Risk Prediction for Cardiovascular Disease using the Framingham Heart Study Dataset

Akshat Gupta [1,4] , Swati Sharma [2,4] , Rajorshi Mondal [3,4]
[1] akshat21515@iiitd.ac.in, [2] swati21568@iiitd.ac.in, [3] rajorshi21187@iiitd.ac.in
[4] Indraprastha Institute of Information Technology, Delhi

1. Introduction

1.1. Motivation

Cardiovascular Disease (CVD) remains one of the leading causes of mortality worldwide. One of the most important aspects of managing CVD is identifying people with high associated risk, and diverting more attention towards this population, which necessitates the need for a precise and effective risk scoring system. Although various risk assessment systems exist, they often tend to cater to specific population demographics. Consequently, there exists a significant opportunity for the enhancement of their performance and applicability.

This inspiration prompted us to contemplate and potentially develop a novel system capable of forecasting individual patient risks related to CVD. The ultimate goal of this study is to streamline the identification of patients at elevated risks of CVD. Such an advancement would facilitate tailored interventions and precise treatment strategies, thus assisting healthcare professionals in potentially saving lives.

1.2. Problem Overview

This study aims to develop an innovative and precise risk assessment system for predicting individual patients' risks associated with Cardiovascular Disease.

The objective is to identify individuals at elevated risk accurately, possibly enabling personalized interventions and targeted treatment plans. The proposed system should be able to precisely and timely determine individuals at high risk of CVD. The system should be able to accurately predict the probable time interval of certainty of CVD from the time of input of a patient's predictors to the model.

2. Literature Review

2.1. Scope

This literature review serves to gain an in-depth understanding of the concepts of Survival Analysis and Competing Risks. It also aims to review past research that involves the use of any form of Medical Records of an individual to predict their risk associated with CVD. This was seconded by an objective of identifying a good dataset for the purpose of this study, and identifying a set of predictors that could potentially result in a precise risk assessment system

2.2. Survival Analysis

Survival Analysis is a statistical approach used to model and predict time-to-event data. In other words, it is a statistical methodology used to predict the time for an event of interest to occur. This event could be any occurrence with a specific start and endpoint, such as the failure of a machine, the onset of a medical condition, or the death of a patient. Key concepts of Survival Analysis include:

a. Survival Function (s(t)) - The survival function represents the probability that an event has not occurred by time t. It gives the likelihood of survival beyond a given time point.

b. Hazard Function (h(t)) - The hazard function describes the instantaneous rate at which events occur, or the conditional probability, at time t given that the event has not occurred before. It helps understand how the risk of the event changes over time.

c. Censoring - In many real-world scenarios, not all individuals experience the event within the study period, or at times opt out of the study before the event has occurred for the said individual. Censoring occurs when the exact event time is unknown but is known to be later than a certain point.

d. Survival Time - This is the time elapsed from a well-defined starting point to the occurrence of the event (or to the point where the sample point was excluded from the study) or to the end of the study period in cases where the event has not occurred.

2.3. Competing Risks

Competing risks refer to situations where there are multiple types of events that can occur over time, and an indi-

vidual may experience one of these events while being at risk of the others. In other words, the occurrence of one event "competes" with the occurrence of other events.

For example, consider patients with a chronic disease who can experience different types of events, such as death from the particular disease, treatment-related side effects, or death from unrelated causes. These events are competing because once a patient experiences one event, they are no longer at risk for the other events.

In the context of survival analysis, competing risks introduce complexity because traditional survival analysis often assumes that individuals are at risk for a single event, typically an event of interest like disease progression or death. However, when competing risks are involved, individuals are at risk of multiple events simultaneously. These could potentially be disease progression, treatment-related side effects, death from unrelated causes, etc. Such cases require specialized modelling techniques that can account for more than one event of interest and accurately predict the time of occurrence of the earliest (or all depending on the use case) of the events of the interest.

2.4. Prediction of cardiovascular risk factors from retinal fundus photographs [5]

This research paper explores the potential use of phenotypic information, particularly regarding vascular health, to identify individuals at a high risk of cardiovascular disease. The authors propose using retinal images to extract additional signals for risk assessment, and applying deep learning techniques, specifically Deep Convolutional Neural Networks (CNNs), on those images for cardiovascular risk assessment.

This study employed two datasets: the *UK Biobank*, consisting of health measurements from 500,000 participants aged 40-69, and the *EyePACS* dataset, a collection of retinal images from diabetic eye disease screenings. The development set consisted of retinal images from *EyePACS* and the clinical validation set (*EyePACS-2K*) included a subset of macula-centered images along with *HbA1c* measurements, short for glycated hemoglobin.

A Deep Neural Network (NN) model based on the Inception-v3 [6] architecture was utilized to predict diverse cardiovascular risk factors such as *Age, Gender, Smoking status, BMI, Blood Pressure*, and *HbA1c* from these retinal images. The average value of each of these indicators was used as a baseline for continuous predictions while *AUC* measured binary classification. The Mean Absolute Error (MAE) for predicting age was notably lower than baseline values. The algorithm also showed high precision in predicting *Systolic Blood Pressure, BMI,* and *HbA1c*.

The algorithm's performance was evaluated against specific error margins and outperformed baseline accuracy, especially for *Age* and *Blood Pressure* predictions. The model successfully inferred *Ethnicity* as a potential risk factor. Additionally, the study examined the impact of *Diabetic Retinopathy* on algorithm performance and found no significant differences between the groups of *EyePACS-2K* dataset.

Despite a limited number of events, the model achieved an *AUC* of 0.70, comparable to established risk calculators. Combining multiple risk factors was more effective in predicting *Major Adverse Cardiovascular Events (MACE)* than individual factors alone. Soft attention was used to identify anatomical regions influencing predictions. *Blood Vessels* were highlighted for *Age, Smoking,* and *Systolic Blood Pressure (SBP)* predictions. Perivascular areas were emphasized for *HbA1c* predictions. Gender predictions focused on the optic disc, vessels, and macula, while *SBP* and *BMI* predictions showed more diffused attention patterns, possibly indicating distributed signals.

This study thus explored the direct correlation between retinal images and cardiovascular events, training a model to predict *MACE* within five years. It also successfully established *Age, Gender, Smoking Status, BMI, Blood Pressure, HbA1c,* and *Ethnicity* as cardiovascular risk factors. This was referred to when selecting the dataset and during the feature selection procedure, verified by the statistical analysis (See Section **3.1**) of a processed dataset to select statistically significant features as predictors for the risk assessment system.

2.5. DeepHit [4]

In this paper, Lee et al. describe the architecture and use of a deep neural network to learn and predict the distribution of first-hitting times. The key aspect of their method is that it smoothly handles situations in which there is a single underlying risk (cause) and situations in which there are multiple competing risks (causes). DeepHit employs a network architecture that consists of a single shared sub-network and a family of cause-specific sub-networks to predict the time for the event of interest. See Fig. 1 for a depiction of the architecture of DeepHit.

The separating aspect of this method to the existing survival analysis models is that

> "it makes *no assumptions* about the form of the underlying stochastic process; it therefore allows for the possibility that...both the parameters and the form of the stochastic process *depend on the covariates*."

The assumption being referred to is the assumption introduced in the Cox Proportional Hazards [2] model. These are

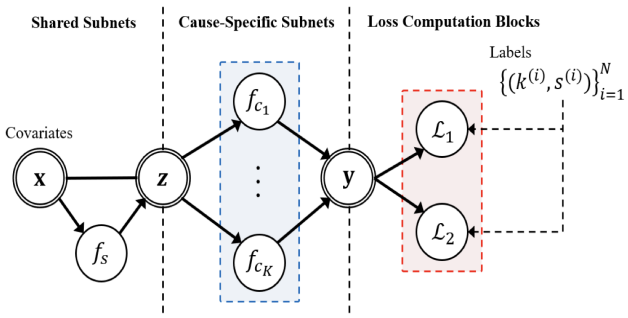

Figure 1. DeepHit Architecture

as follows:

a. survival curves for different strata must have hazard functions that are proportional over the time t

b. the relationship between the log hazard and each covariate is linear

They describe the capability of DeepHit to be able to accurately predict the risk of events even in datasets with competing risks, or right-censoring (patients are lost to follow-up).

For the purpose of their study, they use a dataset with 1) patient's covariates, 2) time elapsed since covariates were first collected, and 3) a label indicating the type of event that occurred. The time indicated here is either the time at which the event occurred or the time beyond which the patient was excluded from the study. Given the number of events-of-interest are K, the DeepHit network uses K cause-specific sub-networks and an additional shared sub-network. The network aims to learn the joint distribution estimate of the K events and the "first hitting time." The first hitting time refers to the time the first event occurred, assuming only one event can occur.

Lee et al. train their network on multiple datasets and compare its performance against conventional Machine Learning and Deep Learning models that underline the same objective. The datasets used were the SEER, METABRIC, UNOS and a synthetically-created dataset of survival data. The synthetic dataset included 3 4-dimensional variables per patient modelled by *N(0,1*, and 2 hitting times, thus 2 events of interest. The dataset was constructed to have 30K patients, among which a random set of 15K were right-censored.

The network architechture included 4 layers - 1 layer for the shared sub-network, 2 fully-connected layers for the cause specific networks, and 1 softmax layer (the output layer). The activation function across layers 1, 2 and 3 was *ReLU*, and the network was trained through back propagation.

Lee et al. used *time-dependent concordance index* (or C^{td} index) as a metric of performance [1]. For the SEER and SYNTHETIC datasets, which have two events (competing risks), DeepHit performed better than conventionally-used models Fine-Gray proportional sub-distribution hazards model and deep multi-task Gaussian process (DMGP) with a mean C^{td} index of 0.684. For the UNOS and METABRIC datasets, which have a single event, DeepHit was compared against two families of models, one that are conventional survival regression models (Cox Proportional Hazards, Threshold Regression, and Random Survival Forests (RSF)) and ones that are used for mortality predictions and derived from machine-learning algorithms such as Logistic Regression, AdaBoost, etc, and the deep neural network DeepSurv [3] (based on the Cox proportional assumption). DeepHit showed significant improvement in performance against all models execept AdaBoost for the UNOS dataset, and RSF for the METABRIC dataset.

3. Implementation

3.1. Data

The dataset identified for the purpose of this study was the Framingham Heart Study Dataset. [7] The dataset contains the information of precisely 11628 patients, with 39 features (or columns) per each patient. Out of the 11628, 2899 patients have *CVD = 1* and the rest 8729 have *CVD = 0*. Clearly, there is a very high class imbalance; such a trend is prevalent across all target variables for this dataset. See Fig. 2 for a visualization of the same.

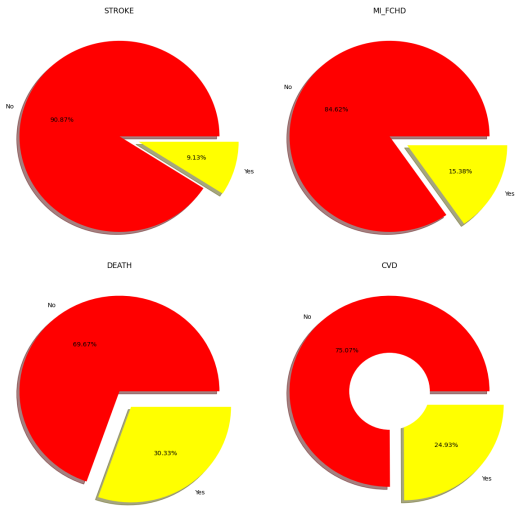

Figure 2. CVD Distribution

The dataset includes features that can be broadly described as their Demographics, Medical History, Administered Medications, Indicators of Events with Timestamps

for each event, and Identification numbers. The *RANDID* column was dropped as it is clearly insignificant towards predicting the risk of *CVD*.

Columns with empty cells i.e. null values were identified. Initially, the columns *educ, HDLC*, and *LDLC* were dropped as these had a significantly high percentage of null values and thus could not be used for making accurate predictions. For the *TOTCHOL* and *HEARTRTE* columns, null values were replaced by their respective mean values. For dependent variables, particularly *CIGPDAY, BMI* and *GLUCOSE*, the median values of the set of sample falling into the dependent categories were used to replace the null values. *BPMEDS* is a dependent variable depending on *SYSBP* and *DIABP*. The null values for *BPMEDS* were replaced by predictions made by a Random Forest Classifier given the input variables as *SYSBP* and *DIABP*.

A new column *DIFCVD* was constructed as *DIFCVD* = *TIMECVD* - *TIME* indicating the time to the event (i.e. *CVD*) from the time of data collection. All columns indicating events and times of events were dropped from the dataset, only keeping the target variables concerned with this study i.e. *CVD* and *DIFCVD*. A subset of features was identified as significant in the prediction of risk associated with *CVD* using Standard t-Test (99% C.I.) and Chi-sqaure Tests (99% C.I.) for regressive and categorical variables (excluding the target variables) respectively. The statistical tests identified two features, namely *HEARTRTE* and *CURSMOKE*, as insignificant, which were then filtered out.

Henceforth, *DIFCVD* will be referred to as the time of event and *CVD* will be referred to as the event.

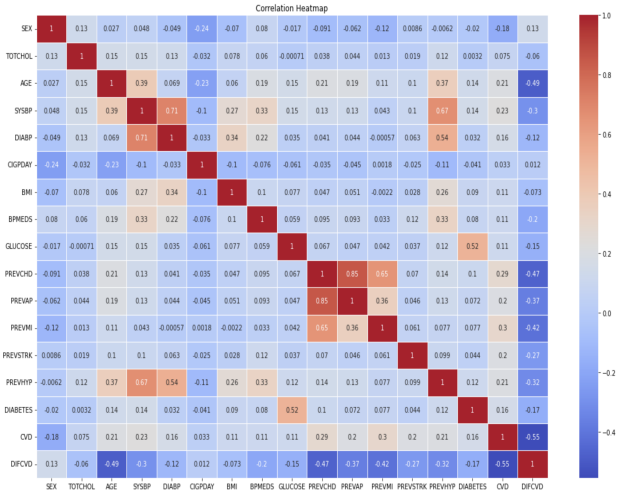

Figure 3. Correlation Heatmap for Processed Dataset

The so-identified features will now be used to predict the event and time of event associated with each patient. The regressive variables in the reduced set of features were standardized and categorical features were one-hot encoded to develop more accurate predictions. See also Fig. 3 for a Correlation Heatmap for the final pre-processed dataset.

3.2. Evaluation Metrics

Survival Analysis models are evaluated using specialized metrics such as the C-score or the C^{td}-index [1] instead of the conventional evaluation metrics such as Accuracy, Precision, etc. The evaluation metric used throughout this study is C-score.

3.3. Algorithm Considerations

When dealing with a dataset that necessitates simultaneous prediction of both a classification (occurrence of the event of interest) and a regression label (time of event), it introduces a multi-output challenge. Traditional machine learning models like Logistic Regression, K-Nearest Neighbors, Random Forest Classifier, and SVM Classifier are inherently designed for either classification or regression tasks, but not both concurrently. Initially, these models were considered for the dataset; however, they proved unsuitable due to the dual prediction requirement and their inability to make accurate predictions for a time-to-event dataset since they assume independence between observations.

Models that cater particularly to Survival Analysis problems were thus explored. These include, but are not limited to, the Kaplan-Meier Estimator (for visualization of the estimation of the survival function), Cox Proportional Hazards Model, and Random Survival Forests. The choice of models is grounded in a thorough literature review and preliminary investigations, which have guided us in identifying these as promising candidates for good results. This experimentation phase was expected to provide valuable insights into the strengths and limitations of the selected models, ultimately informing our decision-making process.

4. Results

In this section, we present the results of our machine learning models for predicting cardiovascular disease (CVD) events. We utilized the Cox Proportional Hazard model (Cox PH) and Random Survival Forest (RSF) to estimate survival probabilities and evaluate the models' performance.

4.1. Cox Proportional Hazard Model

We employed a 5-fold cross-validation strategy for training and testing the Cox PH model. The cumulative distribution functions (CCDFs) corresponding to each split were predicted, and the Concordance Time-dependent (ctd) in-

dices and C-indices were computed for each fold. The results are summarized in Table 1.

Metric	Value
C^{td}-index	0.7303
C-index	0.7585

Table 1. Performance of CoxPH model

The ctd-index measures the predictive accuracy of the model over time, with a higher value indicating better performance. Similarly, the C-index represents the concordance between predicted and observed survival times.

See Fig. 4 for the hazard ratios obtained from the Cox PH model, providing insights into the impact of different features on the risk prediction of CVD events.

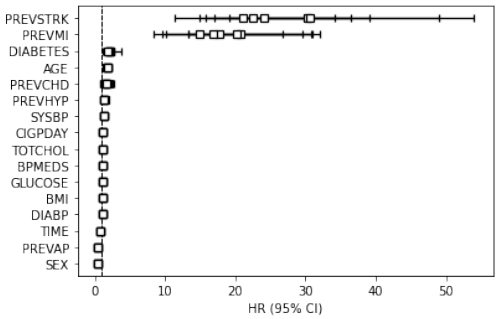

Figure 4. Hazard ratios from CoxPH

4.2. Random Survival Forest

Next, we applied the Random Survival Forest model with similar 5-fold cross-validation. The ctd-indices for each split were calculated, and the mean ctd-index is presented in Table 2.

Metric	Value
C^{td}-index	0.6661
C-index	0.6901

Table 2. Performance of Random Survival Forest

4.3. Kaplan Meier Estimate

The Kaplan-Meier Survival Curve illustrates the estimated survival probabilities at each instance of time recorded within this data with a 95% confidence interval. It is used to visualize the estimated relation of the survival probabilities with passing time, commonly used in survival analysis since they are modelled as time-to-event problems. Each instance of time (say t) corresponds to a certain survival probability (say p_t). It can be interpreted as that the probability of an average individual surviving for a time $T \geq t$ is p_t. See Fig. 5 for the Kaplan-Meier Survival Curve. In our case it models the negative risk associated with Cardiovascular Disease i.e. at time t, the probability of the average individual not developing CVD for a time $T \geq t$ is p_t.

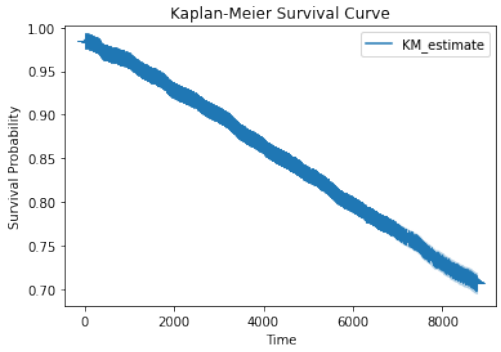

Figure 5. Kaplan Meier Survival Curve

4.4. Mortality Prediction from Traditional ML Models

Lastly, Mortality Prediction models were manufactured from traditional ML models such as ADABoost, Logistic Regression and Random Forests. These, since capable of only solving a regression or classification problem, treated the survival analysis problem as a classification problem with the event of interest being *CVD* i.e. the indicator *CVD* was implicitly used as a target attribute, while not influencing the predictions directly.

Mortality Prediction involved training as many models as there are unique time instances with the labels of each sample that indicated *CVD* before the current time instance under consideration set to 1. More formally, if there are $T = t_1, t_2, ..., t_n$ time instances, we train n models, 1 for each time instance t_i where the label for any sample that indicated *CVD* = $True$ before time t_i is 1, otherwise 0. The label is then treated as the target and the classification model can then be scored by the C^{td}-index as well to measure performance.

The results obtained from running **MP-ADABoost, MP-Logit,** and **MP-RandomForest** are shown in Table 3.

Model	C^{td}-index
MP-ADABoost	0.5456
MP-Logit	0.5800
MP-RandomForest	0.5408

Table 3. Mortality Prediction Models Results

Clearly, **MP-Logit** out performs both the other Mortality Prediction models.

5. Conclusion

In summary, our study leveraged diverse survival analysis models, including the Cox Proportional Hazards model and Random Survival Forest, to predict cardiovascular disease (CVD) risk using the Framingham Heart Study dataset. The Cox model provided insights into proportional hazard relationships, while the Random Survival Forest captured intricate, non-linear patterns.

Key factors influencing CVD risk, such as age, cholesterol levels, blood pressure, and smoking status, were reaffirmed by our models. The Random Survival Forest, with its ensemble nature, excelled in handling complexity of the dataset.

Despite these promising results, challenges and limitations persist, including biases in observational data and the need for careful model validation. Our findings contribute to the understanding of CVD risk, laying the groundwork for refined prediction models and personalized healthcare interventions.

References

[1] Laura Antolini, Patrizia Boracchi, and Elia Biganzoli. A time-dependent discrimination index for survival data. *Statistics in Medicine*, 24(24):3927–3944, 2005.

[2] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2), 1972.

[3] Jared L. Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. DeepSurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1), 2018.

[4] Changhee Lee, William Zame, Jinsung Yoon, and Mihaela Van Der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI conference on artificial intelligence*, number 1, 2018.

[5] Ryan Poplin, Avinash V Varadarajan, Katy Blumer, Yun Liu, Michael V McConnell, Greg S Corrado, Lily Peng, and Dale R Webster. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature biomedical engineering*, (3), 2018.

[6] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015.

[7] Connie W Tsao and Ramachandran S Vasan. Cohort Profile: The Framingham Heart Study (FHS): overview of milestones in cardiovascular epidemiology. *International Journal of Epidemiology*, 44(6), 12 2015.