


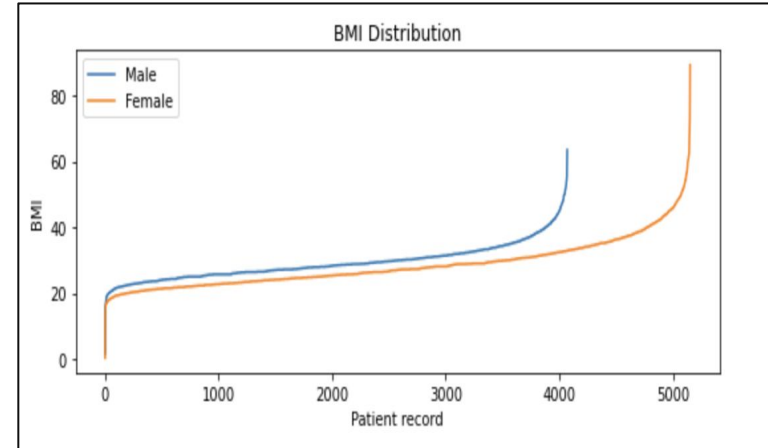
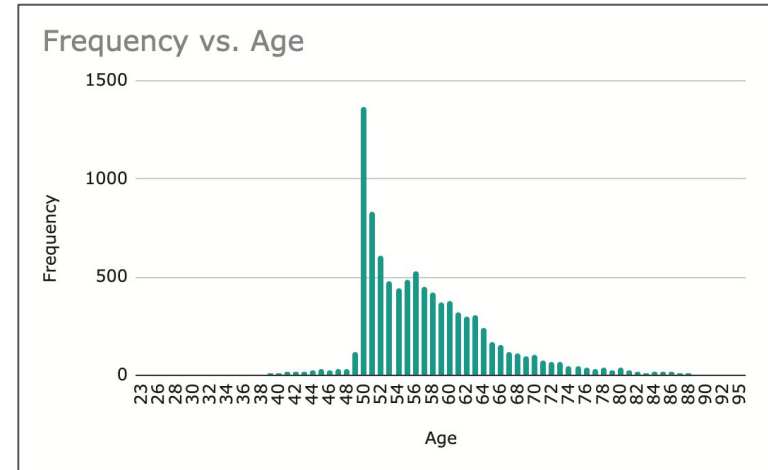
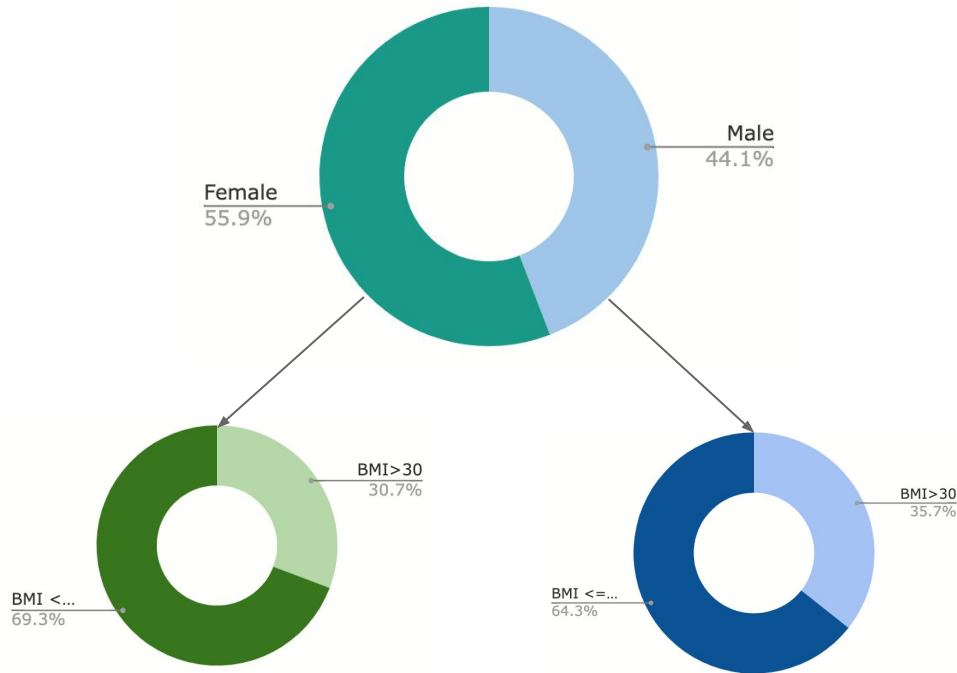
# Ensemble Techniques for Prediction of Adverse Clinical Outcomes using CT Biomarkers



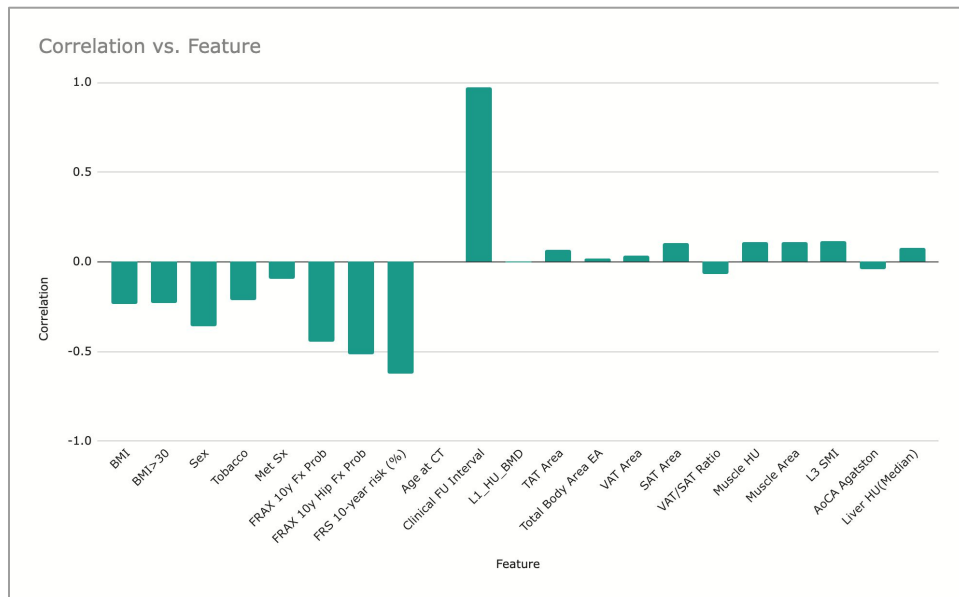
Harshith Gundappa  
Reetuparna Mukherjee  
Sanyam Kakkar



# Data Analysis



# Data Preprocessing



## CT Data

- Null values in the CT data replaced with mean

## Clinical Data

- Replaced null values in BMI column with mean
- Converted categorical data to numerical classes
- Treated blanks in the Tobacco as no tobacco usage

## Dimensionality Reduction

- 95% variance ~ 6 component PCA
- Since the features were < 30, we proceeded without reduction

# Populating sparse Death data

## Approach 1 : Formula Based

*"Compared with individuals of healthy weight (BMI 18.5–24.9 kg/m<sup>2</sup>), life expectancy from age 40 years was 4.2 years shorter in obese (BMI ≥30.0 kg/m<sup>2</sup>) men and 3.5 years shorter in obese women, and 4.3 years shorter in underweight (BMI <18.5 kg/m<sup>2</sup>) men and 4.5 years shorter in underweight women."*

[\[1\]](#)

$$r_{tob} = 9.2$$

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

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

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

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

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

# Populating sparse Death data

## Approach 2 : KNN Based

*"Patients with similar health conditions and data should have similar life expectancy"*

---

### Algorithm 1 Regressor

---

$D \leftarrow \{a_1, \dots, a_n\}$

$L \leftarrow \{a_i \mid \forall i \text{ if } D[a_i]['Death'] \neq null\}$

$U \leftarrow D - L$

**for**  $a_i$  **in**  $U$  **do**

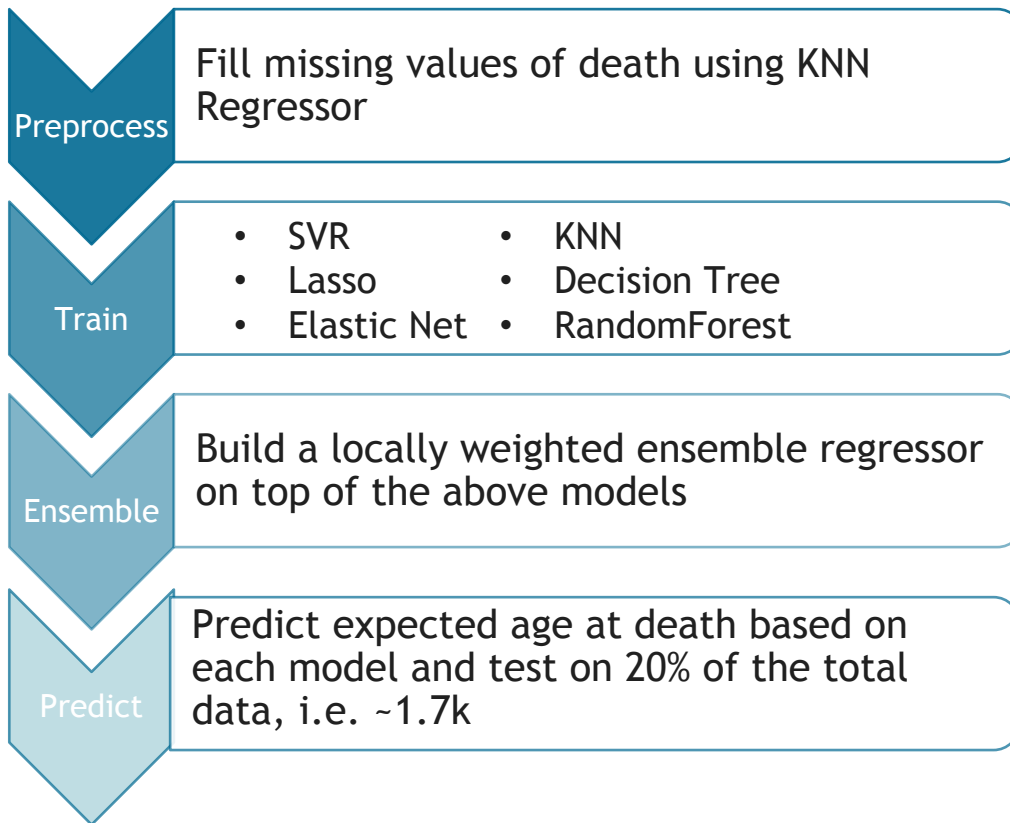
$(d_1, \dots, d_k) \leftarrow kNN(a_i, L)$

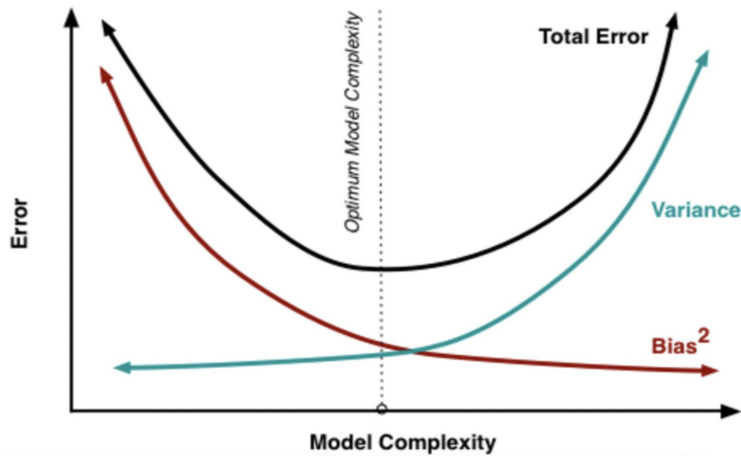
$a_i['Death'] \leftarrow avg(d_1, \dots, d_k)$

**end for**

---

# Methodology





# Regularization in Training

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.

**Problem:** The unbiased OLS would place us on the right-hand side of the picture, which is far from optimal.

**Regularize** - to lower the variance at the cost of some bias, thus moving left on the plot, towards the optimum.

# Regularization

## Lasso

- Adds a penalty for non-zero coefficients
- Lasso penalizes the sum of their absolute values (L1 penalty)
- Many coefficients are exactly zeroed under lasso



Eliminate some features entirely and gives a subset of predictors that helps mitigate multi-collinearity and model complexity

## Ridge

- Decreases the model complexity
- Instead of forcing them to zero, penalize them if they are too far from zero, thus enforcing them to be small in a continuous way.



Decreases the complexity of a model but does not reduce the number of variables, it rather just shrinks their effect

## ElasticNet

A convex combination of Lasso and Ridge

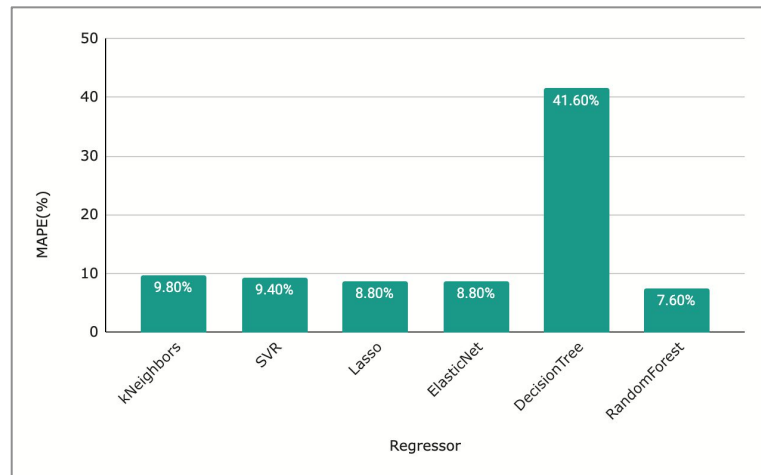
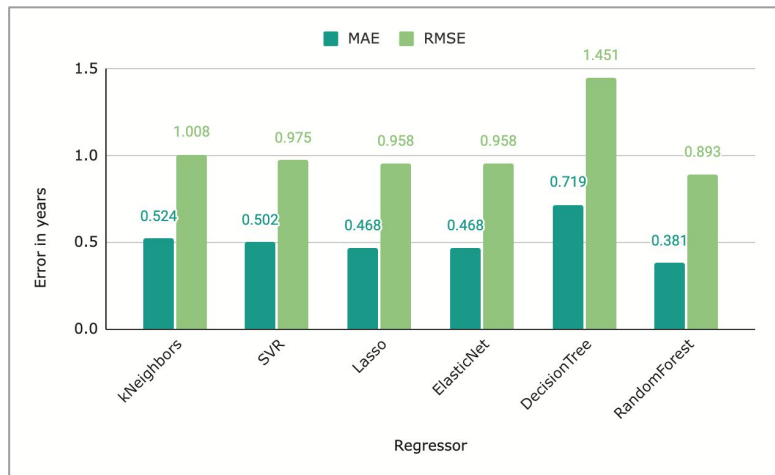


Shrinks some coefficients and set some to 0 for sparse selection



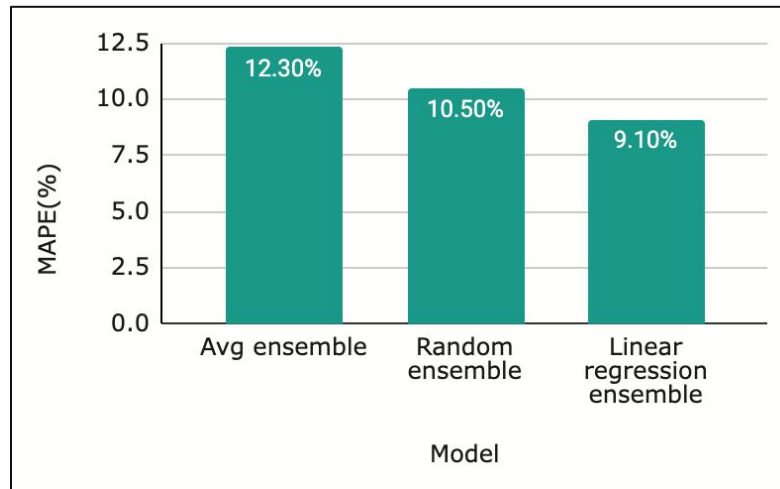
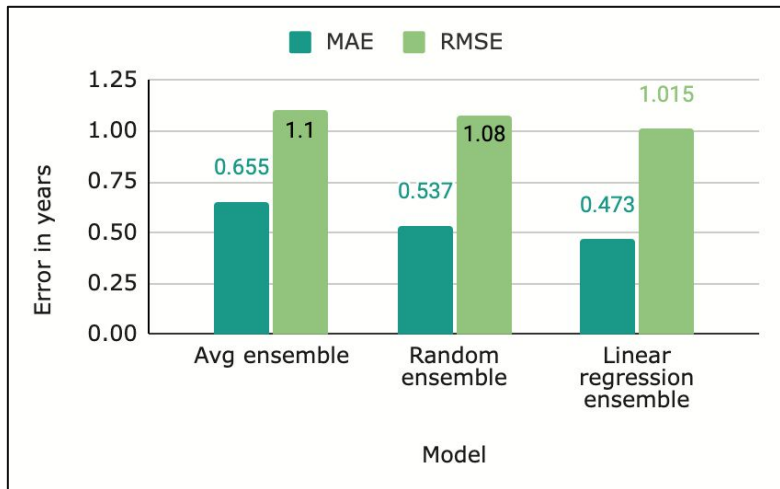


# Death Prediction using CT Data only



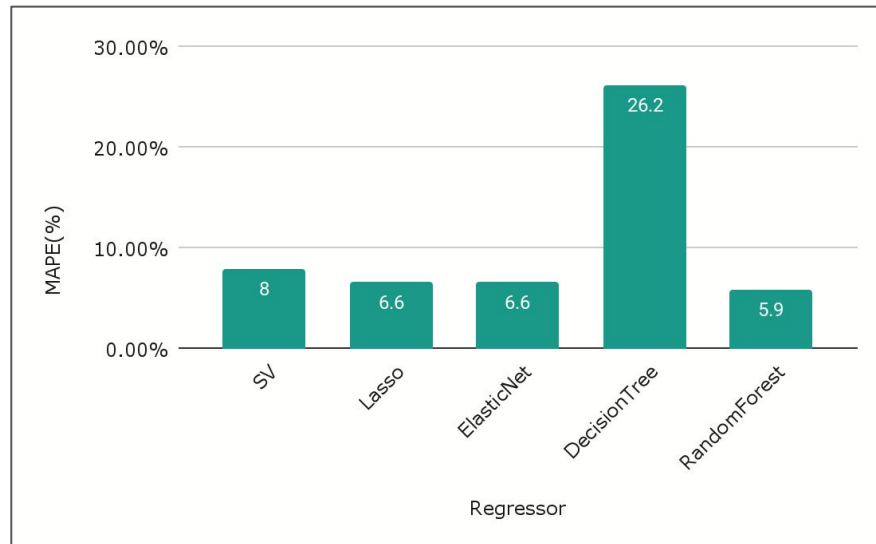
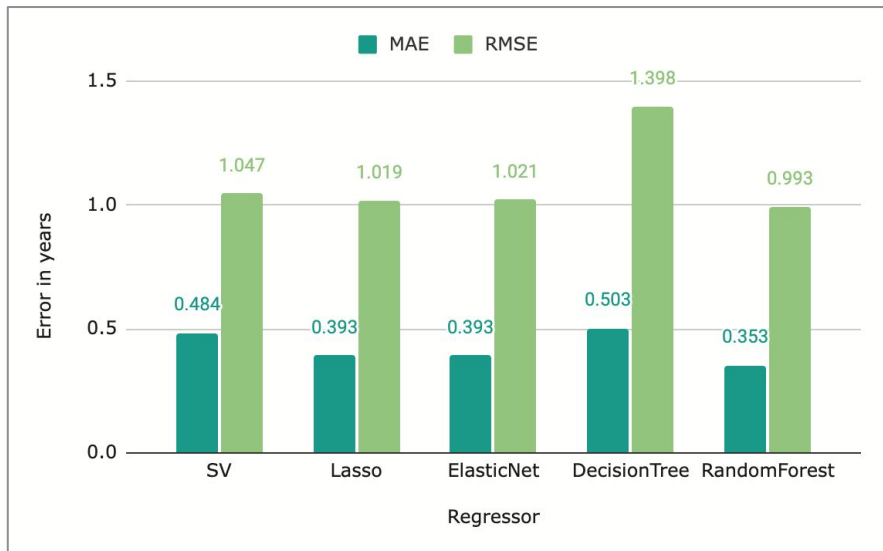
- Random Forest yields the best results
- Tuned the hyperparameters like maximum depth, minimum sample split by comparing the error metrics

# Death Prediction using CT Data only

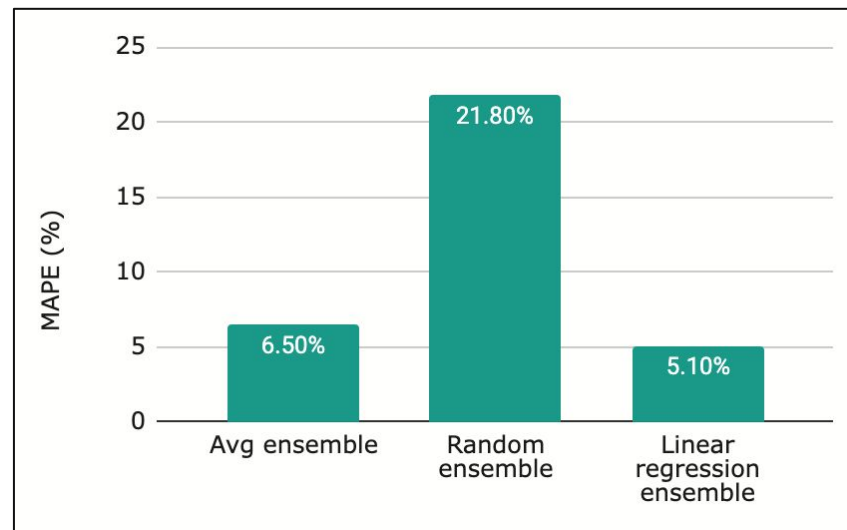
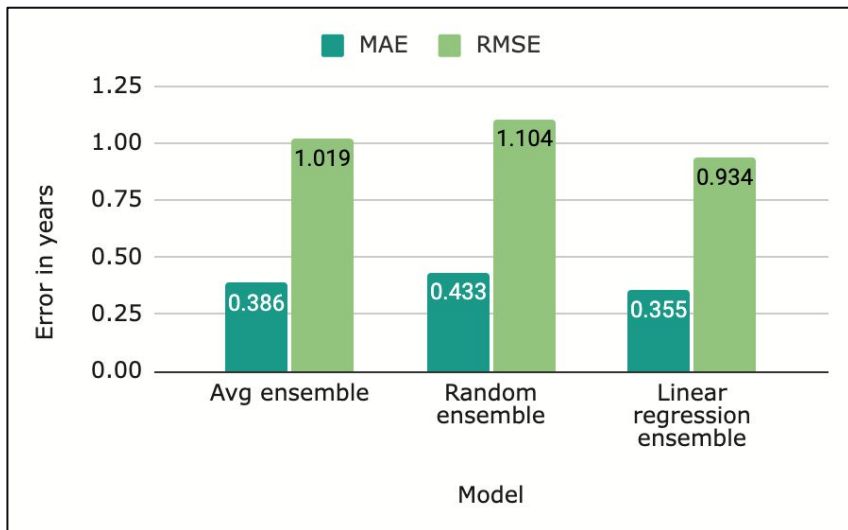


- Linear regression was used to get the weights for each of the models

# Death prediction using CT+Clinical Data

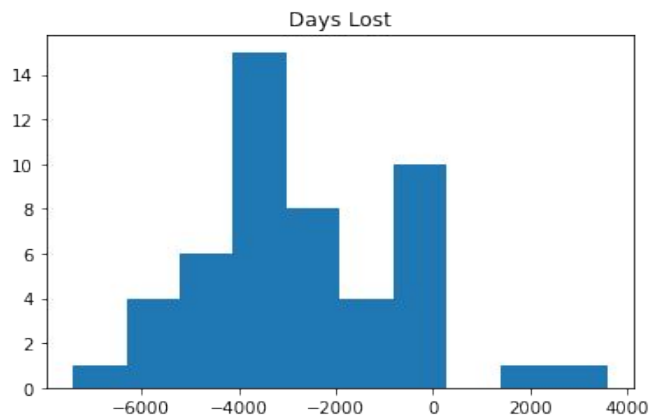


# Death prediction using CT+Clinical Data



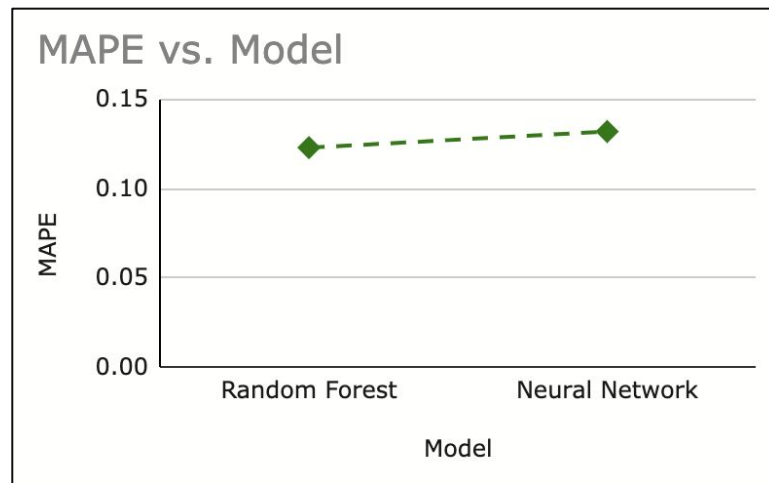
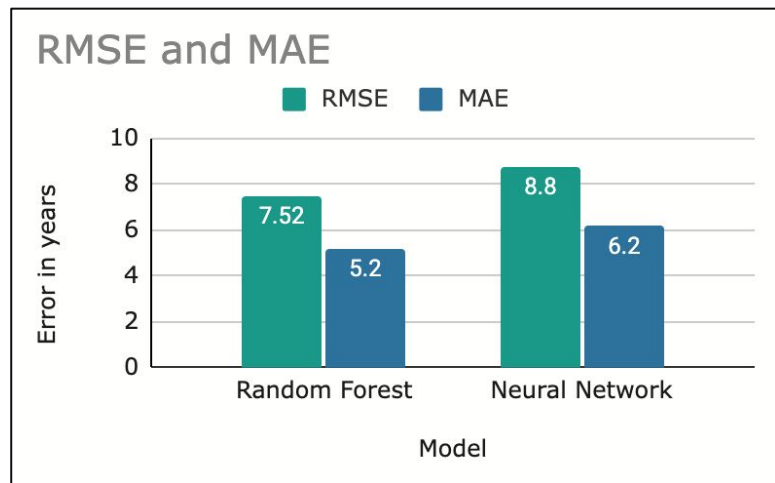
- Adding clinical data reduced MAPE by 4%

# Biological Age Prediction



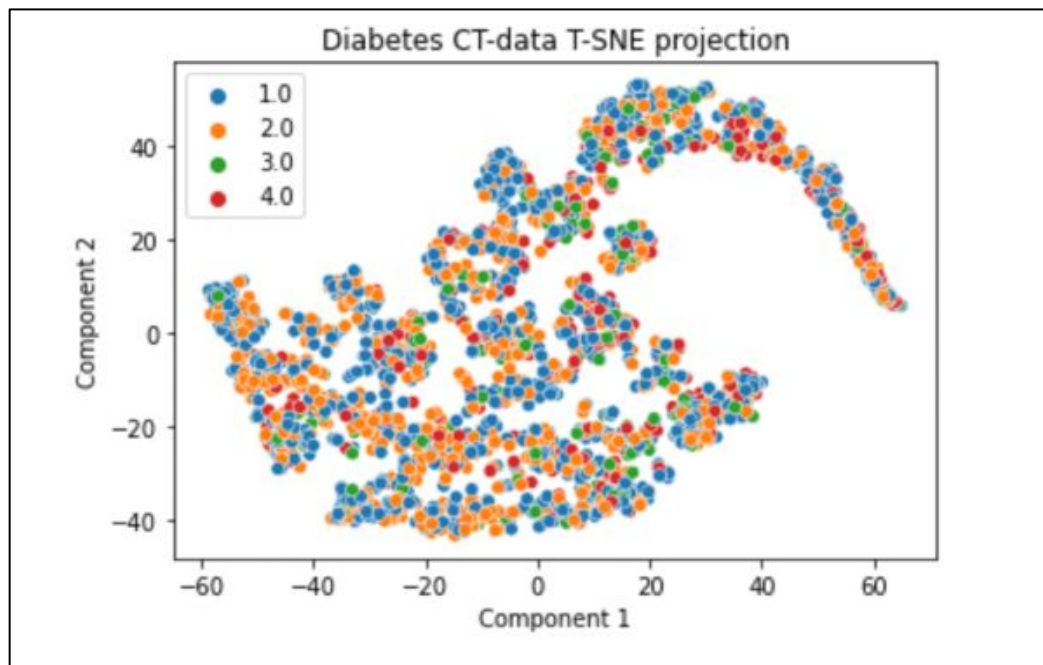
- We assume that if a person is healthy, he will die at the biological age equal to the average life expectancy.
- We then train a model to learn how many years of life was lost due to having a particular clinical condition(ex:Diabetes, Heart Failure). Clinical data and clinical outcomes of dead patient records are used as features vectors to train this model.

# Biological Age Prediction

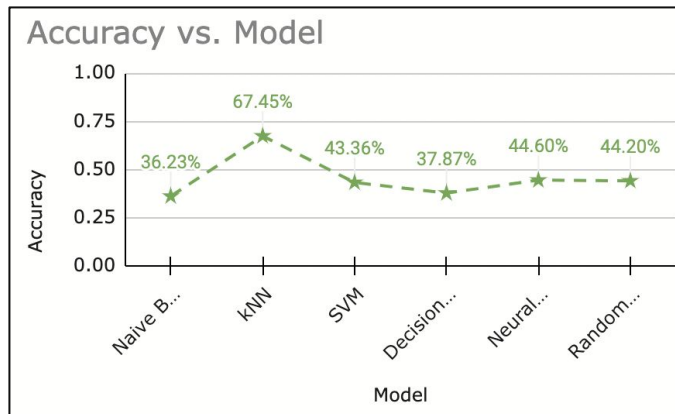
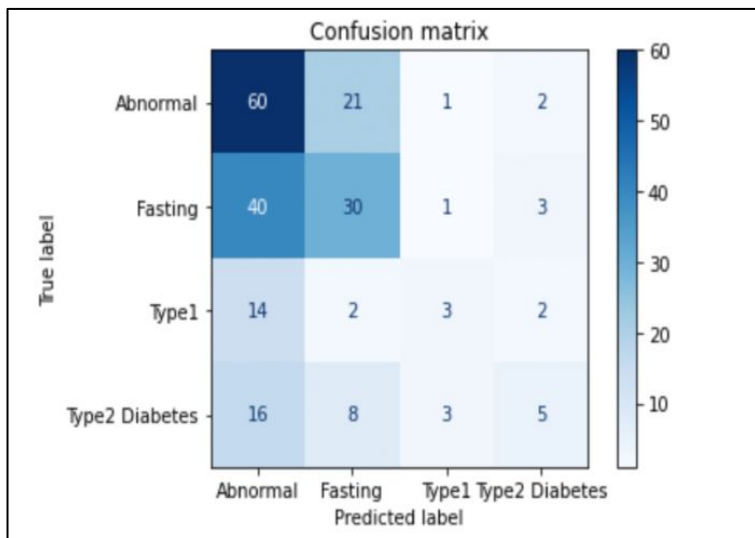


- Random Forest outperforms Neural Networks

# Diabetes Prediction



# Results



	Precision	Recall
Abnormal	<b>0.46</b>	<b>0.71</b>
Fasting	<b>0.49</b>	<b>0.40</b>
Type 1	<b>0.37</b>	<b>0.14</b>
Type 2	<b>0.41</b>	<b>0.15</b>





# Conclusion

- In clinical data, Clinical FU Interval impacts results significantly
- The data is centered around the age group 50-60, so the model may not perform so well on a test set of younger populations. This is in concurrence with already existing literature.
- CT
- Bio age?