MACHINE LEARNING FOR MEDICINE: PREDICTING CLINICAL OUTCOMES FOR PATIENTS WITH LEIGH SYNDROME

Nishaal Ramesh Ajmera
Coauthors: Dr Stephen McGough, Dr Amy Vincent, Dr Conor Lawless, Atif Khan

Newcastle AI Lab
School of Computing
Newcastle University

May 8, 2022

ABSTRACT

Leigh syndrome is a rare disease that affects the central nervous system of patients arising from abnormalities in the mitochondria. The progression of Leigh syndrome is affected by many factors making it difficult for clinicians to make an accurate prognosis. This study aims to provide a prognostic tool to clinicians to predict clinical outcomes and determine relationship between the factors affecting disease progression and clinical outcomes of Leigh syndrome. Four survival analysis machine learning models and a unique deep learning model, DeepHit is used to train Leigh syndrome data with 55 patients to predict survival times. The Newcastle Paediatric Mitochondrial Disease Scale data with 20 patients is used to predict two clinical outcomes, age at feeding inability and immobility using the four machine learning models. The Cox-proportional hazard model with five features was evaluated as the best model to predict patient survival. The Cox-proportional hazards model with elastic net was evaluated as the best model to predict age of patients with feeding inability. Our novel contribution to Leigh syndrome research using survival analysis machine learning and DeepHit could be used to enhance prognosis of patients.

1 Introduction

1.1 Motivation

Leigh syndrome (LS) is a neurometabolic disorder that arises from abnormalities in the mitochondria which is responsible for cellular energy [1]. It is estimated that 1 in 36,000-40,000 children are born with this rare disease. This condition impairs the central nervous system of patients reducing the quality of life as the disease progresses[2]. Generally, the onset of LS is observed in children up to 2 years of age but for some patients it can begin in adulthood. Clinicians diagnose LS by assessing patterns of symptoms such as loss of motor skills, delays in developmental milestones and evaluating brain scans of patients [2, 3]. Genetic causes of LS are highly variable as there are more than 75 associated genes arising from mutations either on the nuclear genome (nDNA) or mitochondrial genome (mt-DNA) [4, 5].

Patients with LS have very poor prognosis. Patients are usually given supportive care to enhance quality of life and manage symptoms such as seizures, lactic acidosis and movement disorders[1]. However, it is challenging for clinicians to accurately make decisions on the best treatment and care options. This is due to the large number of variables that influence the progression of the disease and its low prevalence. The data analysed in this project includes results of diagnostic tests, qualitative clinical assessments, genetic and other

phenotypic data from LS patients. Therefore in this project, this data will be used to extract insights into the prognosis of patients with LS by leveraging Machine Learning models that could aid clinicians to provide better and accurate prognosis.

Machine Learning (ML) in Medicine is a rapidly growing area of research. Clinical applications of ML can enable better care and improved understanding of diseases. ML also has the potential to improve efficiency of clinical practices and detect novel trends in diseases that are not detected through conventional scientific methods. However, at present there are limited deployed ML models in clinical practice. An impeding factor for ML deployment is the interpretability of ML models for clinical applications. Model interpretability can be achieved either by using standard attribution-based models that allow interrogation of input features' importance or using novel, often architecture-based techniques that produce counterfactual outcomes.

In this study we have used survival analysis ML models to predict various clinical outcomes of LS patients. Survival analysis allows the duration to an event and probability of occurrence to be studied[6]. The events investigated in this project are clinical outcomes such as age at requiring feeding support, immobility or death of a patient. Survival analysis also allows the incorporation of censored data. This form of data is ubiquitous in clinical studies due to the unknown duration of patient experiencing an event which can be censored without losing all the information about a patient[7]. Several ML survival models such as cox-proportional hazards, accelerated failure time, random survival forest and gradient boosting have been considered. A novel deep learning model, DeepHit was also trained to model the relationship between covariates and patient survival. The background, methodology and model performance of this project is described in sections 2, 3 and 4.

The aim of this project is to build and evaluate machine learning (ML) models using survival analysis to predict clinical outcomes for patients with Leigh syndrome and explaining the relationship between the features used to train the model with the outcome. We believe that ML and deep learning models can accurately predict LS clinical outcomes and give results that are interpretable for clinicians. Ultimately, we hope to provide a tool that could help clinicians in making treatment decisions.

2 Related Work

Literature available in the domain of mitochondrial diseases specifically LS largely demonstrate findings on scientific experiments, utilize traditional statistical analysis or use computer vision to analyze cellular and mitochondrial structures. Academic papers from recent Machine Learning studies that predicted clinical outcomes and/or survival of patients with other diseases including survival analysis based models were researched to determine interpretable models for LS.

Survival analysis is a domain in statistics that estimates the time to an event of interest such as death, conversion to a disease or even machine failure[8]. It is defined by the survival function, S(t) that outputs the probability of a patient not experiencing the event beyond time,t where T is a positive random time point from the population under study.

$$S(t) = Pr(T > t) \tag{1}$$

In medical settings, patients can have different start and end times of an event. Some reasons for this might be patients entered the study later, are lost to follow-up, experienced a different event or did not experience the event of interest within the time-frame of the study[7]. This introduces a type of missingness called censored data that can be incorporated into survival analysis without removing unobserved values.

In a study on LS with 130 patients some clinical features such as age of onset $\leq$ 6months, epilepsy, failure to thrive, genetic mutations such as SCL19A3, mt.8993 T>G and SURF1 were associated with poorer survival[9]. Sofou *et al*. utilized Kaplan Meier Analysis and multiple logistic regression to understand the significant features contributing to patient survival. Nonetheless, ML models can produce more accurate and informative results to model high-dimensional composite data compared to traditional statistical methods. Therefore, for the purposes of this study ML survival models that were applied on a variety of medical conditions are reviewed.

Machine Learning Survival Models A study on COVID-19 patients demonstrated that ML survival algorithms could predict discharge times of Covid-19 patients using age and sex as covariates[10]. In this study, ML survival models such as Cox-proportional hazards, Coxnet, stagewise gradient boosting (GB) and component-wise GB models were applied. The Cox-proportional hazards models can model the probability of the an event occurring at a given time using the hazard function whilst determining the relationship between covariates and the event using regression. This semi-parametric model can serve as a good baseline model for the patients with LS. The stagewise GB model is a powerful ensemble method that combines multiple learners to give good model predictions. This model however is challenging to interpret relationship between the covariates and event. In another study, Spooner *et al*. weighed different ML survival models and feature selection methods systematically using high-dimensional data to predict time development of dementia [6]. The models were assessed for stability training and test the models with 5 repeats of 5-fold cross validation. This study however uses many boosting algorithms which lack of interpretability as the relationship between covariates an outcome event is not measured. The use of random survival forest to determine important features can be helpful to understand which features most influence the survival of LS patients.

Deep Learning Survival Models In a recent study, Wulczyn *et al*. developed an interpretable deep learning system (DLS) that was built on convoluted neural network architecture with a final Cox regression layer that outputs the risk scores of patients[11]. The DLS model was able to stratify patients by risk and predict survival time for stage II and III colorectal cancer using histopathological images. Another deep survival analysis model based on a Weibull distribution was developed by Nagawa *et al*. to predict age at Alzheimer's conversion[12]. This model achieved high accuracy to predict the sequence of patients converting to Alzheimer's using MRI images. These deep survival models worked well for imaging data however its efficiency on clinical features needs further research. Lee *et al.* developed a novel deep learning survival model called DeepHit proven to perform better than previous Deep Learning Survival models[13]. DeepHit learns the distribution of survival times as the model is trained and unlike stringent parametric models, it allows for the covariates and survival time relationship to change over time. A caveat of this model is that does not output the relationship between covariates and outcome event. Nonetheless, this model can accommodate more than one event of interest in that can be applied in settings where a LS patient might have more than one possible outcome.

3 Methodology

3.1 Overview

The methodology for this project is designed as shown in Figure 1 using the Leigh syndrome Data. In the first arm of the project we use the **clinical data** to predict LS patient mortality with ML survival analysis models and DeepHit. The models are trained in 3 cycles to select the most appropriate clinical features that potentially contribute towards patient survival. We also investigate two clinical outcomes as shown in the second arm of Figure1, time to feeding inability and time to severe immobility which are of interest to the clinicians at Wellcome Centre for Mitochondrial Research Newcastle using the Newcastle Mitochondrial Disease Paediatric Scale (NPMDS) data. These two events are denoted when patients who are unable to feed themselves by mouth are placed on enteral tube feeding and patients who are immobile or wheelchair-bound. Hence, the models used to investigate patient survival, feeding inability and immobility are evaluated separately.

3.2 Data Understanding

The data is obtained from mitochondrial research groups in Newcastle, London and Oxford with clinical data for 57 patients. The data in this study is private to maintain patient confidentiality. Although patients have been anonymised, patients can be identified through age and clinical information since LS has low prevalence. Basic data inspection is done to check the dimensions, number of records and missing value proportions. NMDPS had information on children below 16 years[14] which is analysed seperately in this study as shown in Figure 1.

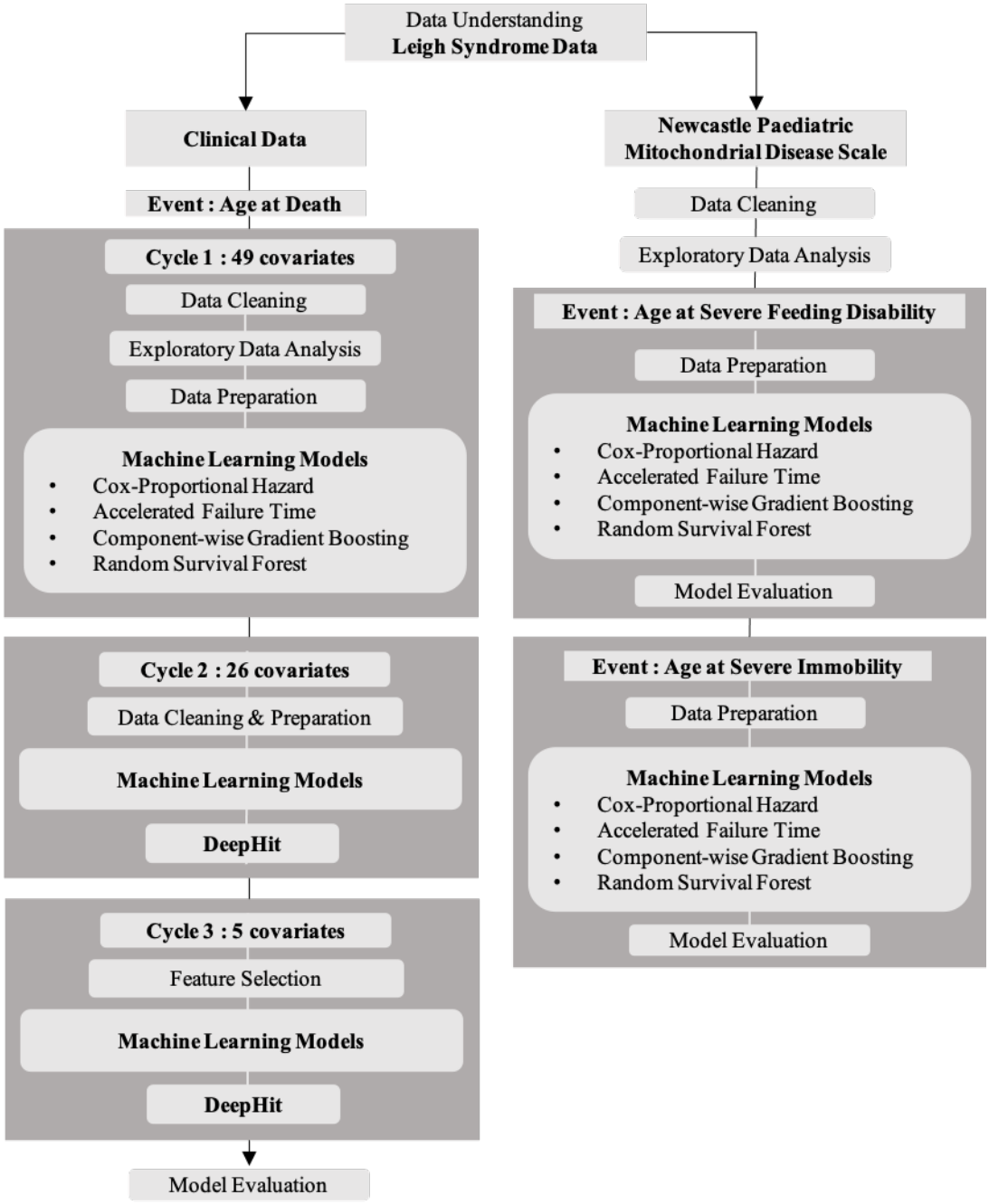


Figure 1: shows overview of methodology

3.3 Data Cleaning

Discussions on data collection and quality with domain-experts revealed that values are not recorded either because patients are not tested for the particular feature at the time of visit, the clinician did not see the need to test the patient for a particular feature or the status of that feature is not transferred from the reports into the database. Since a substantial proportion of the data is missing, a global approach to impute or remove the missing values will deliver biased results. The high proportion of missing data provided reasonable evidence to conclude that the data is either Missing at Random (MAR) or Missing Not at Random (MNAR) [15]. The NMDSA is not analysed further as it had information for only two adult patients.

Clinical Data Features that had $\geq 60\%$ missingness and longitudinal clinical records with missing dates are removed in the first cycle. The various longitudinal clinical assessments and biographic data are joined by patient ID and date to produce a single dataset for analysis. The remaining missingness in data are then imputed using various methods. Missingness in some features such as histochemistry, opthalmology and audiology are filled in by deduction determined through domain understanding. Single imputation methods such as mean and median are applied to continuous features such as blood mineral profiles and respiratory tests by analysing the standard deviation of feature. For some features, forward filling or informative imputation is applied where appropriate (see Supplementary 6.4.2) . Two patients who had clinical assessment dates prior to birth data are also removed. Data for 55 patients are left in the clinical data after cleaning. The missingness threshold is set to $\geq 40\%$ in the second cycle. The datasets are combined by patient ID and date to produce a single dataset and the columns with missing values above the threshold are removed. The remaining missing data is informatively imputed to understand whether it is of clinical significance.
The data is processed to make it compatible for survival analysis. The death of a patient is taken as the event of interest to model the patient survival. The age of patients at death is calculated using the birth and death date in years and the age at last clinical assessment is calculated for patients with unknown death date. An event censoring column is created to denote patients that experienced death and patients that are right-censored at the last clinical assessment date.

NPMDS Data The NPMDS data with 20 patients had repeated clinical features across three age groups. These repeated features are combined after discussions with clinicians and features with $\geq 50\%$ missingness are removed. Informative imputation is performed on the remaining missing values.
The data is processed to make it compatible for survival analysis. Three clinical outcomes; time to death, severe feeding inability and severe immobility are taken as events of interest for survival analysis from this data. The age of patients at death is calculated and for patients with unknown death date, the age at last clinical assessment. The data is processed again to calculate the age when paediatric LS patients could not feed by mouth and age when patients became completely immobile. Three event censoring columns are created for each clinical outcome to denote patients that experienced the events and patients that are right-censored.

3.4 Exploratory Data Analysis

Clinical Data Descriptive statistics such as mean, median, standard deviation and ranges are analysed for continuous features. Simple exploratory data analysis is performed to visualise distribution for some features. The presentation and diagnosis age distributions for patients are analysed to understand the age when patients displayed LS symptoms. The genetic mutation frequencies are also visualised to understand to top genetic mutations in LS. Analysing developmental regression in children with LS is important to understand disease progression. It is postulated that children with LS move down the weight by age percentiles as the disease progresses.Therefore, ages of children under 10 years are calculated and converted to weight by age percentiles using WHO child growth standards data[16]. The trends in weight by age percentiles versus age for each patient are analysed using line plots. Kaplan-Meier estimator (KME) [8] is visualised to understand the LS patient survival times for the clinical data. The probability of death event is measured using the number of patients at risk and number of patients who experienced death at a specific time. These probabilities are then multiplied producing the final estimation of patient survival. The median survival time of patients is estimated to show the age at which 50% of LS patients experienced death.

NPMDS Data Several KME are visualised to understand if moderate and severe stages of some clinical features affected a patients time to death. The effects of feeding ability, mobility and vision on patient survival is analysed through the KME plots. Significance of log-rank test is used to test hypothesis that there is no difference between the two patient groups for each of the four clinical features on the probability of a patient survival. Two additional KME plots are created to measure the age and probability of patients experiencing severe feeding inability and immobility. The median ages of patients are estimated to show the age at which LS patients had a 50% probability of experiencing severe feeding inability and immobility.

3.5 Data Preparation

Clinical Data Feature engineering is carried out to transform the covariates for modelling. In the first cycle, features such as sex, genetics and delivery methods are dummy encoded. Ordinal encoding is applied to birth weight feature and binary encoding to phenotype and histochemistry features such as deafness, hypotonia, developmental delay and mitochondrial complex I-IV. 49 covariates are produced after the clinical features are transformed.
Features containing genetic mutations and patient sex remained after removing the columns with more than 40% missingness. Dummy encoding is applied these features producing 24 covariates in the second cycle.
The clinical data is then randomly divided into training and validation sets in a 80/20 % split with 44 and 11 patients respectively in each dataset.

NPMDS Data The data is already ordinally encoded to show the different clinical ratings for each feature. Dummy encoding is applied to the sex feature in the NPMDS data. For the both the clinical events investigated the data is randomly divided into training and validation sets in a 70/30 % split with 14 and 6 patients respectively in each dataset.

3.6 Machine Learning Models

Cox Proportional Hazards Model The data is trained using Cox Proportional Hazards (Cox-PH) [17] as a baseline semi-parametric model. Cox-PH is a cross over between Kaplan-Meier estimator and multiple linear regression. We utilized this model to the predict survival probability and determine the relationship between clinical features of LS patients and survival using the clinical data. 51 and 24 features are passed as covariates to train the model and calculate the hazard function in the first and second cycle respectively. This model measures survival through the hazard function which calculates the risk of LS patients experiencing death. A ridge penalty value of 0.1 is applied to the model to assist the model in converging and avoid multicollinearity. The coefficient values are computed to understand influence of the various clinical features on patient survival.

T he data is scaled feature-wise to further enhance the fit of model. A regularized Cox-PH model with elastic net (Coxnet) is applied to select features that are most influential on LS patient survival. The data is trained using 5-fold cross-validation to obtain a more stable model. The elastic net hyperparameter, α is tuned from a range of values. The best performing model with the optimum hyperparameter is selected and the non-zero coefficients are analysed.

Cox-PH and Coxnet with hyperparameter tuning and 5-fold cross-validation is also applied to the NPMDS data in two cycles to investigate the relationship between probability of severe feeding inability and immobility with other clinical features.

Accelerated Failure Time Model (AFT) [18] is selected to quantify the relationship between clinical features and LS patient survival times. This is a regression model that analysed which features can potentially accelerate or decelerate the time to death for LS patients. A ridge penalty of 1.0 is applied to the coefficients to deal with multicollinearity given the high-dimension of features. In this study, there are 36 patients who are right-censored in the clinical data. Therefore, to eliminate any potential bias by the low number of uncensored patients, each patient is weighted by the inverse probability of censoring (IPC) to effectively take right-censored data into account and improve the efficiency of the model. The coefficient values are then extracted to understand the most prominent coefficients.

AFT with IPC ridge penalty is also performed on the NPMDS data in two cycles to investigate whether other clinical features could accelerate or decelerate the age of patients with severe feeding inability and immobility.

Componentwise Gradient Boosting This model is chosen as an ensemble method instead of the stagewise gradient boosting (GB) model to improve the accuracy of LS patient outcome[19]. Although, the stagewise GB had the best prediction results in the paper by Nemati *et al*[10], the componentwise GB model allows interpretability of clinical features and patient survival times. This model evaluates the clinical feature

coefficients by updating the coefficients one at a time. The Cox-PH partial likelihood loss function is used to optimise the gradient. The aggregated coefficient values are then computed for 49 and 24 covariates respectively.

Componentwise GB is also used to train the NPMDS data with the two clinical outcomes; severe feeding inability and immobility. The aggregated coefficients are computed to understand which features most influenced the outcome age and probabilities.

Random Survival Forest Survival analysis based random forest model [20] is used to train the clinical and NPMDS data. Random survival forests builds trees by ranking the clinical features using the magnitude of log-rank test as the splitting criteria.
The clinical data is trained by setting the hyper-parameters to 50 trees, a minimum of 8 patients to split at an internal node and a minimum of 4 patients to be at a leaf node in the first and second cycle. At the leaf nodes, the cumulative hazards are calculated and is grown until each terminal node has a more than or equal to a patient clinical event count.

50 trees with minimum of 6 patients to split at an internal node and 4 patients at a leaf node is are used to train the two clinical outcomes using the NPMDS data.

Then, permutation is used to measure the feature importance of the clinical features using the validation data. This is an alternative to the standard feature importance in survival that measures the decrease in evaluation score; concordance-index when a feature is removed.

3.7 DeepHit Model

The DeepHit model developed by Lee *et al*. is adapted in this analysis to provide a scalable and more robust architecture to predict LS patient survival given its efficacy in predicting survival using healthcare data[13]. DeepHit is chosen in this project as it does not assume any underlying stochastic process between the covariates unlike the previous models that make strong assumptions. Instead, it learns the covariate distribution and relationship with survival times as the model is trained. This model contains a two hidden networks with the first being a shared sub-network and a group of cause-specific sub-networks (see Supplementary Figure 6). The number of cause-specific sub-networks can be modified according to the number of possible events of interest /competing risks. We used DeepHit to model the LS clinical data with one event of interest, death of a patient in the second and third cycle as shown in Figure 2. First, the hyperparameter tuning is

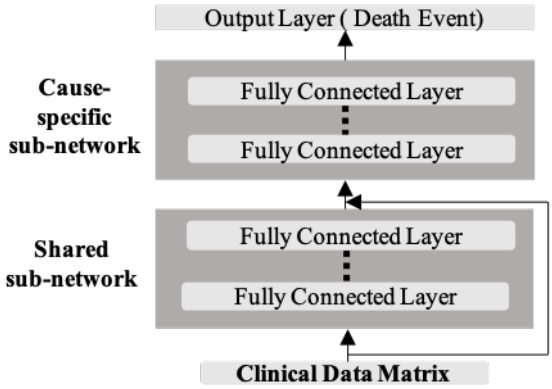


Figure 2: shows adapted DeepHit network for Death event of Leigh syndrome patients

performed with one training validation split and four random search combinations of hyperparameters. The optimized parameters are saved and applied to train the model. In the second cycle with 49 covariates, the shared sub-network contained 300 layers with 5 exponential linear unit (ELU) activation functions and the cause-specific sub-network had 200 layers with 3 ELU activation functions after tuning and optimization. In the third cycle with 5 selected covariates, the shared sub-network contained 100 layers with 3 Tanh activation

functions and the cause-specific sub-network had 300 layers with 1 Tanh activation functions. The final output layer of DeepHit produces a vector of values with survival probabilities of LS patients at different ages within observation times of this study using a softmax function.

3.8 Feature Selection

Feature selection is performed on the high-dimensional clinical data to reduce the complexity of the model. Selection of a subset of features could potentially improve the accuracy of the model and enable easier interpretation of covariates and patient survival relationship. The top 5 common covariates with high absolute coefficients values or high feature importance ranking from the ML models in the second cycle are selected. These features are then used to train the ML and DeepHit model in the third cycle (see Figure 1).

3.9 Model Evaluation

The primary metric used to evaluate the models is the concordance index (C-index)[21]. C-index demonstrates the ability of survival models to rank the survival or clinical event times correctly based on the patient risk scores. Therefore, it computes the proportion of patient pairs where a patient with a higher survival time or longer time to clinical event has a higher probability of survival or higher probability of not experiencing the clinical event as predicted by the model. The Brier scores[22] at 4 years, 10 years and 17 years is used as a secondary metric to analyse the prediction accuracy of probabilities within the defined time of the data. For the NPMDS data, the integrated Brier score was used to measure the prediction accuracy. The time-dependent area under the ROC curve (AUC) is then compared to understand the sensitivity and specificity of the model. This cumulative/dynamic AUC measures the ability of a model to differentiate between patients that experience an event at specified time or after a specified time.

The DeepHit model was evaluated differently compared to the ML models. For this model, the patient ages at which model evaluation is performed can be defined. Therefore using this architecture we compute the time-dependent C-index at 4, 10, 17 and 30 years. The 30 year C-index was taken as an overall C-index metric to compare with the other models.The Brier score at 4, 10, 17 and 30 years were also calculated to assess the ability of the model to predict survival at various times in a patients life.

4 Results

4.1 Exploratory Data Analysis

Exploratory data analysis was performed to visualise distribution of some features. The highest number of patients diagnosed with LS is at 2 years (Figure 3a) and generally most patients are diagnosed between infancy to 2 years. 5 patients in this cohort presented with LS symptoms 1 year of age and most patients presented LS symptoms within the first 2 years as shown in Figure 3b. This shows that this cohort also follows the same LS onset pattern as with most LS patients at population level[3]. The top three genetic mutations found in the patient cohort are SURF1, NDUFV1 and m.8993T>G mutation as shown in Figure 3c. SURF1 encodes a mitochondrial protein that assists in the assembly of a crucial complex, COX required for energy metabolism[23]. Mutations in this gene is one of the most common genetic causes of LS. Mutation in NDUFV1 subunit causes deficiency of the mitochondrial complex I [24]. The m.8993T>G is a mutation on the mitochondrial genome that is causes deficiency in mitochondrial ATPsynthase, the most important complex that makes the energy producing molecule, ATP. It is observed in the birth weight chart that most LS patients weighed between 3000 to 3500 g at birth (Figure 3d).

The KME in Figure 4 shows survival times and probabilities of patients with LS from the clinical data. At 10 years the plot shows that 8 patients are at risk of dying from Leigh syndrome with a probability of 79%. The shaded area in Figure 4 shows generally shows wide 95% confidence interval for estimating patient survival especially after 17 years. The median age of patients cannot be computed as patients were censored at survival probability of 52%. Figure 5a shows KME of time to feeding inability where 50% of patients require complete enteral feeding by 7.6 years. The 95% confidence interval in Figure 5a increases as the age

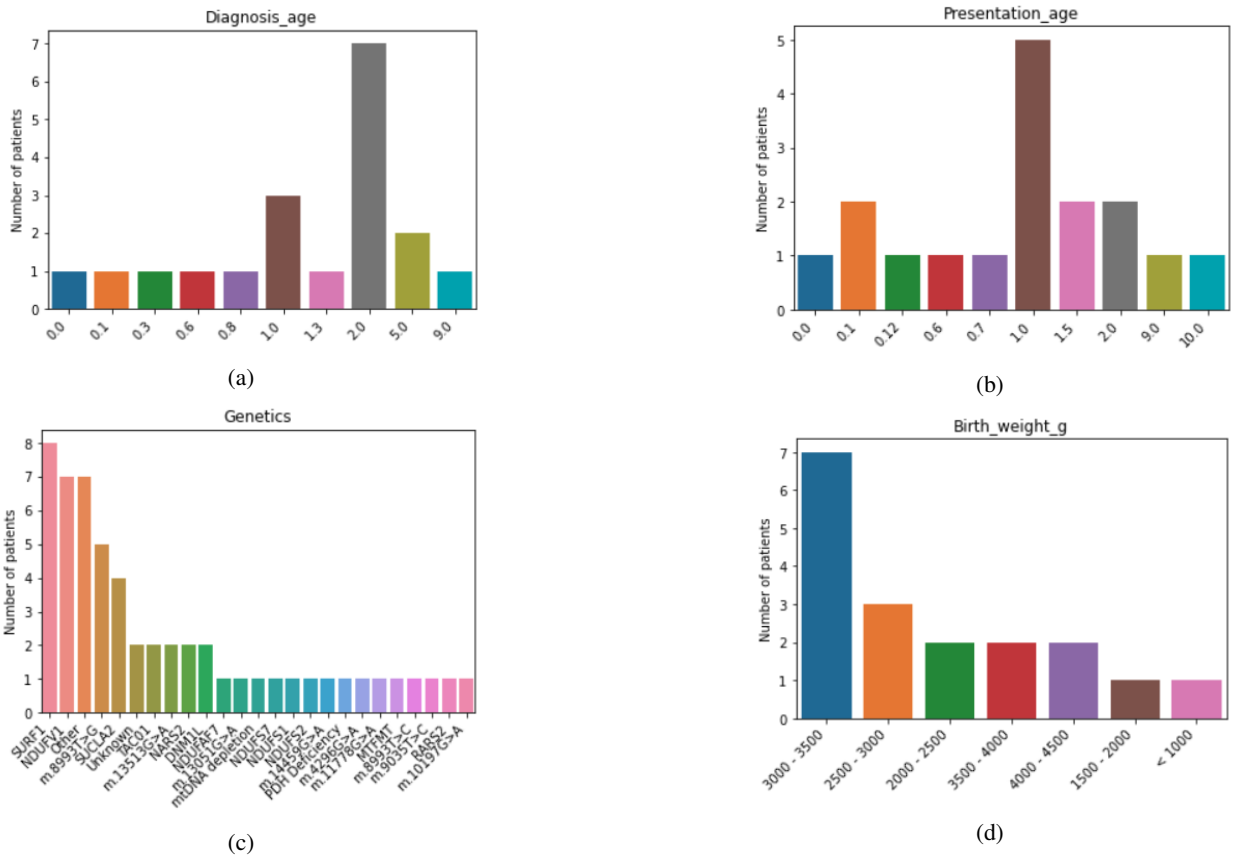


Figure 3: Distribution of some features

increases. Figure5b shows that 50% of LS patients in this cohort were wheelchair bound or immobile by 6.8 years. The 95% confidence interval is about the same as the age of patients increases in Figure 5b.

4.2 Performance Metrics

Clinical Data In the first cycle with 49 covariates the Coxnet, Accelerated Failure Time and Componentwise GB model had similar levels of performance with C-Index of 0.8888 outperforming the baseline Cox-PH model as shown in Table 1. The Coxnet and Componentwise GB model in cycle 1 had similar Brier scores at different ages and the time-dependent AUC for these models was also 1.0 from 5 years to 25 years. The Cox-PH, Coxnet and Componentwise GB model had C-index of 0.9444 in the second cycle with 26 covariates. However, the baseline Cox-PH model had the best Brier scores that were lowest across all time points when compared to the other models. The DeepHit model had a good C-index of 0.90 but the Brier scores of the model were higher than that of Cox-PH. In the 3rd cycle, the Cox-PH model trained with 5 covariates was evaluated as the best performing model for this study. It had the lowest Brier scores compared to all other models and an all time AUC of 1.0.

NPMDS Data The Coxnet model has perfect C-index of 1.0 with an integrated Brier Score of 0.3954 when predicting age at feeding inability (Table 2. The C-index of the other models in Table 2 are highly variable ranging from 0.3 to 0.6. The random survival forest model performs the best compared to the other models when predicting age at immobility with a C-index of 0.5 and integrated Brier score of 0.2014 (see Supplementary 6.5 for detailed results).

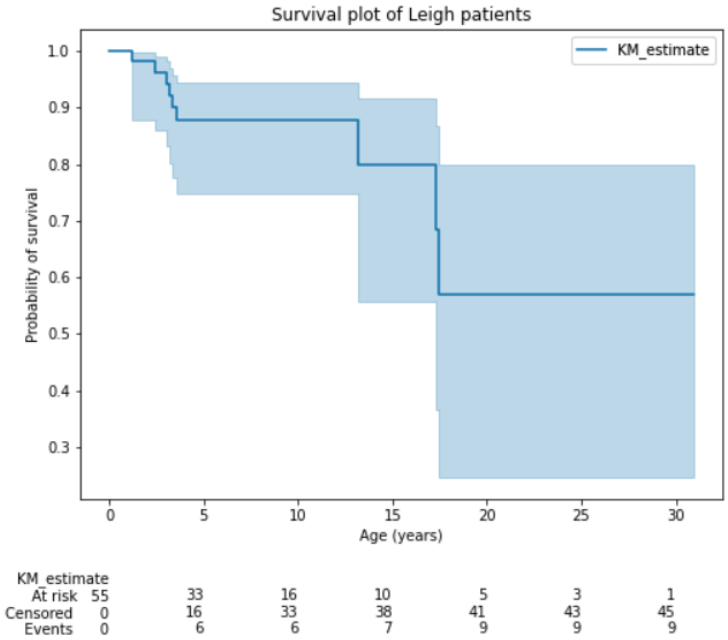
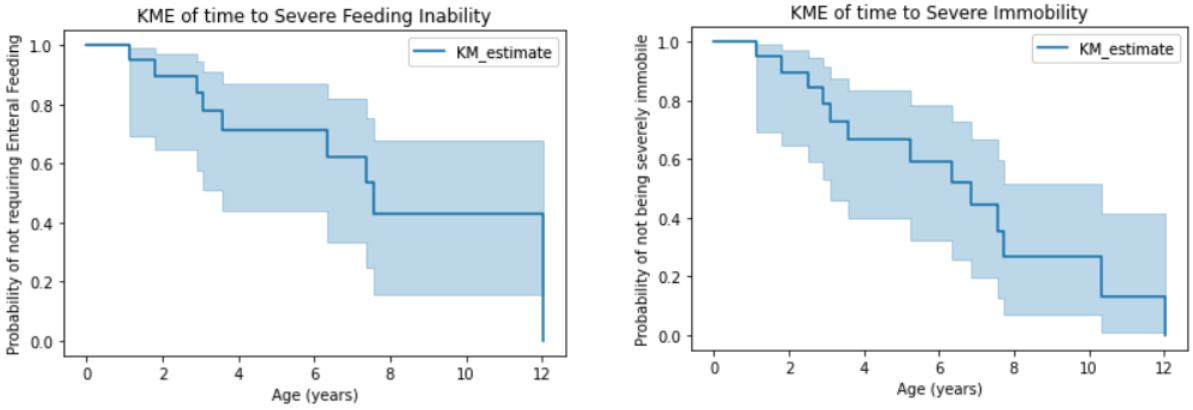


Figure 4: Kaplan Meier of 55 Leigh syndrome patients from the training set



(a) (b)

Figure 5: KME of two clinical outcomes investigated

4.3 Discussion

The coefficients of the best performing CoxPH model for the clinical data demonstrates that the 5 clinical features have strong relationship with the hazard rate of LS patient survival (Table 3). The strongest relationship is seen with the NDUFS1 mutation, where this mutation increases the hazard of a patient dying from LS at any given age by a factor of 3.1283. The NDUFS1 mutation can cause disruption in the electron transport chain which is vital to produce ATP [25]. This could potentially decrease energy production therefore increasing risk of patient dying from LS. Similarly, the mitochondrial 10197G>A mutation and SURF 1 mutations could increase the risk of dying by causing inefficiencies and problems in energy production. The sex feature has the fourth strongest relationship with the survival hazard by a value of 1.707. It is inferred that patients who are males have lower chances of survival. This CoxPH model which is recommended to predict patient survival times assumes that every patient has the same hazard rate. The effect of the covariates on the survival times is also assumed to be constant with time. However, the effect of some covariates such as weight, blood measurements and mobility could change with time. In the

Table 1: clinical data performance metrics

	Model	C-index	4 year Brier score	10 year Brier score	17 year Brier score
Cycle 1: 49 covariates	Cox-PH	0.3330	0.0931	0.0931	0.0931
	Coxnet*	0.8888	0.0506	0.0532	0.0363
	AFT	0.8888			
	Componentwise GB*	0.8888	0.0506	0.0525	0.0373
	Random Survival Forest	0.7778	0.0701	0.0748	0.0689
Cycle 2: 26 covariates	Cox-PH*	0.9444	0.0201	0.0152	0.0197
	Coxnet	0.9444			
	AFT	0.1667			
	Componentwise GB	0.9444	0.0472	0.0152	0.0197
	Random Survival Forest	0.8333	0.0777	0.0807	0.0778
	DeepHit	0.9000	0.3580	0.2701	0.0446
Cycle 3: 5 covariates	Cox-PH**	0.9444	0.0163	0.0090	0.0086
	Coxnet	0.1111	0.0931	0.0931	0.0931
	AFT	0.1667			
	Componentwise GB	0.9444	0.0472	0.0489	0.0340
	Random Survival Forest	0.9444	0.0673	0.0698	0.0713
	DeepHit	0.9000	0.2897	0.1627	0.0282

Table 2: NPMDS performance metrics

	Model	C-Index	Integrated Brier
Feeding Inability	Cox-Ph	0.6666	0.4142
	Coxnet*	1	0.3954
	AFT	0.3333	
	Component GB	0.6666	0.3500
	Random Survival Forest	0.5	0.2226
Immobility	Cox-Ph	0.4545	0.3592
	Coxnet	0.4545	0.2149
	AFT	0.6363	
	Component GB	0.2272	0.2354
	Random Survival Forest*	0.5454	0.2014

second and third cycle (Table 1) the AFT model is the worst performing model. For the NPMDS data, the AFT model is the worst performing model when the event investigated is feeding inability. The AFT model is a parametric regression-based model that makes strong assumptions about the covariate and outcome relationship. Therefore, this assumption that a covariate can either accelerate or decelerate the time to an event by a constant might not be true. The DeepHit model was applied to clinical data to understand if this model works to predict clinical outcomes of rare diseases. This model is a good was to learn the covariate distributions without making any assumptions. However, given the small dataset in this study, the model high C-index and low Brier scores could be a result of overtraining.

Features	Coefficient
NDUFS1	3.128
m.10197G>A	2.943
SURF1	2.825
Sex	1.707
m.8993T>G	1.436

Table 3: coefficients of Cox-PH model with 5 covariates

The coefficients of the best performing Coxnet model for the NPMDS data demonstrates that vision as rated by the patient has the strongest relationship with the hazard of requiring feeding support, then development of a patient followed by problems in communication (Table 4). A potential interpretation of the relationship between poor vision feeding inability could be that the neurodegeneration of patients reached a certain stage where both vision and feeding capacity of patients is affected severely. This could also explain the relationship for developmental regression and communication with the feeding inability. Random survival

forest model shows that the most important feature in determining patient immobility is the development of a patient (Table 5) . This relationship shows that patients that do not reach developmental milestones will have mobility issues. Features such as gastrointestinal issues, seizures respiratory problems, sex and growth of patient also are important to determine the age at immobility. The random survival forest model that performed best to predict age at immobility using the NPMDS data shows random prediction of patient sequence of experiencing the event with a C-index that is close to 0.5. Therefore, more analysis is required to recommend this model.

Feature	Coefficient
Vision - (patient rating)	0.1957
Development	0.1229
Communication	0.0147

Table 4: coefficients of Coxnet model for Feeding Inability

Feature	Importance
Development	0.2303 ± 0.3033
Gastrointestinal	0.1576 ± 0.1236
Seizures	0.0970 ± 0.1403
Respiratory	0.0970 ± 0.1688
Sex	0.0364 ± 0.1599
Growth	0.0121 ± 0.1305

Table 5: coefficients of random survival forest model for Immobility

5 Conclusion

Leigh syndrome is a neurometabolic disorder that has a large array of outcomes and multigenic causes that affect the disease progression. We have used two Leigh syndrome datasets with 55 patients and 20 patients to train and evaluate machine learning survival models that were selected to predict clinical outcomes and determine relationships between covariates and outcomes. We propose that machine learning survival models such as Cox-proportional hazards model can be used to predict patient survival, time to feeding inability and immobility. We have also shown that patient gender and some genetic mutations such NDUFS1, m. 10197G>A, SURF1 and m.8993T>G greatly influences a patients survival. Additionally, a patients vision, development and communication could affect a patients time to feeding inability.

Future Work Data with a higher number of patients could improve the performance of the models is this study. The clinical outcomes investigated with the NPMDS data will benefit from more paediatric patient records to give better predictions. Data that is more complete with regular and coherent records across different times of patient information and clinical assessments could reduce the missingness in the data. Thus, more features that are potentially vital to predict patient survival and clinical outcomes surface when trained using the models in this study. In future, covariates that have a time-dependent effect could be modelled using time-varying cox-proportional hazards model since the data is in a longitudinal format. The DeepHit model could be scaled up using data with more patients and multiple competing clinical outcomes could be investigated. For example, the time to feeding inability or immobility could be investigated together using the DeepHit architecture. With more robust data and improvisations suggested, the results of this study could be greatly enhanced.

6 Supplementary Document

6.1 Hypothesis

We hypothesise that machine learning and deep learning architectures incorporated with survival analysis can accurately predict clinical outcomes for patients with Leigh syndrome. We believe that these models can

also demonstrate the relationship between selected features and the clinical outcome of interest. This could potentially help clinicians to understand risk factors for an outcome and also an understanding of the chances of a patient experiencing a clinical outcome.

6.2 Project Objectives

This project will ultimately aid clinicians in their decision-making to prescribe the most appropriate treatment or care. Therefore, the following objectives are listed to achieve this:

- Research on the integration of survival analysis in machine learning and its application on disease progression by reading research publications that are relevant.
- Understand the processes and methods used to collect LS patient data through data analysis and interaction with clinicians.
- Wrangle the raw clinical data by combining the various features with duration of measurement for survival analysis.
- Perform exploratory data analysis to understand the distribution of features using bar charts and survival plots.
- Train the Leigh syndrome data with selected survival analysis models that are suitable for the size of the data and research domain.
- Determine features that influence progression of the disease through domain-driven approach via interaction with clinicians and statistical approaches.
- Evaluate the results of the model using survival analysis metrics to determine the best performing model.

6.3 Background

6.3.1 Related Work

Several ML approaches were reviewed apart from survival analysis to understand the applications of ML in studying clinical outcomes and patient survival.

Classification and Clustering Methods ML models have been extensively used in medical settings where the outcome is either a discrete variable or clusters of an expected outcome. Studies have shown that algorithms such as logistic regression, random forests and support vector machines have good predictive performance to predict binary outcomes of patient such as survival with heart failure[26], malignancy and survival of breast cancer[27, 28]. Deep learning architectures such as artificial neural networks have also been used to classify and cluster cancer types, recurrence and survival using genome expression, clinical features or medical imaging data[29]. Traditional classification and clustering techniques are well-researched methods that have can produce good predictive performance. These models has the potential to be deployed as diagnostic tools in medical settings however it often lacks reliable prognostic information such as probability of the event or time to the event.

Time-series Modelling In clinical settings, the ability to forecast medical events over time is crucial for clinicians to make treatment decisions and set patient expectations. Therefore, time-series analysis can be an appropriate method for analysis longitudinal medical data. Various time-series models such as linear dynamics systems and hidden markov models have been used to predict trajectories of patients when admitted to hospital, detection of sepsis in newborns[30]. In some studies, deep learning frameworks such as continuous-time autoregressive (CAR) model, recurrent neural network (RNN) and time-aware long short-term memory (T-LSTM) were applied to predict intensive care unit (ICU) mortality rate, forecast Alzheimer's and COVID-19 disease progression[31, 32]. Medical data for a particular diseases is highly composite in nature with a mixture of multivariate clinical features and longitudinal data. Longitudinal data is often irregularly sampled with high amount of missing values. For the purposes of this project, time-series

analysis will not easily fit the irregular sampling and high-dimensionality of this data. Many LS patients are lost to follow-up which cannot be taken into account the aforementioned ML models.

6.3.2 Model Descriptions

Survival analysis was developed in 1662 century by John Graunt, a statistician which at the time was called lifetables[33]. This method was used to investigate mortality rates for various diseases which was then termed survival analysis. This branch of statistics however has been successfully applied in other fields such as economics, engineering and finance.

A package called lifelines was built by Cam *et al*. [34] to run ML survival models in python or R. This package contains an array of non-parametric, semi-parametric and parametric survival models. Recently, another package was built on top of scikit-learn that would allow the adaptation of survival analysis into ML models whilst keeping the basic functionalities of scikit-learn[35]. Whilst lifelines offers great statistical outputs of survival models, scikit-surv offers more common ML based architectures for survival analysis such as random forests and boosting models. Additionally, scikit-surv has seamless and versatile functionalities like scikit-learn.

Kaplan-Meier Estimator The most common form of non-parametric survival analysis model is the Kaplan-Meier Estimator (KME). KME estimates the survival function in stepwise intervals of time given the observed event duration[8]. KME can visualise survival times and probabilities well, but it cannot accommodate the effects of covariates on the survival times and probability. The KME is defined using the function below:

$$S(t) = \prod_{t>t_i} (n_i - d_i)/n_i \tag{2}$$

where d_i are the number of events investigated at time, t and n_i is the number of patients at risk of experiencing the event prior to time,t.

Cox-proportional hazards Survival analysis can also model the relationship between features and time to event using regression with some modifications. A semi-parametric regression model that is widely used for multivariate data is Cox-proportional hazards model (CoxPH) [17]. This model is defined by a hazard function that is a derivative of the survival function:

$$\lambda(t|x) = \lambda_0(t)\, exp(\beta_1 x_1 + ... + \beta_n x_n) \tag{3}$$

The hazard function outputs the probability of the event occurring at a certain time,t given the event has not happened up to time t. This function is take in covariates ($x_1....x_n$) with coefficients ($\beta_1.....\beta_n$) that measure that strength of the covariates within the exponential component. The baseline hazard is represented as $\lambda_0(t)$ which is the hazard when all covariates are 0. This function can also be converted to a multiple linear regression model by taking the log of the hazard function to compute the linear combination of covariates and the baseline hazard being the intercept that varies with time. Therefore, CoxPH allows us to understand the magnitude of influence a particular clinical feature has on the rate of the event investigated at a particular age of patient; hazard rate. A positive coefficient for a certain clinical features might suggest a worse prognosis and a negative coefficient potentially is a protective feature on the event of interest. This model makes some assumptions about the covariates. It assumes that the covariates have an exponential effect on the survival. In this study, the model will assume that different patients have the same hazard rate[36]. The CoxPH model is vulnerable to high-dimensional data as it can easily cause overfitting. This model is also not able to handle multicollinearity between covariates. Therefore, to make the model more efficient regularization terms can be added to the model.

Accelerated Failure Time (AFT) Model AFT model is parametric regression model that assumes that a particular covariate can either accelerate or decelerate the time to an event of interest by a respective coefficient value. The model is generally defined by the function below:

$$log\, T = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n + \epsilon \tag{4}$$

In this model, coefficients $(\beta_1, .., \beta_n)$ is added to the covariates $(x_1 ... x_n)$ and ϵ is the independent random disturbance term. Log is applied to the regression function as survival times or ages are positive. This particular model computes the accelerated failure time with the samples weighted by the inverse probability of censoring . Ridge penalty is applied on the coefficients, $(\beta_1, ..., \beta_n)$ and censoring is assumed to be random.

Componentwise Gradient Boosting This model is a boosting model that evaluates the coefficients either by fitting the gradient by updating the coefficients one at a time[19]. This model optimizes many loss functions such and feeds it into the input matrix using regression methods such as Cox-PH or least squares. It combines predictions of multiple base learners producing a powerful model. The ability of this model to compute the aggregated coefficients can enhance interpretability of covariates and the event of interest for LS patients.

Random Survival Forest Random forest [20] is an ensemble method that builds multiple decision trees and combines them to get a more accurate and stable prediction. Normally, random forests can be used for classification and regression methods for machine learning. Random forests splits features according to a specified splitting criterion such as information or Gini index that will rank the features with the most important feature at the top of the tree and the least at the bottom. However this model is modified for survival analysis. For survival analysis, this model outputs cumulative hazard function through multiple trees[20]. This way the proportional constraint of Cox-PH can be avoided. The splitting criterion for feature values is the log-rank test.

DeepHit The DeepHit model developed by Lee *et al*. is a multi-task network that can also accommodate a single task/event [13]. The model has a shared sub-network and a group of cause-specific sub-networks that

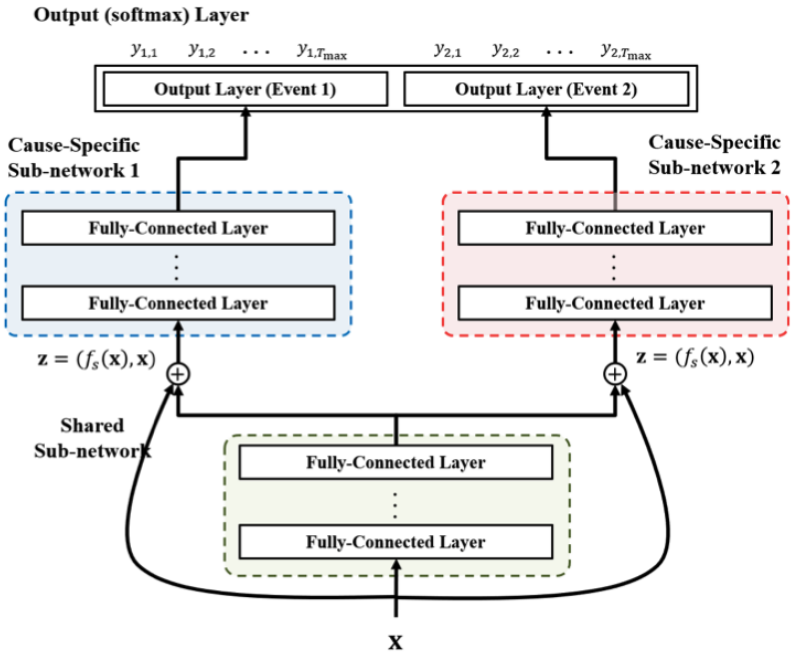


Figure 6: shows DeepHit architecture[13]

can be specified by the number of events investigated as shown in Figure 6. This model is a type of multi-task learning model with some modifications. A multi-task learning model is a type of neural network that has the ability to learn and predict multiple outcomes using covariates that are different for each outcome and shared covariates for all the possible outcomes. The covariate matrix is connected to the shared sub-network and the cause-specific network. This additional input at the cause-specific network trains the model to learn the covariate distribution better when there are more than one event of interests. The output layers contains one

softmax function to learn to combined distribution when multiple events are present. The loss function of DeepHit can effectively handle censored data. It is a sum of the log-likelihood of the joint distribution of the first hitting time and event of interest and a cause-specific ranking loss function. The first term therefore is used to gain information about uncensored events and the second computes the censoring bias by interpreting that patients did not experience the event at the time of censorship.

6.3.3 Leigh Syndrome Data

Within the Wellcome Centre for Mitochondrial Research, a team led by paediatric neurologist Prof Robert McFarland carry out research into mitochondrial diseases. They routinely collect valuable clinical data, metadata and tissue samples from patients with mitochondrial diseases. We obtained this anonymised LS patient data from Mitocohort, a health research authority that contained information of patients from Newcastle, Oxford and UCL,London.

6.4 Methodology

6.4.1 Data Understanding

The data contained biographic information and various clinical assessments such as genotype, phenotype, histochemistry, blood results and cardiac tests stored longitudinally in different datasets with varying number of patients. The data also contained two semi-quantitative clinical rating scales designed by Newcastle to assess mitochondrial diseases. The Newcastle Mitochondrial Disease Scale for Adults (NMDSA) data had clinical information for adults and . Many features in NMDSA and NMDPS stored similar information as the other clinical data but are assessed differently. Therefore, the NMDPS and NMDSA data are analysed differently in this study.

6.4.2 Data Cleaning

Clinical data In Cycle 1, features with $\leq 60\%$ missingness were imputed using various methods as expanded in Table 6.

NPMDS data Two functions were created in python to combine NPMDS data with repeated columns across the different age ranges (refer to code). After applying the 50% threshold for missingness the remaining columns were imputed using informative imputation.

6.4.3 Technical Implementation

Data from Newcastle, Oxford and London are stored in 20 different excel sheets containing information about quantitative and qualitative clinical assessments collected sequentially over time per centre. The various excel sheets are loaded as individual dataframes using Python within Jupyter Notebook. Dataframes with same clinical assesments such as phenotype, blood results and histochemistry from the different centres are combined creating 20 dataframes using pandas library. The data was further cleaned through the two missingness threshold 60% and 40%. Of the 20 dataframes, two contained NMDSA and NPMDS ratings that are treated differently by concatenating the similar columns and setting a missingness threshold of 50%.

The ages of patients is calculated by substracting the birth date from the data of the event of interest. This results in the number of days in datetime format which is then converted to interger type and then divided by 365 to get the age in years in float type. Feature engineering and data splitting is carried out using scikit-learn library.

Exploratory data analysis of the feature distributions was carried out using seaborn [37] and matplotlib[38] packages. Kaplan-Meier estimators are constructed using lifelines library [34]. The number of risk patients and patients experiencing the event at various timestamps is also specified on the plots.

The Cox-PH model was implemented using the scikit-surv library [35]. Cross-validation and hyperparameter tuning was carried using scikit-learn for Coxnet model. The AFT model was also applied from scikit-surv using the IPCRidge algorithm. The componentwise GB algorithm and random survival forests are from

Feature	Imputation method
Histology Normal	Forward filling by ID and Informative filling
Complex I	Forward filling by ID and Informative filling
Complex II	Forward filling by ID and Informative filling
Complex III	Forward filling by ID and Informative filling
Complex IV	Forward filling by ID and Informative filling
Complex V	Forward filling by ID and Informative filling
Blood Heteroplasmy	Mean by Year and Informative filling
Weight	Forward, Backward and Informative filling
Predicted Forced Vital Capacity	Median and Informative filling
Deafness	Forward filling by ID and Informative filling
Hypotonia	Forward filling by ID and Informative filling
Developmental Delay	Forward filling by ID and Informative filling
Haemoglobin	Mean by Year
Sodium	Mean by Year
Potassium	Mean by Year
Urea	Mean by Year
Creatinine	Mean by Year
Alanine	Mean by Year
Bilirubin	Mean by Year
Alkaline Phosphatase	Mean by Year
Adjusted Calcium	Mean by Year
Neuro Slow wave	Forward filling by ID
Neuro Sharp wave	Forward filling by ID and Informative filling
ECG Normal	Forward filling by ID and Informative filling
ECHO Normal	Forward filling by ID and Informative filling
Ophthalmology Normal	Forward filling by ID
Audiology Normal	Forward filling by ID
Sensorineural loss right	Forward filling by ID
Sensorineural loss left	Forward filling by ID
Diagnosis age	Informative filling
Presentation age	Informative filling
Gestational age	Informative filling
Birth weight	Informative filling
Genetics	Informative filling
Consanguinity	Informative filling
Diagnosis basis	Informative filling
Delivery method	Informative filling
Resuscitation	Informative filling

Table 6: Feature imputation methods for cycle 1 with 49 covariates

the same libraries. The scikit-learn attribute of feature importance is not available in scikit-surv random survival forests since the log-rank test is used as the splitting criteria. Therefore, permutation from the ELI5 library[39] is adapted to calculate feature importances of the covariates.

DeepHit [13] available as a full package in R and Python however there is no documentation or API for the packages. Therefore, the source code of the entire architecture had to be implemented in jupyter notebook. The github repository was cloned and the network scripts in python files are run in the jupyter notebook. The hyperparameter and network settings are defined and the model was trained for two cycles with 24 and 5 features.

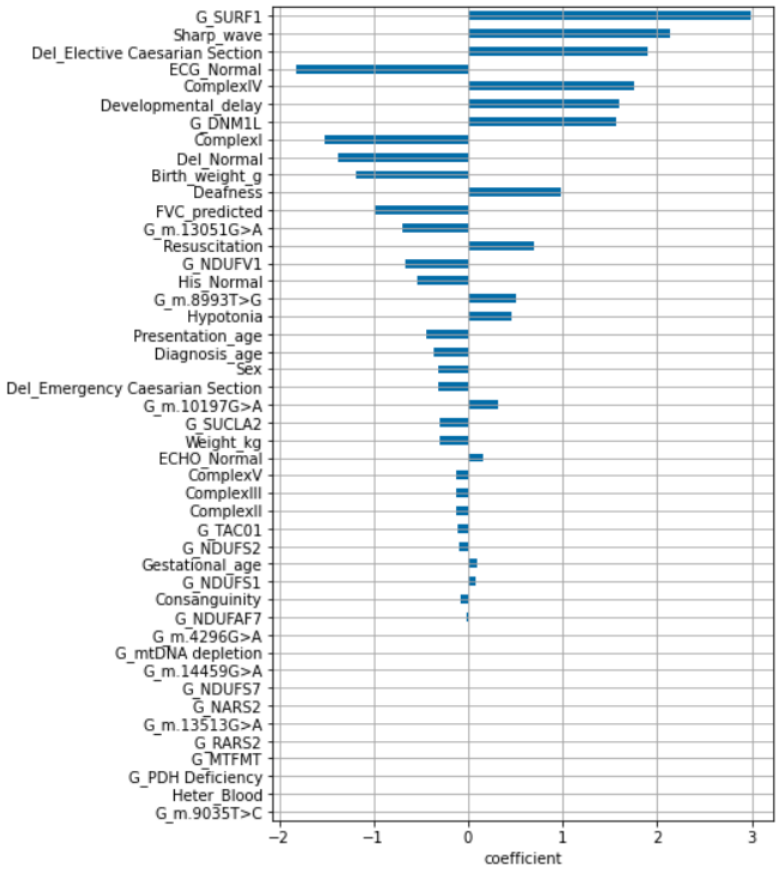
The concordance index and time-dependent AUC could easily be implement for all the models using the scikit-surv library. However, the intergrated brier score could not be computed for the AFT model as the IPCRidge algorithm in scikit-surv did not have the attribute to predict survival function.

6.5 Results

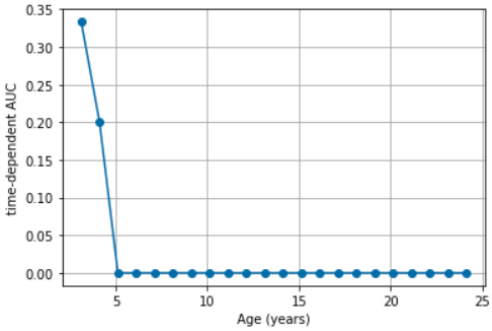
Figures 7 - 23 show the coefficient values, hyperparameter tuning for Coxnet and time dependent AUC results for each cycle and clinical outcomes.

6.6 Reflection

This project was an exciting and challenging project. It was interesting to learn about new techniques and incorporation of survival analysis into Machine Learning. I also expanded my knowledge on different applications of Machine Learning in Medicine. Leigh syndrome is a rare disease with very small number of patients which made it difficult to work out how Machine Learning can be applied for clinical outcome prediction. At the initial stage, it was a challenge to figure out the best way to combine and clean the dataset. I learnt various approaches to deal with missing data which includes using ML algorithms to impute healthcare data. Throughout this project, I also learnt the importance of interpreting ML models in clinical practice. Many ML models could give very accurate predictions but it needs to have prognostic value to be applied in clinical practice. I worked in a team of with people of different areas of expertise which enhanced my collaboration and communication skills.

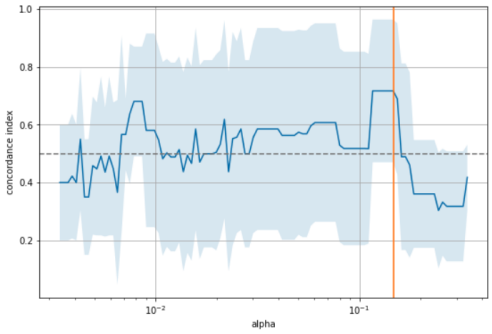


(a)

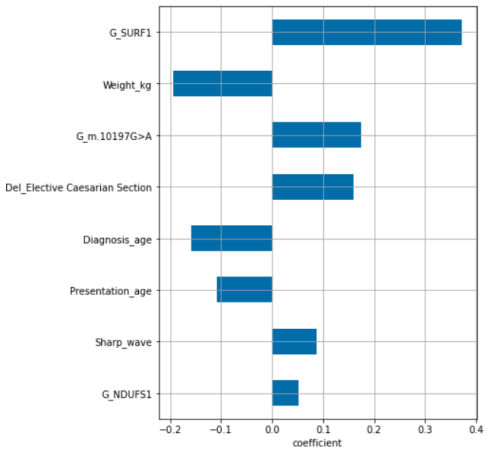


(b)

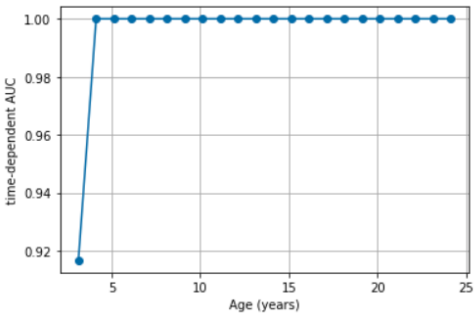
Figure 7: Cox proportional hazard Cycle 1



(a)

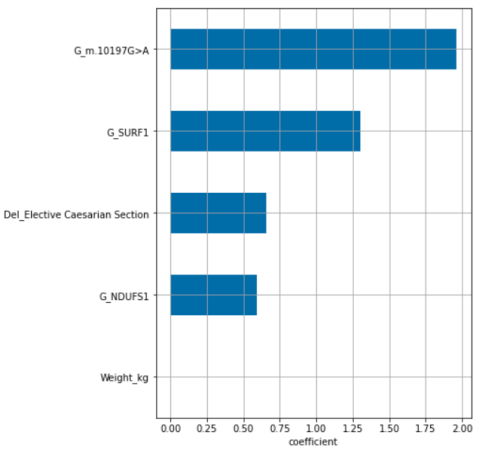


(b)

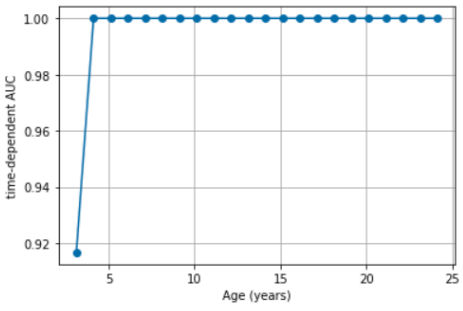


(c)

Figure 8: Coxnet Cycle 1

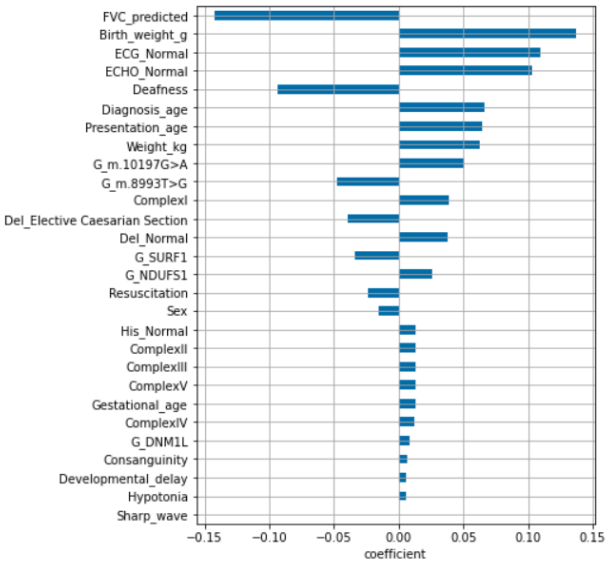


(a)

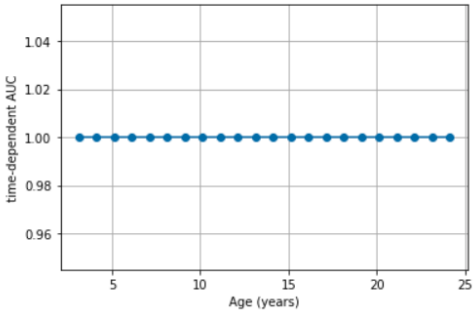


(b)

Figure 9: Componentwise gradient boosting Cycle 1



(a)

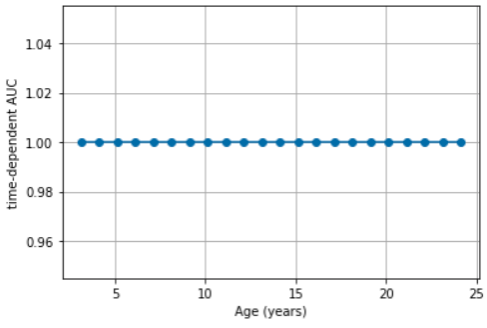


(b)

Figure 10: Accelerated failure time Cycle 1

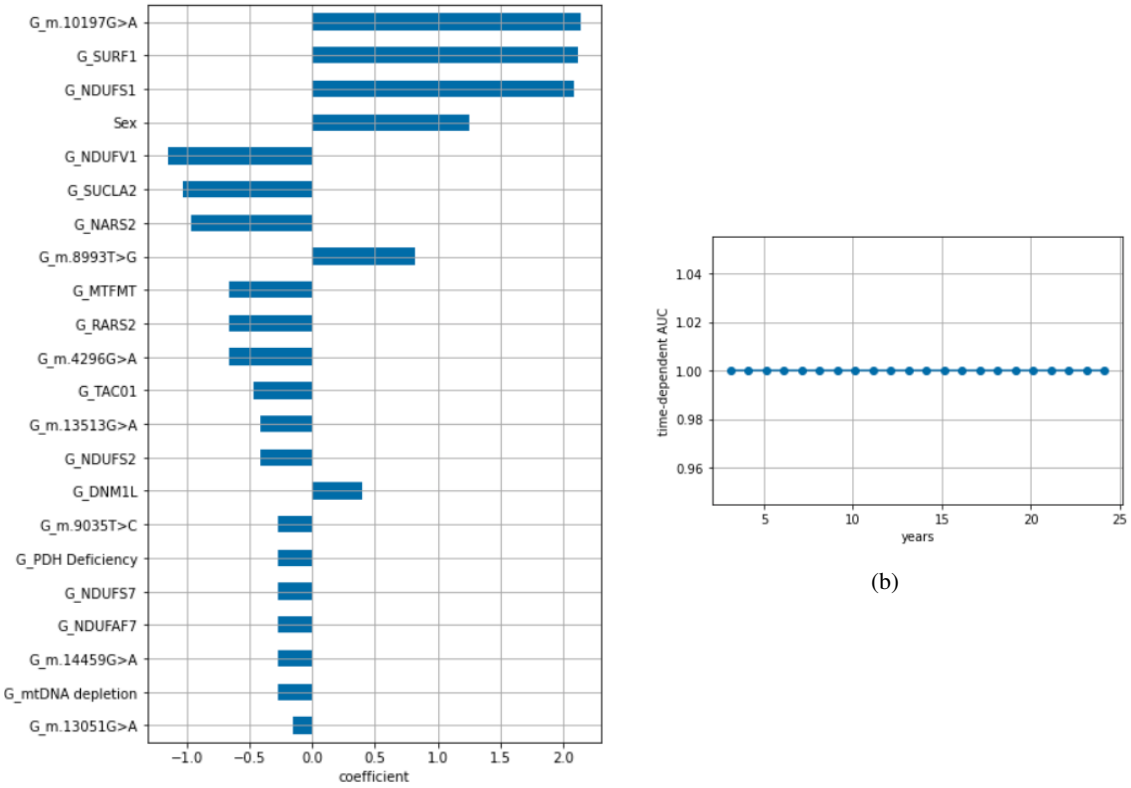
Weight	Feature
0.3000 ± 0.3823	Deafness
0.1556 ± 0.3014	Weight_kg
0.0741 ± 0.1048	G_SURF1
0.0519 ± 0.1374	Developmental_delay
0.0370 ± 0.1932	Presentation_age
0.0296 ± 0.0983	Sex
0.0222 ± 0.0889	Consanguinity
0.0074 ± 0.0554	FVC_predicted

(a)



(b)

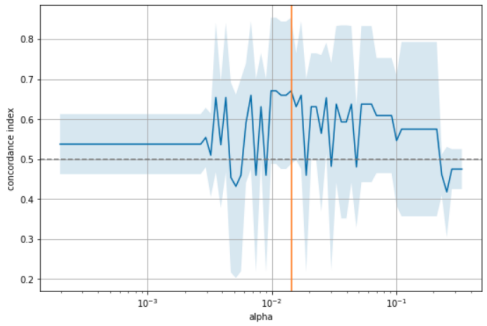
Figure 11: Random survival forest Cycle 1



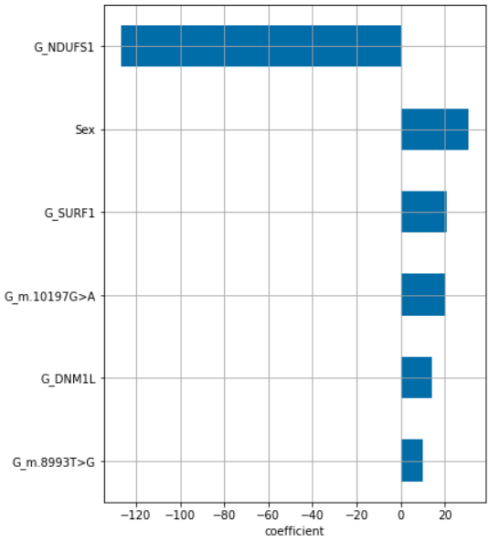
(a)

(b)

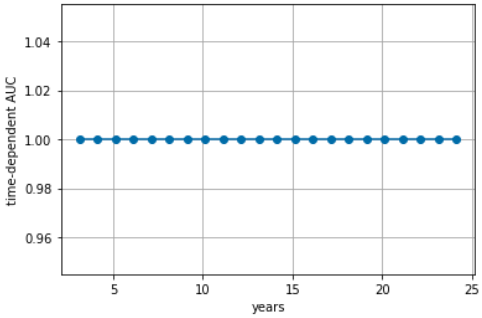
Figure 12: Cox proportional hazard Cycle 2



(a)

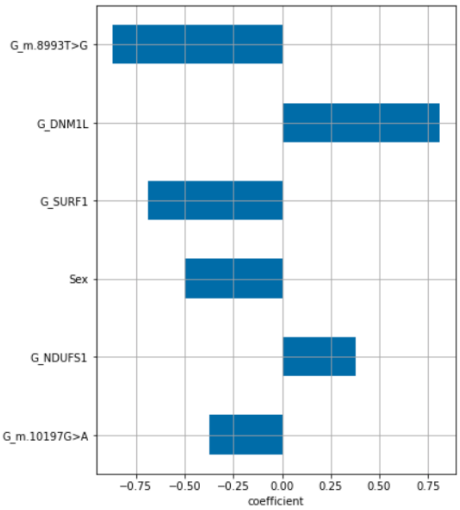


(b)

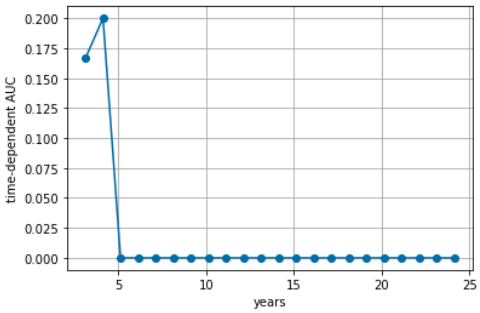


(c)

Figure 13: Coxnet Cycle 2

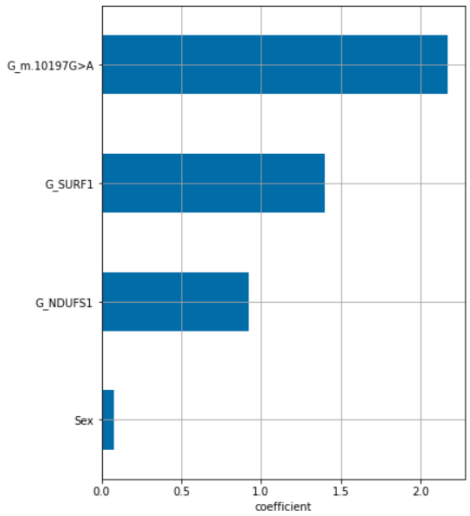


(a)

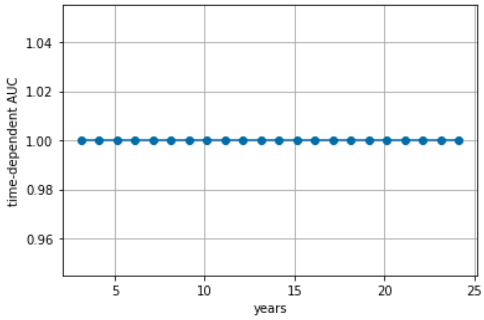


(b)

Figure 14: Accelerated Failure Time Cycle 2



(a)

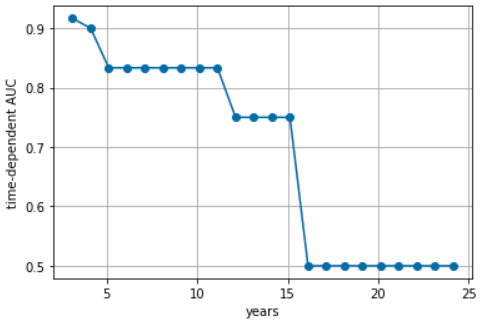


(b)

Figure 15: Componentwise gradient boosting Cycle 2

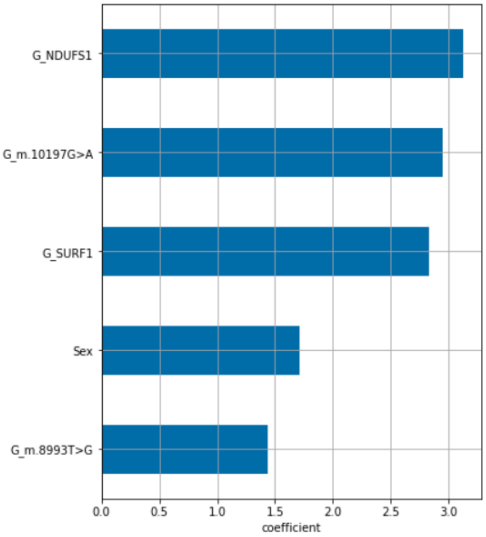
Weight	Feature
0.2556 ± 0.4441	Sex
0.1444 ± 0.6029	G_NDUFV1
0.0593 ± 0.1248	G_m.8993T>G

(a)

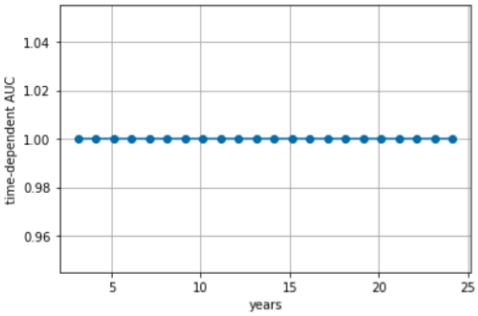


(b)

Figure 16: Random survival forest Cycle 2

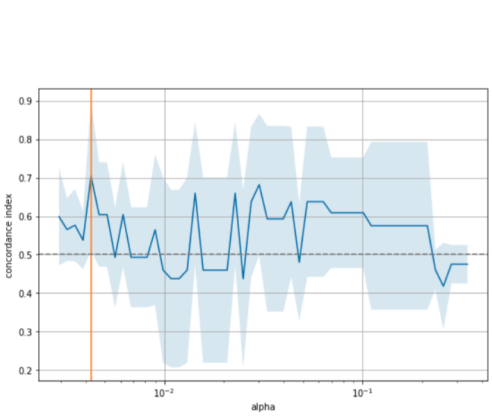


(a)

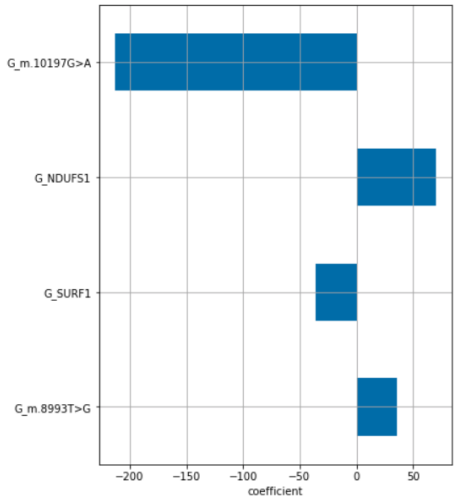


(b)

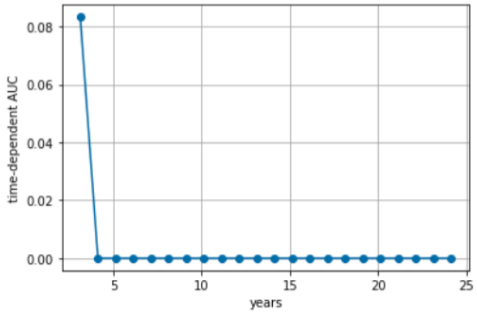
Figure 17: Cox proportional hazard Cycle 3



(a)

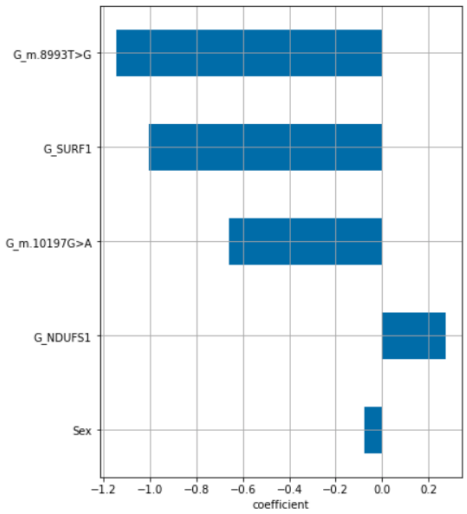


(b)

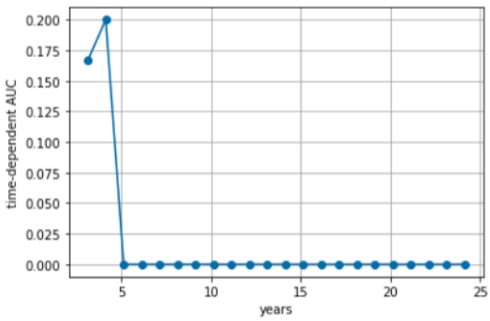


(c)

Figure 18: Coxnet Cycle 3

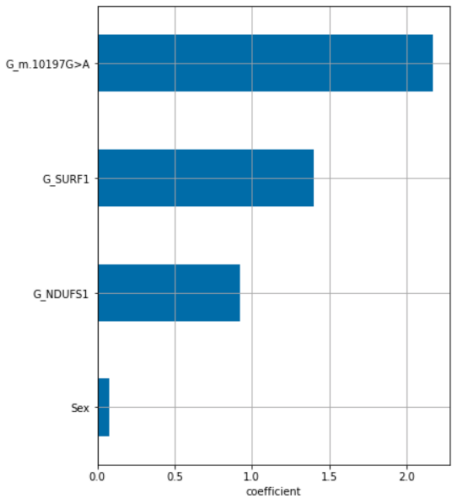


(a)

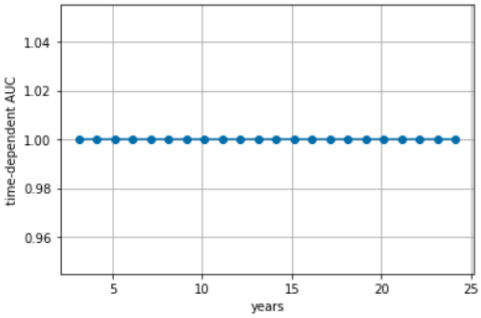


(b)

Figure 19: Accelerated Failure Time Cycle 3



(a)

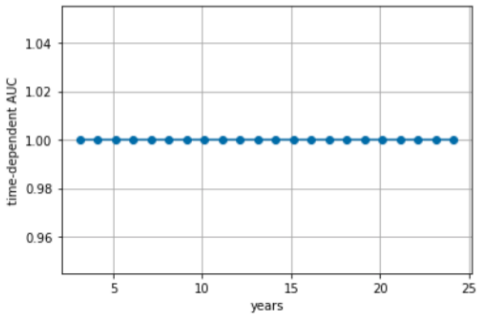


(b)

Figure 20: Componentwise gradient boosting Cycle 3

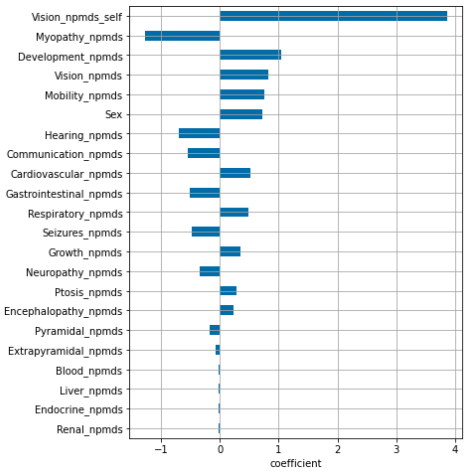
Weight	Feature
0.3333 ± 0.3828	G_SURF1
0.0741 ± 0.1325	Sex

(a)

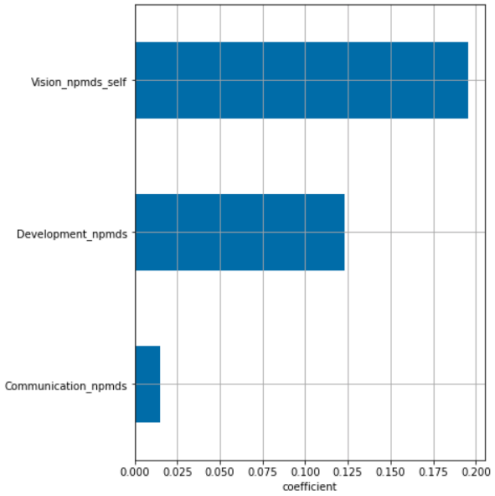


(b)

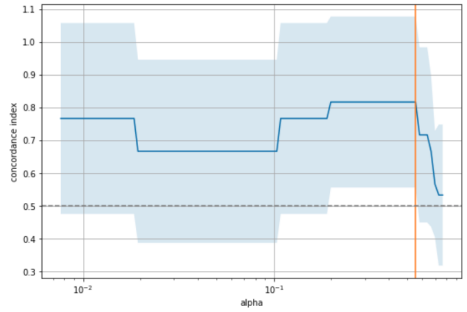
Figure 21: Random survival forest Cycle 3



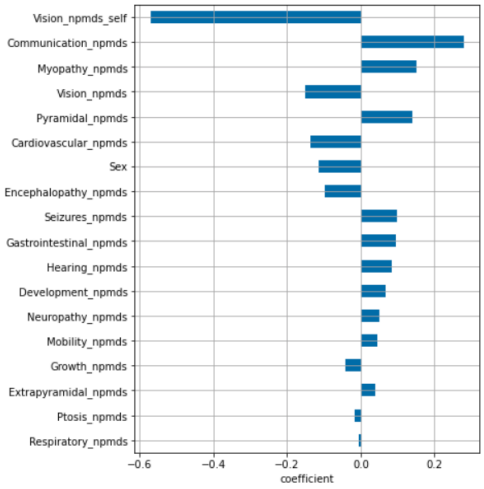
(a) Cox proportional hazard coefficients



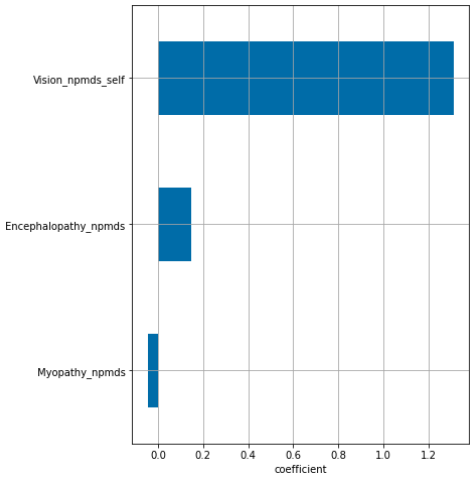
(b) Coxnet coefficients



(c) Coxnet tuning parameter

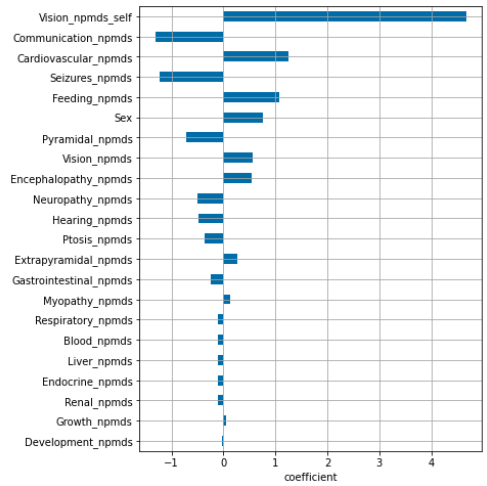


(d) Accelerated failure time coefficients

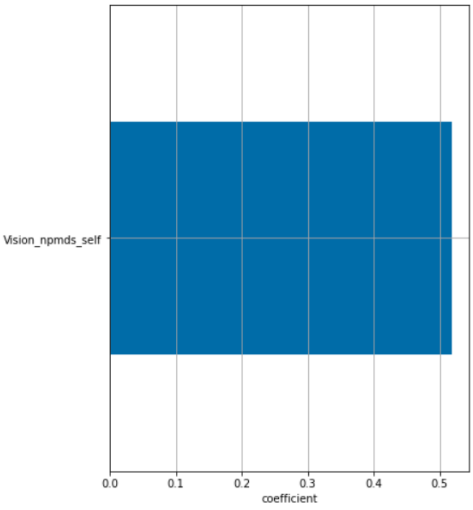


(e) Componentwise gradient boosting coefficients

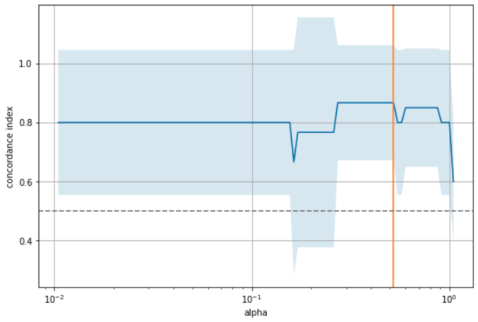
Figure 22: Model tuning and coefficients for feeding inability



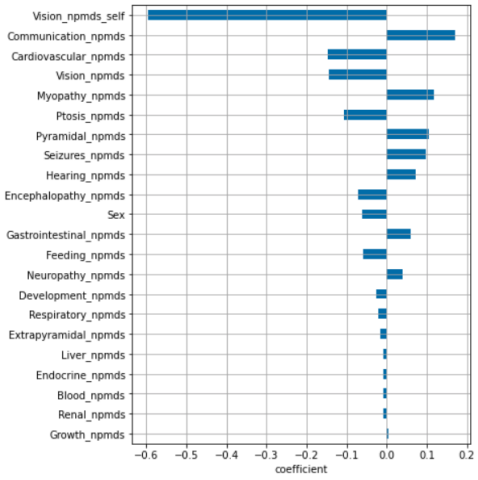
(a) Cox proportional hazard coefficients



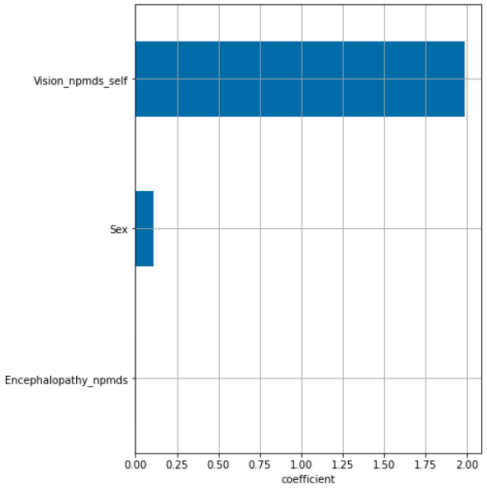
(b) Coxnet coefficients



(c) Coxnet tuning parameter



(d) Accelerated failure time coefficients



(e) Componentwise gradient boosting coefficients

Weight	Feature
0.2303 ± 0.3033	Development_npmds
0.1576 ± 0.1236	Gastrointestinal_npmds
0.0970 ± 0.1403	Seizures_npmds
0.0970 ± 0.1688	Respiratory_npmds
0.0364 ± 0.1599	Sex
0.0121 ± 0.1305	Growth_npmds

(f) Random survival forest feature importance

Figure 23: Model tuning and coefficients for Immobility

References

[1] Manuela Schubert Baldo and Laura Vilarinho. Molecular basis of Leigh syndrome: a current look. *Orphanet Journal of Rare Diseases*, 15(1):31, January 2020.

[2] Nicole J. Lake, Matthew J. Bird, Pirjo Isohanni, and Anders Paetau. Leigh Syndrome: Neuropathology and Pathogenesis. *Journal of Neuropathology & Experimental Neurology*, 74(6):482–492, June 2015.

[3] Fabian Baertling, Richard J. Rodenburg, Jörg Schaper, Jan A. Smeitink, Werner J. H. Koopman, Ertan Mayatepek, Eva Morava, and Felix Distelmaier. A guide to diagnosis and treatment of Leigh syndrome. *Journal of Neurology, Neurosurgery, and Psychiatry*, 85(3):257–265, March 2014.

[4] Nicole J. Lake, Alison G. Compton, Shamima Rahman, and David R. Thorburn. Leigh syndrome: One disorder, more than 75 monogenic causes. *Annals of Neurology*, 79(2):190–203, February 2016.

[5] S. Rahman, R. B. Blok, H.-H. M. Dahl, D. M. Danks, D. M. Kirby, C. W. Chow, J. Christodoulou, and D. R. Thorburn. Leigh syndrome: Clinical features and biochemical and DNA abnormalities. *Annals of Neurology*, 39(3):343–351, 1996. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ana.410390311.

[6] Annette Spooner, Emily Chen, Arcot Sowmya, Perminder Sachdev, Nicole A. Kochan, Julian Trollor, and Henry Brodaty. A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. *Scientific Reports*, 10(1):20410, November 2020. Number: 1 Publisher: Nature Publishing Group.

[7] Censored (Patient)Censored (patient). In Wilhelm Kirch, editor, *Encyclopedia of Public Health*, pages 104–105. Springer Netherlands, Dordrecht, 2008.

[8] T. G. Clark, M. J. Bradburn, S. B. Love, and D. G. Altman. Survival Analysis Part I: Basic concepts and first analyses. *British Journal of Cancer*, 89(2):232–238, July 2003. Number: 2 Publisher: Nature Publishing Group.

[9] Kalliopi Sofou, Irenaeus F. M. De Coo, Pirjo Isohanni, Elsebet Ostergaard, Karin Naess, Linda De Meirleir, Charalampos Tzoulis, Johanna Uusimaa, Isabell B. De Angst, Tuula Lönnqvist, Helena Pihko, Katariina Mankinen, Laurence A. Bindoff, Már Tulinius, and Niklas Darin. A multicenter study on Leigh syndrome: disease course and predictors of survival. *Orphanet Journal of Rare Diseases*, 9(1):52, April 2014.

[10] Mohammadreza Nemati, Jamal Ansary, and Nazafarin Nemati. Machine-Learning Approaches in COVID-19 Survival Analysis and Discharge-Time Likelihood Prediction Using Clinical Data. *Patterns*, 1(5):100074, August 2020.

[11] Ellery Wulczyn, David F. Steiner, Melissa Moran, Markus Plass, Robert Reihs, Fraser Tan, Isabelle Flament-Auvigne, Trissia Brown, Peter Regitnig, Po-Hsuan Cameron Chen, Narayan Hegde, Apaar Sadhwani, Robert MacDonald, Benny Ayalew, Greg S. Corrado, Lily H. Peng, Daniel Tse, Heimo Müller, Zhaoyang Xu, Yun Liu, Martin C. Stumpe, Kurt Zatloukal, and Craig H. Mermel. Interpretable survival prediction for colorectal cancer using deep learning. *npj Digital Medicine*, 4(1):1–13, April 2021. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Colon cancer;Computer science;Prognostic markers Subject_term_id: colon-cancer;computer-science;prognostic-markers.

[12] Tomonori Nakagawa, Manabu Ishida, Junpei Naito, Atsushi Nagai, Shuhei Yamaguchi, Keiichi Onoda, and on behalf of the Alzheimer's Disease Neuroimaging Initiative. Prediction of conversion to Alzheimer's disease using deep survival analysis of MRI images. *Brain Communications*, 2(1), January 2020.

[13] Changhee Lee, William Zame, Jinsung Yoon, and Mihaela van der Schaar. DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), April 2018. Number: 1.

[14] C. Phoenix, A. M. Schaefer, J. L. Elson, E. Morava, M. Bugiani, G. Uziel, J. A. Smeitink, D. M. Turnbull, and R. McFarland. A scale to monitor progression and treatment of mitochondrial disease in children. *Neuromuscular Disorders*, 16(12):814–820, December 2006.

[15] James D. Dziura, Lori A. Post, Qing Zhao, Zhixuan Fu, and Peter Peduzzi. Strategies for Dealing with Missing Data in Clinical Trials: From Design to Analysis. *The Yale Journal of Biology and Medicine*, 86(3):343–358, September 2013.

[16] World Health Organisation. WHO child growth standards: growth velocity based on weight, length and head circumference: methods and development, November 2009.

[17] D. R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972. _eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1972.tb00899.x.

[18] L. J. Wei. The accelerated failure time model: A useful alternative to the cox regression model in survival analysis. *Statistics in Medicine*, 11(14-15):1871–1879, 1992. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.4780111409.

[19] Torsten Hothorn, Peter Bühlmann, Sandrine Dudoit, Annette Molinaro, and Mark J. Van Der Laan. Survival ensembles. *Biostatistics*, 7(3):355–373, July 2006.

[20] Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3), September 2008. arXiv: 0811.1645.

[21] Frank E. Harrell, Jr, Robert M. Califf, David B. Pryor, Kerry L. Lee, and Robert A. Rosati. Evaluating the Yield of Medical Tests. *JAMA*, 247(18):2543–2546, May 1982.

[22] E. Graf, C. Schmoor, W. Sauerbrei, and M. Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17-18):2529–2545, September 1999.

[23] Inn-Chi Lee, Ayman W. El-Hattab, Jing Wang, Fang-Yuan Li, Shao-Wen Weng, William J. Craigen, and Lee-Jun C. Wong. SURF1-associated Leigh syndrome: a case series and novel mutations. *Human Mutation*, 33(8):1192–1200, August 2012.

[24] Samantha E. Marin, Ronit Mesterman, Brian Robinson, Richard J. Rodenburg, Jan Smeitink, and Mark A. Tarnopolsky. Leigh syndrome associated with mitochondrial complex I deficiency due to novel mutations In NDUFV1 and NDUFS2. *Gene*, 516(1):162–167, March 2013.

[25] Yang Ni, Muhammad A. Hagras, Vassiliki Konstantopoulou, Johannes A. Mayr, Alexei A. Stuchebrukhov, and David Meierhofer. Mutations in NDUFS1 Cause Metabolic Reprogramming and Disruption of the Electron Transfer. *Cells*, 8(10):1149, September 2019.

[26] Davide Chicco and Giuseppe Jurman. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Medical Informatics and Decision Making*, 20(1):16, February 2020.

[27] Meriem Amrane, Saliha Oukid, Ikram Gagaoua, and Tolga Ensarİ. Breast cancer classification using machine learning. In *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*, pages 1–4, April 2018.

[28] Vaishnavi Subramanian, Tanveer Syeda-Mahmood, and Minh N. Do. Multimodal fusion using sparse CCA for breast cancer survival prediction. *arXiv:2103.05432 [cs, eess, q-bio, stat]*, March 2021. arXiv: 2103.05432 version: 1.

[29] Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, and Dimitrios I. Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13:8–17, January 2015.

[30] Farah Shamout, Tingting Zhu, and David A. Clifton. Machine Learning for Clinical Outcome Prediction. *IEEE Reviews in Biomedical Engineering*, 14:116–126, 2021. Conference Name: IEEE Reviews in Biomedical Engineering.

[31] Mostafa Mehdipour Ghazi, Lauge Sørensen, Sébastien Ourselin, and Mads Nielsen. CARRNN: A Continuous Autoregressive Recurrent Neural Network for Deep Representation Learning from Sporadic Temporal Data. *arXiv:2104.03739 [cs, stat]*, April 2021. arXiv: 2104.03739.

[32] Chenxi Sun, Shenda Hong, Moxian Song, Hongyan Li, and Zhenjie Wang. Predicting COVID-19 disease progression and patient outcomes based on temporal deep learning. *BMC Medical Informatics and Decision Making*, 21(1):45, February 2021.

[33] Liberator Camilleri. History of survival analysis, March 2019.

[34] Cameron Davidson-Pilon, Jonas Kalderstam, Noah Jacobson, Sean Reed, Ben Kuhn, Paul Zivich, Mike Williamson, AbdealiJK, Deepyaman Datta, Andrew Fiore-Gartland, Alex Parij, Daniel WIlson, Gabriel, Luis Moneda, Arturo Moncada-Torres, Kyle Stark, Harsh Gadgil, Jona, JoseLlanes, Karthikeyan Singaravelan, Lilian Besson, Miguel Sancho Peña, Steven Anton, Andreas Klintberg, GrowthJeff, Javad Noorbakhsh, Matthew Begun, Ravin Kumar, Sean Hussey, and Skipper Seabold. CamDavidsonPilon/lifelines: 0.26.0, May 2021.

[35] Sebastian Pölsterl. scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn. 21(212):1–6, 2020.

[36] Laura Löschmann and Daria Smorodina. Deep Learning for Survival Analysis, February 2020.

[37] Michael L. Waskom. seaborn: statistical data visualization. *The Open Journal*, 6(60):3021, 2021.

[38] J.D Hunter. Matplotlib: A 2D graphics environment. *IEEE COMPUTER SOC*, 9(3):90–95, 2007.

[39] Mikhail Korobov and Konstantin Lopuhin. ELI5. 2017.