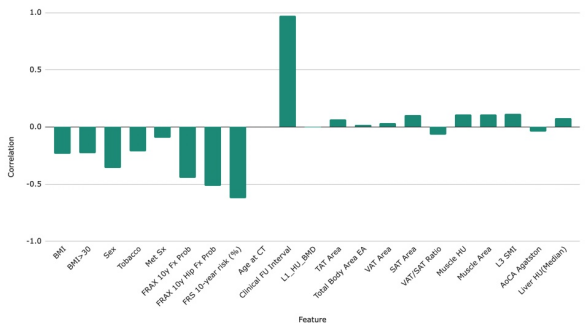

Figure 3: Correlation of features with Death column

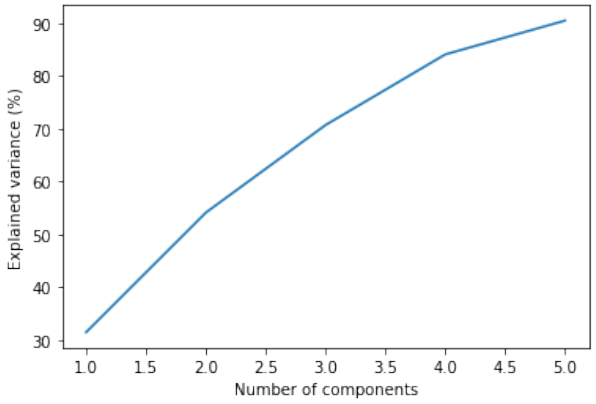

Figure 4: Explained Variance vs Number of component

4.3 Death Prediction

In this part, Death after CT scan was predicted using CT data and then CT and clinical data.This problem was divided into 3 parts.
1. Filling missing values in death column
2. Death prediction from CT Data
3. Death prediction from CT Data and Clinical Data

4.3.1 Filling missing values in Death column

The first problem at hand was that there were only 550 records which contained information about whether the patient has died or not, and if he/she did, how many days after the scan did it happen. For people with missing death values, we assume they are still alive and we try to come up with a way to estimate at what age would they die. The problem of sparse data was solved using two techniques. The first technique is based on coming up with a formula to fill the "Death" column. The underlying intuition in the formula stated below is that BMI and tobacco usage have a significantly detrimental effect on health. We used normalised correlation of these fields with "Death" column. We then calculate the reduction factor for each of these factors, for instance, we reduce 4.2 years from the life expectancy if the individual is an overweight male. Similarly, we reduce 10 years for a person with a history of smoking. These values were decided upon based on [3]. We then combine them into a formula to fill the missing values for estimated age at death.

$$r_{tob} = 9.2$$

$$r_{bmi} = \begin{cases} 4.2 & if\ gender = male \\ 4.4 & otherwise \end{cases} \quad (1)$$

$$Avg_{m/f} = \begin{cases} 76.2 & if\ gender = male \\ 81.2 & otherwise \end{cases} \quad (2)$$

$$\alpha = \frac{corr(bmi, death)}{corr(bmi, death) + corr(tobacco, death)}$$

$$\beta = \frac{corr(tobacco, death)}{corr(bmi, death) + corr(tobacco, death)}$$

$$\Delta = Avg_{m/f} - (r_{bmi} \times \alpha + r_{tob} \times \beta + Age)$$

The second approach was motivated by the fact that people with similar health condition should have similar life expectancy. To this end, we chose K-Nearest Neighbors to pre-fill missing values in the "Death" column. Euclidean distance was used as metric and we narrowed down on the best k value = 17 (lowest RMSE).

A cross validation on the already available records with death information revealed that the second approach performed around 22 per cent better than the first. So we proceed to use this KNN augmented data for further training.

Algorithm 1 Regressor

$D \leftarrow \{a_1, \ ... \ , a_n\}$
$L \leftarrow \{a_i \ \forall i \ if \ D[a_i]['Death'] \ != null\}$
$U \leftarrow D - L$
for $a_i \ in \ U$ **do**
  $(d_1, ..., d_k) \leftarrow kNN(a_i, L)$
  $a_i['Death'] \leftarrow avg(d_1, ..., d_k)$
end for

4.3.2 Death Prediction using CT Data

For predicting the age at death, first we use only CT data which includes measures like Aortic Calcification, Total Adipose Tissue area, Visceral Adipose Tissue area, etc. After cleaning up records with null entries, we obtain a set of 8878 tuples. We use different regression algorithms to train and test on an 80-20 split on these records. As the model complexity (which in the case of linear regression can be thought of as the number of predictors) increases, estimates' variance also increases, but the bias decreases. The unbiased OLS would place us on a region which is far from optimal. So we explored models with L1 and L2 regularization to help us reach the optimum results. Different Regression algorithms like Random Forest Regressor,Linear Regression with regularization (Lasso, ElasticNet),SVM Regressor were fitted on the data. It is evident that Random forest yields the best results with the least RMSE and MAE value. The random forest hyperparameters were tuned using GridSearchCV. Our training setup has another layer involved on top of these regressors. We build an ensemble of all of these models. We tried three types of ensembling - average ensembling, random ensembling and the last one being locally weighted ensemble regression which uses linear regression to assign weights to the underlying models.

Three different ensemble techniques was used to further decrease the error.

4.3.3 Death Prediction using CT and Clinical Data

Adding clinical data to the dataset incorporates features like BMI and tobacco usage which have a high impact on mortality. We used the same techniques as mentioned in 4.3.2 and compare the results obtained in section 5 below.

4.4 Diabetes Classification

Other than death, we also built models to predict the Diabetes condition of a patient using CT data. Rest of the clinical outcomes, with the exception of Diabetes, are very sparse, leading to inadequate data for machine learning purposes. We, therefore, focus only on the diabetes classification task. Out of the total 9K

records in the dataset, 2.5K patient's diabetes condition is recorded. We use this data to train a model to predict the diabetic condition of the patient. The four class labels the models are trained on are Type-1 Diabetes, Type-2 Diabetes, Fasting and Abnormal.

In order to visualize the vector space of the CT data, we run the t-SNE algorithm. Figure[5] shows our results. As shown in the plot, we do not see any clusters in the t-SNE diagram and the various class labels are distributed throughout the vector space. We suspect this to be one of the reasons behind the low accuracy levels of the ML models.

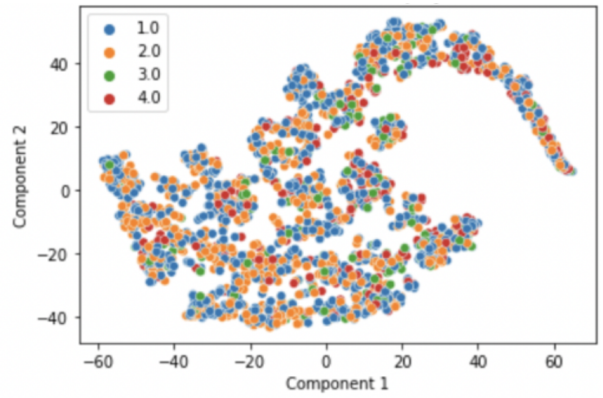

Figure 5: t-SNE visualization of Diabetic patients' clinical data. 1.0, 2.0, 3.0 and 4.0 corresponds to Abnormal, Fasting, Type 1 Diabetes and Type 2 Diabetes class labels, respectively.

As shown in Figure [13], we built six different models to fit the data. KNN gave the best results with an accuracy of 67%. The optimal value for K was found to be 9. The KNN model employs a uniform weight distribution strategy and the training CT data was transformed using Z-score normalisation.

4.5 Biological Age

4.5.1 Determining Patient's Biological Age

In order to compute the biological age of a patient, we make the assumption that when a person with no pre-existing condition dies, their biological age would be the average life expectancy. The first step of the approach involves creating a new dataset with only the dead patients. We measure how many years of life was lost/gained for each of the dead patient. The difference between the person's chronological age at death and the average life expectancy gives this value. We then train a model to learn the relationship between the number of years lost/gained and the rest of the clinical data. The 'X' in the new dataset consists of fields such as BMI, Sex, Tobacco, Alcohol-Abuse and the rest of the clinical data. We also consider clinical outcomes in our X. This is important because clinical outcomes such as Heart Failure, Diabetes, etc play a crucial role in the overall health of the patient. The 'Y' is the number of years lost/gained.

We train a Polynomial Regression Model to fit this data. We got the best results on a degree 3 PR model. Figure[6] shows our results. On a K-fold cross validation set(K=10), we got an average MAE score of 1.2 years. After training on the dead patients' records, we determine the biological age of rest of the patients using the PR model. The output of the PR model would be the number of years the patient might lose/gain given their pre-exising condition. We then add this value to their chronological age. This would be the biological age of the patient.

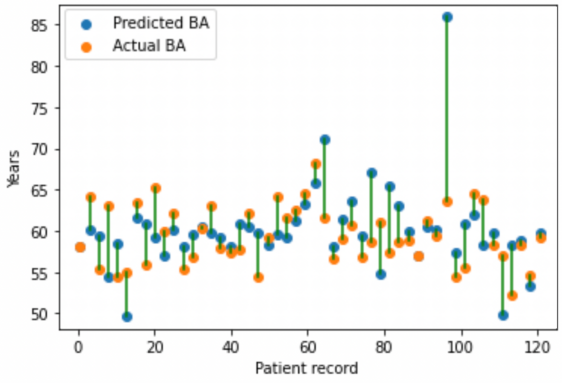

Figure 6: Polynomial Regression Model - Predicted vs. Actual values of Biological Age. The gap between the blue and the orange dots indicates the degree of error.

4.5.2 Biological Age Prediction

After determining the biological age of the patients using clinical data as explained above, we train two different models to predict a person's biological age using CT data. The two models used for this task were :

- Multi-Layered Perceptron(MLP)
- Random Forest Regressor

Our neural network consists of 3 hidden layers with ReLU activations. The dimensions of the hidden layers are as follow: First layer - 64 neurons, second layer - 32 and the third layer - 1. The last layer's output is the

model's predicted value for biological age. We trained the multi-layered perceptron on 100 epochs. Adam optimizer with a learning rate of 0.001 was utilized. The optimizer had a weight decay of 0.25. Batch size used was 64. Training data was normalized as per Z-score normalization and Kaiming He's [5] technique was employed to initialize the weights of the network. The model was tested on 10 per cent of the total data and the results are shown in Figure[15] and Figure[16].

We also utilized Random Forest Regressor to predict the biological age. We got the best results when the maximum depth of the trees was limited to 20. The model employed 300 estimators and the minimum sample leaf size was 6. We used K-fold cross validation to generate the MAE, RMSE and MAPE values. The value of K chosen was 10.

5 Results

We describe the results of our experiments on each of the subproblems here. We tested the performance of our cross-validated models over 50 iterations each on 20 per cent of the data. All the experiments were run on Google Colab.

5.1 Death Prediction with CT Data

Figure [7] shows the performance of the Linear Regression Ensemble Model. We only plot a subset of the test set for clarity. Figure[8] and [9] list out the error metrics for the models. The weighted ensemble technique gave the minimum MAE of 0.47 years.

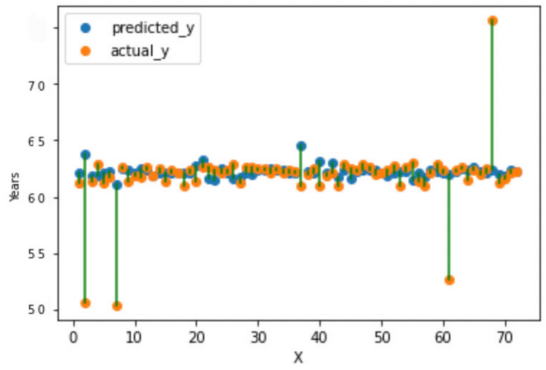

Figure 7: Locally Weighted Ensemble Regression with CT-Data:Predicted vs. Actual values of Death Age. The gap between the blue and the orange dots indicates the degree of error.

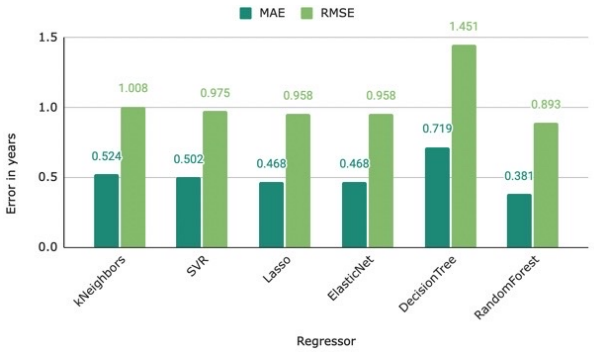

Figure 8: Error comparison for different algorithms for Death prediction using CT-Data

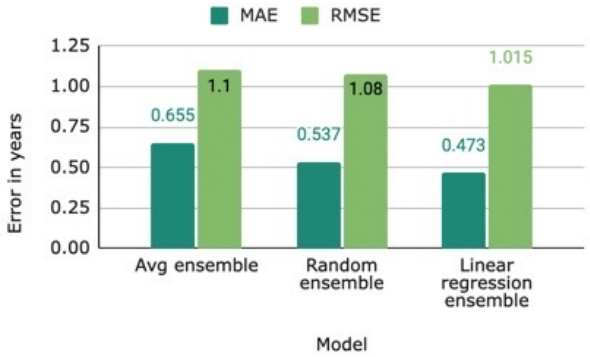

Figure 9: Error comparison for different Ensemble for Death prediction using CT-Data

5.2 Death Prediction with CT and Clinical Data

By adding clinical data to the training set, we note that although the underlying regression models do not show a very significant change in accuracy, the ensemble model benefits by 4 per cent in MAPE.

5.3 Diabetes Classification

Figure [13] shows the acccuracies of six different models on the Diabetes Classification task. KNN gives the best results out of all the models, with an accuracy score of 67%.

5.4 Biological Age Prediction

The results of our experiments with the two models (Neural Networks and Random Forest) can be found in

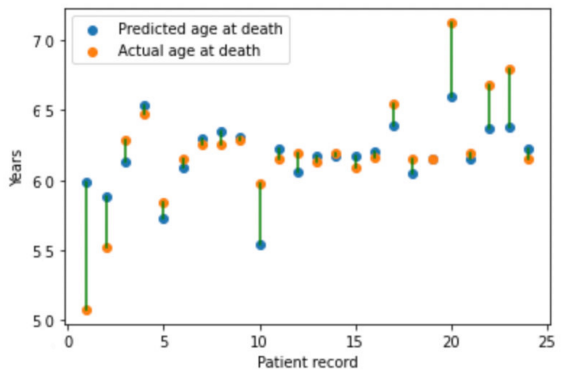

Figure 10: Performance of Locally Weighted Ensemble Regression on CT and clinical Data

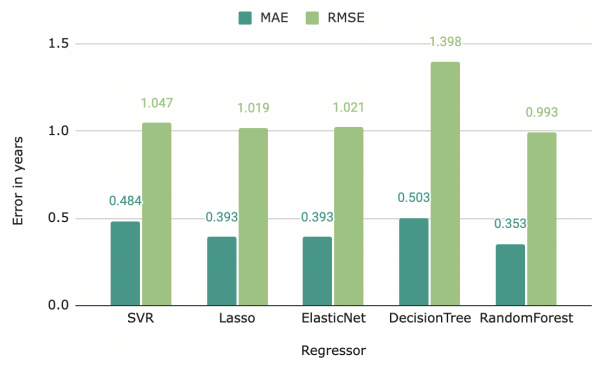

Figure 11: Error comparison for different models for Death prediction using CT and Clinical Data

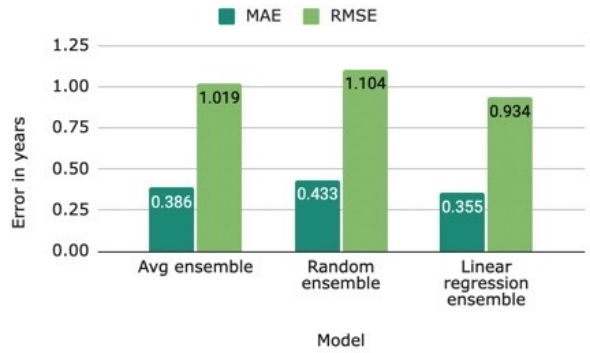

Figure 12: Error comparison for different Ensembles for Death prediction using CT and clinical Data

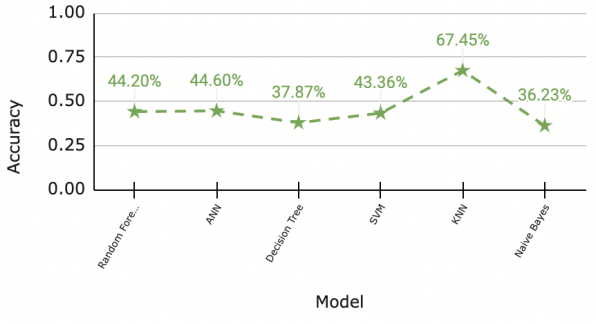

Figure 13: Accuracies of various models on diabetes prediction task

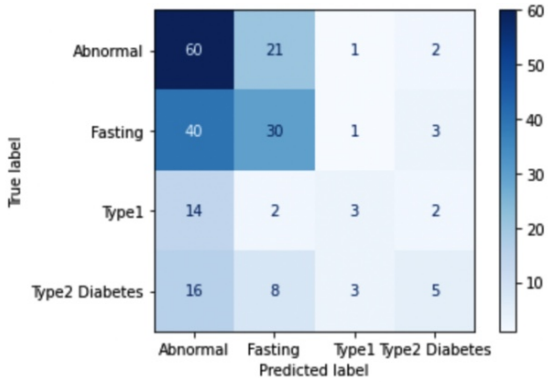

Figure 14: Confusion Matrix-Diabetes Classification

Class label	Precision	Recall
Abnormal	0.47	0.71
Fasting	0.49	0.40
Type-1 Diabetes	0.37	0.14
Type-2 Diabetes	0.41	0.15

Table 1: Diabetes Classification:Precision-Recall

Figure[15] and Figure[16]. Random Forests outperform Neural Networks by a MAPE margin of 4%.

6 Conclusions and Future Work

We have studied CT and clinical data, and tried to build models to use these as biomarkers to predict adverse clinical outcomes, primarily death and diabetes, and developed an estimator for biological age. We reckon that our data is concentrated around the age group of 50-60 years, and our models might, therefore, inherently include some bias. Since the accuracy of biomarkers differ with age, we plan to test this model on a more diverse distribution

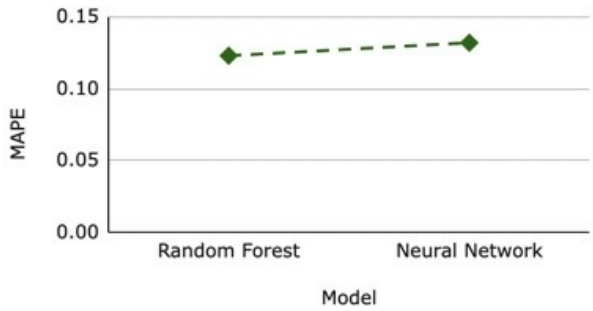

Figure 15: BA Prediction:Mean Absolute Percentage Error(MAPE) of Random Forest and ANN models

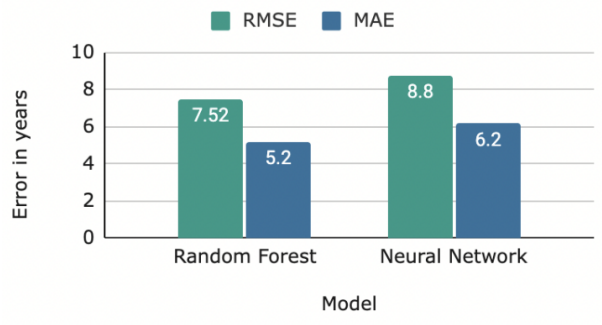

Figure 16: BA Prediction:RMSE and MAE of Random Forest and ANN models

of age. A adolescent cohort might help us identify areas of improvement for our models. We leave that for future work. We also plan to use Cox Index and Proportional Hazards Models on the data.

Although the search for biomarkers for biological age prediction and estimation of mortality is an active domain of research, the vast pool of candidate biomarkers from diverse areas of medical science, along with the lack of pristine data pertaining to only a single marker has made it a difficult problem to solve. There have been studies exploring genetic,physiological and metabolic markers, radiology, to us, appears to be one of the fields which have not yet been extensively explored in this regard. It may not be very intuitive to use tomography data for prediction of clinical outcomes other than those closely related to the ailments for which the scan was actually performed. However, machine learning models can be harnessed to derive insights from such data to discover novel relations between tomography data and not-so-closely related outcomes.

References

[1] George T Baker III and Richard L Sprott. Biomarkers of aging. *Experimental gerontology*, 23(4-5):223–239, 1988.

[2] Daniel W Belsky, Avshalom Caspi, Renate Houts, Harvey J Cohen, David L Corcoran, Andrea Danese, HonaLee Harrington, Salomon Israel, Morgan E Levine, Jonathan D Schaefer, et al. Quantification of biological aging in young adults. *Proceedings of the National Academy of Sciences*, 112(30):E4104–E4110, 2015.

[3] Krishnan Bhaskaran, Isabel dos Santos-Silva, David A Leon, Ian J Douglas, and Liam Smeeth. Association of bmi with overall and cause-specific mortality: a population-based cohort study of 3· 6 million adults in the uk. *The lancet Diabetes & endocrinology*, 6(12):944–953, 2018.

[4] Maryellen L Giger. Machine learning in medical imaging. *Journal of the American College of Radiology*, 15(3):512–520, 2018.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, 2015.

[6] Naomi Hirota, Shinya Suzuki, Takuto Arita, Naoharu Yagi, Takayuki Otsuka, and Takeshi Yamashita. Prediction of biological age and all-cause mortality by 12-lead electrocardiogram in patients without structural heart disease. *BMC geriatrics*, 21(1):1–8, 2021.

[7] Juulia Jylhävä, Nancy L Pedersen, and Sara Hägg. Biological age predictors. *EBioMedicine*, 21:29–36, 2017.

[8] Arun S Karlamangla, Burton H Singer, and Teresa E Seeman. Reduction in allostatic load in older adults is associated with lower all-cause mortality risk: Macarthur studies of successful aging. *Psychosomatic medicine*, 68(3):500–507, 2006.

[9] Petr Klemera and Stanislav Doubal. A new approach to the concept and computation of biological age. *Mechanisms of ageing and development*, 127(3):240–248, 2006.

[10] Morgan Levine. Modeling the rate of senescence: Can estimated biological age predict mortality more accurately than chronological age? *The journals of gerontology. Series A, Biological sciences and medical sciences*, 68, 12 2012.

[11] Luke Sagers, Luke Melas-Kyriazi, Chirag J Patel, and Arjun K Manrai. Prediction of chronological and biological age from laboratory data. *Aging (Albany NY)*, 12(9):7626, 2020.