PREDICTING THE OUTCOME OF KIDNEY TRANSPLANTS USING MACHINE LEARNING METHODS

by

Syed Asil Ali Naqvi

Submitted in partial fulfilment of the requirements
for the degree of Master of Computer Science

at

Dalhousie University
Halifax, Nova Scotia
April 2020

Table of Contents

List of Tables

List of Figures

Abstract

The prediction of the survival of kidney grafts is based on the procedure of matching kidney donors and recipients. Machine learning can be effectively used to analyze the appropriate donor-recipient attributes from a high-dimensional transplantation dataset in developing the prediction models. In this study, we analyzed 52827 deceased donor cases from year 2000-2017 using a large dataset of kidney transplant recipients. In our approach, we divided the patients in 3 different time-cohorts— patients with graft failure in year 1, between years 2-5, and more than 5 years. The intent was to investigate the changes in the significance of patient attributes towards graft success across multiple time-periods. We applied machine learning approaches to predict the status of the graft as either failed or survived in three different time-cohorts; and to predict the risk of graft failure as either high, medium or low following a kidney transplant surgery. We experimented with 5 classification algorithms (i.e. random forest, adaptive boosting, artificial neural network, logistic regression and support vector machine). In addition to developing the prediction models, we also analyzed the changes in the significance of the features over the study. Our results indicate that support vector machine and adaptive boosting combined with SMOTE provided the best area-under-the-receiver-operating-characteristic-curve (AUROC). The cross-validated AUROC scores for predicting the graft status were 85%, 66%, and 84% in 1st and 2nd and 3rd cohort, respectively, whereas the F1-Micro score for the risk of graft failure was 62%. The feature importance scores were calculated using Gini impurity and permutation based techniques to identify the important predictors and analyze their changing contribution in predicting the results for the three different time-cohorts; we noted a change in the significance of attributes across the three different time cohorts (e.g. the number of years on dialysis before transplant was an important attribute in only 1st and 2nd time-cohorts, whereas, the recipient's age and recipient's diabetes status were important in only 3rd cohort).

List of Abbreviations Used

ESRD	End Stage Renal Disease
ML	Machine Learning
KM	Kaplan Meier
CPH	Cox Proportional Hazard
RSF	Random Survival Forest
SVR	Support Vector Regression
DT	Decision Trees
RF	Random Forests
LR	Logistic Regression
UNOS	United Network of Organ Sharing
SMOTE	Synthetic Minority Oversampling Technique
RUS	Random Under Sampling
ANN	Artificial Neural Network
CFS	Correlation Feature Selection
PC	Pearson Correlation
SA	Sensitivity Analysis
RFE	Recursive Feature Elimination
ANOVA	Analysis of variance
PCA	Principal Component Analysis
AUROC	The Area Under the Curve
ROC	Receiver operating characteristic
RFECV	Recursive feature elimination by cross validation
PV	Paired Variables
MLP	Multi-layer Perceptron
PRC	Precision Recall Curve
USRDS	United States Renal Data System

Acknowledgements

I would like to thank my God, who got me this far; who blessed me with the right people to help me during the course of my study. It gives me great pleasure to express my deepest respect and sincere thanks to my advisor Professor Syed Sibte Raza Abidi for his encouragement, valuable suggestions, discussion and guidance throughout my graduate studies. He was patient with my writing style and taught me how to explain my thoughts and be articulate while presenting them. Without his guidance and persistent help this thesis would not have been possible.

In addition to his mentorship, the success of this project required a lot of guidance and assistance from many people and I feel indebted to express my gratitude to all these people who contributed and supported me in the completion of this project.

I owe my deep gratitude to Dr. Karthik Tennankore and Dr. Amanda Vinson for their supervision from domain experts perspective. Their useful critiques, advice and assistance were invaluable in order to accomplish this project. In addition, their pleasant attitude made this collaboration joyful and productive the same time.

A special thanks to Syed Faizan, Raja Rashid and Patrice Roy for the brainstorm session at the beginning of our projects.

Last but not the least, I owe my sincere thanks to my family and friends both back home and here, for their continued love and support whose love and sacrifice for me is beyond anything I will ever understand.

1. Introduction

1.1. Motivation

Kidneys are vital for the health of an individual. The overarching purpose of the kidney is to filter the waste products from the blood and produce hormones and urine [1]. It becomes a serious health condition when the performance of kidneys starts to deteriorate because the waste substances, such as urea and creatinine, gradually begin to accumulate and become toxic for the body. When the performance of kidneys leads up to failure, it is called End Stage Renal Disease (ESRD) [2]. Today, more than 10% of the global population is affected by ESRD and its ensuing morbidities (such as cardiovascular disease, etc.). If timely measures are not taken to deal with this problem, premature death becomes an inevitable outcome [3][4].

Kidney transplantation or dialysis are two main treatments of kidney failure [5]. Kidney transplantation is the optimal form of kidney replacement therapy treatments with respect to improving the patient's length and quality of life [6][7]. However, whilst kidney transplantation has shown great survival results in the past two decades, the graft rejection rate is still considerably high and even when the kidney starts to function properly after transplant, there is a high probability of long term graft failure. In the US, there are an increased number of dialysis patients who have been reported to have a transplant failure after a few years of their surgery [8]. Kidney allocation is based on a number of donor-recipient related factors and it is very important to study individual factors that are responsible for graft failure in the short and long term. Studies have been conducted on the implications of these factors in the clinical domain, but the complex interaction amongst these factors still requires a thorough understanding. For this, modern predictive techniques appear an effective solution.

Prediction tools and models have gained attention in the recent years [9]. Pre-transplant donor-recipient factors have been incorporated in these prediction tools and models to find hidden interactions and predict the outcome of the transplant. Machine Learning (ML) has played an effective role in the development of these prediction mechanisms and there has been growing evidence that the application of supervised ML techniques can further aid in improving predictive accuracy.

1.2. Research Objectives

This study aims to leverage ML methods to improve the accuracy of predicting the outcome of kidney transplants and to empower nephrologists in selecting the best deceased donor kidney during the organ matching process. It intends to address the prediction of graft survival and graft status over short, medium and long term because they are usually considered by the nephrologists before allocating the organ. Our focus in this research is on the classification techniques which have the ability to make class predictions in different time-cohorts.

Specifically, this research seeks to:

a) identify appropriate supervised classification methods and develop prediction models to predict graft survival and graft status in short, medium and long term.
b) analyze the changing effect of significant donor-recipient related predictors over the period of the study.

The main challenge that we face during this research is the selection of appropriate features from a high dimensional feature space to build a robust machine learning model.

1.3. Solution Approach

We aim to attain our research objectives by exercising binary and multiclass classification approaches. Our solution approach comprises of standard data mining methodology which is based on data preparation, feature selection, prediction modelling and evaluation. The task of data preparation involves the discovery or identification of outcome variables that hold vital implications for the graft survival. Specifically, this thesis explores and investigates the potential of breaking the dataset into binary and multiclass classification problems that can effectively perform data analytics.

1.4. Contribution

This research can potentially yield benefits for nephrologists, researchers and recipients. The contributions of the thesis are summarized as follows:

- This research provides a detailed study on the implications of preparing the dataset as a binary or a multiclass problem.

- It provides an analysis of results to discover the best model among all and evaluation of results with the baseline results
- It identifies and provides an analysis of important features during different time-cohorts

Most existing studies are based on predicting the status of the graft (failed or survived) in overlapped time-cohorts, thus, those features which are important during a particular frame of time cannot be identified. Our research has addressed a gap in the research literature by involving both overlapped and non-overlapped time cohorts along with an exploratory study on multiclass classification approach to develop the prediction models and analyze the changing relevance of features over time.

1.5. Organization of the thesis

This thesis is composed of five chapters. Whilst the first chapter introduces the thesis, the second chapter presents the basic concepts of survival analysis and reviews the literature pertaining to ML based prediction modelling and feature selection techniques used in this thesis, as well as the conventional statistical techniques for the analysis of high dimensional survival data.

The third chapter describes the data mining methodology for analyzing high-dimensional survival data from the classification perspective. Three different approaches of breaking the dataset to select significant features and develop prediction models are discussed.

The fourth chapter provides the results and analysis: a) to discuss the classification approaches considered in the research; b) to discover the best model among all; and to evaluate the changing relevance of the feature over time.

The thesis concludes in Chapter 5 with a summary of the study's contributions, limitations and suggestions for future research.

2. Background and Related Works

In this chapter we provide the background required to understand the rest of this thesis. We start by discussing the survival analysis. This includes a brief overview of statistical techniques such as the Kaplan-Meier estimator, and the proportional hazards model and the problem of censored data. Next, we discuss the machine learning approaches to survival analysis which encapsulates the resampling techniques, prediction modelling, selection of significant features, and evaluation metrics used in the literature.

2.1. Survival Analysis

Survival analysis is a widely used method in the domain of healthcare to analyze the time-to-event patient data. The event of interest, particularly in healthcare is an adverse outcome, which can be a death of a patient, re-occurrence of a disease or a relapse after a surgery, is measured within a specific time frame called survival time [10]. The survival time is calculated as the span from the beginning of the follow-up to the occurrence of an adverse event. The likeliness of the adverse event tends to increase with the passage of time, however, it is also highly likely that the patient stops to follow-up before the occurrence of the event. If that is the case, the event of interest is said to be missing in the observation and the time of the outcome is said to be right censored [11].

An observation is considered as censored data when the information of a participant is missing before the end of the study due to any reason that is unrelated to the study. In contrast to other statistical techniques, survival analysis approaches have the capability to deal with the censored data.

There are broadly two kinds of censored data in survival analysis: interval and point. Since most of the survival data is point censored, we will only refer to the point censoring in the discussion to follow. Point censoring, which is also known as right censoring, occurs when the participant in study stops to follow-up during the course of the study or the participant completes the study without the occurrence of the adverse event. Various techniques have been developed by statisticians to deal with the problem of right-censored data. These includes complete data analysis, imputation of the data, removal of all censored observations and treating the censored data as non-event [12][13]. In machine learning approaches, the patients who are censored in the later part of the study are considered to be

non-failure/successful, however, those who are censored in the early stages are either right-censored and discarded from the analysis or are generally considered twice in the dataset, one as experiencing the event and one as event-free [11][14].

Independent censoring, random censoring and non-informative censoring are three assumptions that are taken into consideration whilst analyzing survival data. These assumptions establish that the participant who dropped out of the study does not do so due to the reasons related to the study [15].

2.1.1. Conventional Techniques in Survival Analysis

The two primary functions in survival analysis that are used to calculate the probabilities related to event of interest are survival function S(t) and hazard function h(t). The survival function is important to survival analysis because it provides survival probabilities for any point of time beginning from the induction of the patient to the occurrence of adverse event. Here T
is the positive random variable indicating the time from the beginning of the event till the survival.

$$S(t) = P(T > t), t \geq 0.$$

On the other hand, the hazard function H(t) is a conditional failure rate of an individual which is conditional that the individual survived up until time *t*. The value of hazard function can range from 0 to infinity, and can be either increasing, decreasing or constant.

$$h(t) = lim_{dt \to 0} \frac{P(t \leq T < t + dt | T \geq t)}{dt}, t \geq 0.$$

Both survival function and hazard function are closely related to each other. The general formula is express as follows:

$$h(t) = \frac{f(t)}{S(t)} = \frac{-dlogS(t)}{dt}$$

and

$$S(t) = e^{-H(t)}$$

We can define the cumulative hazard function as a survival function in the following way:

$$S(t) = e^{-H(t)} = e^{-\int_0^t h(u)du}, t \geq 0$$

Kaplan Meier (KM) model and Cox Proportional Hazard (CPH) model are two of the most used statistical techniques in survival analysis. The KM model is a simplest nonparametric test used to plot the survival curve and estimate the probability of both censored and uncensored individual surviving at a given time period. The method is also known as "Product Limit Estimate" because it involves computation of probabilities for the occurrence of adverse event in all points of time. The final estimate is provided by multiplying the successive probabilities by any earlier computed probability of survival [16]. KM estimates are used in conjunction with Log Rank test to compare the statistical difference between the groups in the study [17]. Since the KM model is based on one single covariate, it is only suitable to use if the groups being compared are reasonably similar.

When multiple covariates are needed to be factored in the analysis, CPH model is utilized to do the needful.

CPH model is a semi-parametric regression model in survival analysis which is used to fit the survival data with multiple covariates on baseline hazard function $h(t)_0$. The model is semi-parametric because the baseline is calculated using KM estimate which is a non-parametric function. The model is fitted in the form below:

$$h(t) = h(t)_0 \exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)$$

where *t* is determined by the hazard function $h(t)$ and covariates are determined by *x*. The baseline hazard function $h(t)_0$ assumes that the covariates are all zero at the start of study. The value of baseline hazard varies during the course of the study but the coefficients of covariates remain constant throughout the period. Due to this constant effect of covariates during the study, this is also named as proportional hazards model. The quantity of $exp(\beta)$ is the ratio of hazards which can multiplicatively impact the risk of the event by increasing or decreasing it depending on the change in the number of covariates, irrespective of time.

It can also be written as $h(t)/h(t)_0$. The values of β are adjusted to maximize the Cox partial likelihood. The Cox partial likelihood is simply the product of probabilities for the events of an individual at all points of time which are conditional to all the individuals who are at risk at that particular point of time.

Significant research in medical domain has been done with variations of cox based models in the survival analysis of different organ transplants [18][19][20]. Despite the limitations such as the assumption of the proportionality of the hazards etc., the model is widely used for its flexibility with the point censored data.

2.2. Machine Learning in Survival Analysis

Machine Learning (ML) approaches in survival analysis has provided us more robust alternatives. The capability to deal with high dimensionality of risk factors by finding the complex interactions between them is arguably the main difference between the ML methods and the traditional statistical methods. The ML methods are generally classified into regression and classification methods. The main difference between a regression and classification method is the type of the target variable. In regression methods, the target variable is continuous whereas in classification methods, the target variable is categorical. We will review the regression and classification methods in the sections below.

2.2.1. ML Regression Methods in Survival Analysis

The popular ML regression methods which are developed particularly for survival analysis includes survival ensembles, support vector methods and Multi-task Logistic Regression [21][22][23]. Survival ensembles consist of different types; however, the most prevalent method is the Random Survival Forest (RSF)[24][25].

RSF has been used alongside Cox models in medical literature [22][24]. The non-parametric nature of the algorithm helps to deal with the time-varying effect of the variables. Apart from this, RSF has the ability to determine the important features from a high-dimensional feature space in the presence of right-censored data. The study by Pölsterl, Sebastian et. al [24] developed a neighborhood graph which is used to create a low-dimensional representation of the data which is close to the neighborhood of the high-dimensional data. They used RSF to determine the local neighborhood relations in the

presence of right-censored data and perform feature selection from a high dimensional feature set.

Support vector methods, on the other hand, have been categorized into ranking, regression, and combined approaches. Research shows that ranking approaches outperforms the other two approaches [21]. Traditional Support Vector Regression (SVR) models have been widely adopted in the literature but there are few instances where they are applied in survival analysis. The presence of right-censored data is the main issue when applying a traditional SVR because it requires the data to have a response variable [26].

In addition to survival ensembles and support vector methods, survival neural networks have also been proposed in the past decade but they are, in essence, the generalized form of Cox models for nonlinear functions [27]. However deep learning models which are based on dense neural networks have been adequately used to perform survival analysis in organ transplants.

A number of studies have been conducted with deep neural networks to predict the time of survival. In the study by K. Matsuo *et al.* [28], survival risk of patients with cervical cancer was computed using a deep learning model. The model used a subnetwork of deep neural networks with a single output node and was trained on 3 disjoint subsets of feature space. The results showed that the model outperformed the different variants of Cox proportional hazard such as CoxLasso and CoxBoost for each subset of feature space. Likewise, in the study by M. Luck et al. [29], probability of survival was also predicted using deep learning in kidney graft survival analysis and the model again outperformed the conventional Cox proportional hazard model.

2.2.2. ML Classification Methods in Survival Analysis

Classification methods have been extensively adopted in studies on organ transplants. The time of survival in classification methods are converted into a binary or multiclass variable as per the desired outcome. The desired outcome is generally the status of the graft or the risk of graft failure in an arbitrary time period. Discussion on this desired target outcome is done in the subsection Binary and Multiclass Approaches, below. Apart from the prediction of the class variable, classification methods have proven their utility in selecting the significant features from a high-dimensional feature space using multiple wrapper

methods (See section Feature Selection). However, unlike Cox models and other survival methods, the classification methods do not provide any proper mechanism to deal with the right-censored data; therefore, multiple different ways have been adopted in studies to deal with this problem. The study by Kazim et al. [14] on kidney transplants discarded all the right-censored data before seven years from the time of transplant and included the rest in the low risk group. In another study on the prediction of heart transplant outcomes by A. Dag et al. [30], the dataset was broken down into three different time cohorts (one year, five years, and nine years) in order to predict the status of graft. All those patients who did not have any graft failure during that particular time-cohort were dropped from the analysis and all the patients beyond that time cohort were considered as successful transplants.

In our study, the research questions are answered using classification approaches instead of regression approaches because we wanted to predict the risk of graft failure in short, medium and long term. The regression approaches have two main problems: a) the methods to make predictions and identify important features in survival data were relatively few, and b) the results were very difficult to evaluate, since the metrics (e.g. Mean Squared Error) have little interpretability. The classification methods that we have used in this research are stated in the subsection Supervised Machine Learning, however, the most frequently used classification techniques in survival analysis are reviewed below along with their advantages and limitations.

The two classification algorithms which have been the first choice in several studies with survival data are Decision Trees (DT) and Random Forests (RF). Both DTs and RFs have high interpretability and fast computational power, but they have failed to compete with their more robust counterparts such as SVMs, ANNs and even Logistic Regression (LR) in producing better results [31][32][33][34][35]. A combination of ANN, SVM, LR and Bayesian belief networks were also used and compared in different datasets on heart, liver and kidney transplants [30][14][36][37]. The application ANNs on survival data has shown significant improvement in reducing the error rates in comparison to the conventional ML models. However, the two main constraints that it generally pose are the requirement of high computational power and bigger datasets [38][39]. The study by D. Medved et al. [40], trained a straightforward neural network with 2 layers each with 128 nodes using

Keras framework on a United Network of Organ Sharing (UNOS) heart transplant dataset. The results showed a marked difference of F1 macro score that rose to 0.68 from 0.271 when compared with the baseline model.

The study by C. Lee et al. [41] designed DeepHit which is a novel approach to predict the survival of a patient who possesses a risk of comorbidities. DeepHit has the ability to handle situations where there is a single underlying risk (occurrence of adverse event due to single cause) as well as multiple competing risks (occurrence of adverse event due to multiple causes). It utilizes multiple layers of shared and adverse-event specific sub networks to train the neural network. Although the aim was to develop a network for multiple adverse events, the network also worked well for single adverse event. The study used concordance index metric to measure the survival probabilities in any given time point, however they did not train the networks for different time-cohorts as we have generally seen in studies taking a classification approach to survival problems.

2.2.3. Class Imbalance

The biggest issue that comes with survival data - when it is approached as a classification problem - is the handling of class imbalance. This problem occurs when the distribution of outcome variable in the dataset is uneven. In survival data of organ transplants, irrespective of whether the outcome of the prediction is the risk of graft survival or the status of the graft, the number of failed grafts monotonically increase with time. The workaround to deal with this imbalance is to use over and under sampling techniques.

Synthetic Minority Oversampling Technique (SMOTE) and Random Under Sampling (RUS) have been the popular techniques to deal with the class imbalance. SMOTE oversamples the minority class by generating data based on the neighboring observations. The k-NN algorithm is used to determine the number of nearest neighbors in the oversampling process. The technique allows the users to specify the type of variables (e.g. categorical and continuous) to perform accurate oversampling. The study by Blagus et. al. [42] investigated the behavior of SMOTE on high-dimensional imbalanced data and found that it does not change the class-specific mean values whereas it decreases the data variability and introduces correlation between samples. The study on heart transplantation [30] used SMOTE to oversample the minority class in the three different time cohorts and the results were significantly improved when compared to the original dataset. Though it

improved the scores, it is a computationally expensive technique in comparison to its other counterparts [43]. RUS is relatively an easier technique. The way it works is that it selects the samples from majority class to even out the imbalance. Although, it is relatively faster than SMOTE, it fails to include important information [44].

2.2.4. Binary and Multiclass Approaches

There are broadly two approaches in medical literature to predict the outcome of the graft: a) binary and b) multiclass. The first approach - which is a binary approach - is widely adopted to predict the status of the graft in different time points - we call them time-cohorts. In this thesis, we have categorized the binary approach into overlapping and non-overlapping cohorts. These terms overlapping and non-overlapping have been frequently used throughout this thesis.

The overlapping cohorts are defined as those where the starting point for the all the time-cohorts is the beginning of the study, whereas the non-overlapping cohorts are defined as those where each cohort is mutually exclusive. The binary approach with overlapped cohorts has an extensive body of literature. As an example of the approach, deep learning was used to predict the heart transplantation outcome in three different time cohorts [40]. The time cohorts were formed for 0-180 days, 0- 365 days, 0-730 days, where all the instances in the prior cohorts were included in the later cohorts. Each of these cohorts were based on the number of individuals who either had experienced the adverse event or had remained safe during that time. The target variable was then used to train the classification models to predict the status of the graft by the end of that time-cohort.

The second approach is a multiclass approach which involves the classification of the dataset into different risk groups. These risk groups are considered as the target variable used in the prediction process. In this study, three risk groups (high, medium, and low) were determined to predict the graft survival on kidney transplants dataset [14]. The first risk group comprises of patients who had a failure in the first year, the second group consists of patients who had failure between the second and seventh year, and the third group has all the remaining patients. An example of both overlapping binary and multiclass approach is shown in this study by J. Li et al. [45], where Bayes net classifiers were used to classify the status of the graft as well as the survival of the graft on renal transplantation

dataset; however, the dataset that they used in the analysis was very small in comparison to other recent studies.

To the best of our knowledge, non-overlapping cohorts have not been studied in the literature. This would be the first study where non-overlapping cohort are considered for developing prediction models and analyzing the changing relevance of features over time. The motivation to formulate non-overlapping cohorts came from the fact that feature importance scores calculated in studies with overlapping cohorts only tell about the features which either emanate their influence in short term or long term. They cannot find out the features which are specifically important during a specific time frame. An arbitrary cohort based on long term transplants (e.g. 0-10 years) would intrinsically include the short term (e.g. 0-1 years) and medium term (e.g. 0-5 years) cases within itself hence, it would not be possible to find out the important features with an arbitrary time periods of 2-5 years, without involving the bias of the data from 0-2 years. Multiclass approaches also lack in this analysis because the significant features which are selected after the feature selection represents the whole dataset rather than a particular class of the outcome variable. A recent study on heart transplantation by A.Dag et al. [30] analyzed the changing significance of features for three overlapping time-cohorts (1-, 5-, and 9-year). They deduced that certain type of features perform well in long term as compared to the short and medium term. For instance, the socio-economic factors were more influential in 9-year time-cohort as it was covering major variation of the data. As mentioned above, this interpretation has one key issue: it cannot substantiate whether the socio-economic factors were actually influential in long-term or the merger of cohorts have made them more responsive to the target variable. The influence of factors in different intervals can only be substantiated if the analysis is done with non-overlapping cohorts with data relevant to that particular time period only with no previous data having the potential to create a bias.

2.2.5. Feature Selection

As mentioned in the sections above, feature selection is an important aspect of ML approaches. The survival data constitutes a plethora of features which can be primarily dichotomized into clinical and socio-economic features. With the advancement in technology and better recording tools, the feature space is constantly growing with

additional features, but it does not mean that all the features are equally important hence it is essential to remove the redundant features to prevent the perplexity during the model training. A study by A. Agrawal et. al [46], using UNOS dataset of lung transplants reduced the feature set from 50 features to 8 features by applying a feature selection technique. Selection of important features not only helped in reducing the feature space, it also improved the accuracy of the trained models by reducing the errors of making the wrong predictions. The Artificial Neural Network (ANN) model provided the AUROC score of 59% when it was trained with 62 features, however the score rose to 66% when the feature set was reduced to only 12 features. There were a number of models that were trained but ANN showed the biggest difference before and after selecting the important features.

Classification methods provide a high utility of analyzing the importance of the features. Backward/forward feature selection and sensitivity analysis are two of the frequently used feature selection techniques to understand the significance of features in the dataset.

A study on lung transplants stated above used feature importance scores to select the important features after applying Correlation Feature Selection (CFS) on the complete feature space. The technique CFS is an extension of simple Pearson Correlation (PC). It adapts a greedy approach to find a subset of features which have a high correlation with the class variable and are weakly correlated amongst each other [46].

Features tend to perform differently over time. Effect of predictors were found to be different in the long and short term [18]. The Information Fusion (IF) technique applied in heart transplant study above helped to analyze the performance of variables in three-time cohorts. Seven groups were formulated based on the significance of variables [30]. Another study by J. Yoon et. al [13], provided a complete transition of feature significance using a heat map.

The different feature selection techniques which are related to the classification methods are discussed below.

2.2.5.1. Sensitivity Analysis

Sensitivity Analysis (SA) is a powerful technique which is used to enhance the accuracy of the model by selecting those features which have the maximum influence on the output's

variability [47]. In other words, the sensitivity of the variable is calculated by taking the ratio of the prediction error from the time when the variable is included in the error to the time when the variable is not included. The equation below defines the sensitivity measure (S_i) used in the feature selection process:

$$S_i = \frac{V_i}{V(y)} = \frac{V(E(y|x_i)}{V(y)}$$

where y is the output variable (graft status or the risk of graft failure), $V(y)$ is the unconditional output variance, and E is the expectation operator, which calls for an integral over all predictor variables except x_i.A further integral operator is implied over x_i by the operator V_i. The importance of a specific variable is then computed as the normalized sensitivity, as described by Saltelli et al. [47].

The term SA has multiple connotations attached to it in different disciplines [48][49][50], but in our research, we will follow the definition that has been stated above followed by the equation. It can be further simplified as understanding the activeness of input factors.

Studies on kidney and heart transplants mentioned above analyzed the sensitivities of the predictors using a number of different ML algorithms and finally adopted the IF technique [14][30][51]. The IF technique creates a fused model of the sensitivities of each predictor by taking a weighted sum of each predictor by the evaluation measure of the model. The equation below explains how the fused model is computed.

$$S_{\theta(fused)} = \sum_{i=1}^{r} \lambda_i S_{i,\theta}$$

The λ_i is the evaluation measure of the trained model which is multiplied by the sensitivity of the predictor. A weighted average is calculated by training models to finally generate a set of $S_{\theta(fused)}$ fused sensitivities. The evaluation measure employed in kidney dataset is an f-measure whilst the authors in heart transplantation chose AUROC score as weights of the predictors. The classifiers that were used to create a fused models were decision trees,

artificial neural network, support vector machines bootstrap forest and logistic regression [14][30].

2.2.5.2. Backward and Forward Feature Selection

Backward feature elimination is a feature selection technique which keeps removing the features with the least information until the model stops to make further improvement. It begins with all the features in the feature space and removes each feature one by one to check the best score when the feature is not included during the training phase of the model [40]. The features removed in backward elimination are usually based on the p-values. The study by Lee et al. [52] performed simple statistical analysis with backward stepwise variable selection based on p-values with cutoff of less than 0.1 to select the important features. Recursive Feature Elimination (RFE) is a ML variant of backward feature elimination. It involves the feature ranking system to remove the weak features. The feature ranking system used in recursive feature elimination can be based on multiple feature importance metrics such as sensitivity analysis, information gain and linear discriminant analysis [53][54][55]. The study by Escanilla et al. [53] performed RFE by sensitivity testing for SVM's with nonlinear kernel. The SVM's with linear kernel provides the coefficients for all the predictors which are then used in the RFE. All the algorithms which provide *feature importance* can be used in RFE, however, all the other algorithms can be utilized in RFE process with the help of sensitivity analysis. Another study used RFE by 10-fold cross validation using the coefficients of a Logistic Regression model to produce the results [56].

Recursive feature elimination becomes a greedy sub-optimal method when multiple features are set to remove on each step. The study by Guyon et al. [57] explained how RFE can miss an important singleton feature in an effort to get the best subset of the features whilst eliminating multiple features at a time.

Forward feature selection is very similar to backward feature elimination except that the features are added into an empty feature set with the priority given to most significant feature. The process stops when the performance of the model stops to improve. Both forward feature selection and backward elimination have been used with a greedy approach in a study by Medved, et al. [58]. They selected an optimal feature set by performing a

forward and backward search on a set of 482 variables from UNOS dataset for heart transplants. A threshold value of 0.0001 was used to evaluate the difference in the score by adding or removing the variable in forward selection and backward elimination, respectively. The point of interest in the study is the way how forward and backward search is employed to find the local optimum. The process starts off with an empty set and new variables are added into it using forward selection until the score stops to improve. Once the score stops to improve backward elimination comes into action on the set of variables formed by forward selection. The process continues back and forth until the score of both the backward and forward searches stops to improve by the threshold amount.

2.2.5.3. Feature Importance Scores

In DTs, the split of a node is based on the condition of the impurity. In classification problems, the criterion is either Gini impurity or information gain. When a dataset is trained with a tree classifier, the contribution of each feature in reducing the weighted impurity can be easily analyzed. In RF, the only difference is that the reduction in weighted impurity is based on the average of all the trees generated in the training process rather than a single tree. Computing feature scores is a fast calculation in terms of time and size of the memory, however, it has one major problem of inflating the importance of continuous features in comparison to categorical features. The continuous features always have the largest mean decrease in comparison to categorical features and even in categorical features, the features with high-cardinality are preferred over others.

The workaround to deal with this intrinsic bias of impurity based feature scores is to use permutation based feature scores. The permutation based feature importance is measured by re-shuffling one of the predictors in the feature set whilst training a model. The drop in the outcome explains the dependency of the feature in the model. It is defined to be the decrease in a model score when a single feature value is randomly shuffled [59]. Although, this technique is reasonable effective, it is a computationally expensive in comparison to impurity based feature importance. Also, it overestimates the importance of the features which are highly correlated therefore it is important to handle the issue of collinearity before using this technique.

2.2.5.4. Miscellaneous Techniques

There are a numerous feature selection and feature reduction techniques which have been used in medical literature. The most common statistical technique which is used in the studies is PC. The study by A. J. Aljaaf *et al.* [60] used PC along with analysis of variance (ANOVA) to find the correlation among the different variables in the dataset. The study found a strong positive correlation between urea and creatinine level but chose to remove urea from the feature set because creatinine was more likely to correlate with the class variable then urea [4].

Among feature reduction techniques PCA and auto-encoders have been used in few studies. The study by M.Zafar et al. [61] used auto-encoders to perform representation learning on the continuous features and input the resultant feature set along with the remaining categorical features for the conventional feature selection. The significant features from the final feature set were used for model training. The hyper parameters of the stacked auto-encoders were supposed to be adjusted for best accuracies. Since the stacked auto-encoders were the first stage of the process, tweaking the parameters meant reiterating the whole process all over again. They compared the results with random forest before and after using auto-encoders and concluded that auto-encoders outperformed, significantly.

Principal Component Analysis (PCA) is one of the effective feature reduction techniques. It works by finding the dimensions which have the highest variance and then retaining their variance in a new variable. Interestingly, PCA was among the least used techniques in the domain of survival analysis of organ transplantation. A study by J. Lasserre et al. [62] on predicting the renal transplantation outcome used PCA and Relief-F on reducing the feature space of 36 variables and concluded that PCA and Relief-F performed worse on their dataset. In another study by Raji et al. [36], PCA was used on a set of 197 attributes of liver transplant data. 27 attributes out of the whole feature space were ranked based on the standard deviation. Association rule mining was done before and after performing PCA to ensure the improvement in new rules. The important variables were eventually used in ANN for training and testing.

To summarize the findings on PCA, it was interesting to see that it was rarely used technique to reduce the dimensionality of the data. The studies which used PCA in feature reduction did not highlight any significant performance improvement specifically due to

the use of PCA. In fact, one study claimed that PCA had worked contrary to any improvement and had actually worsen the performance.

2.2.6. Supervised Machine Learning

The review of the classification methods is already provided in the section ML Classification Methods in Survival Analysis above. Here we have provided the background for five ML classification methods that we have used in this research.

2.2.6.1. Random Forest

Random Forest (RF) [59] is an ensemble machine learning method based on multiple decision trees. The final decision in RF is based on the majority class outputted by the individual decision trees in the forest. It is a fast and easy to use algorithm mainly used in classification problems, however it is now widely used in survival analysis. Random Survival Forest which is a variant of Random Forest has shown its significance with right censored data in medical literature. The technique that underpins RF is bootstrap aggregation (bagging). Bagging makes RF a robust algorithm resistant to overfitting.

RF has the ability to handle high dimensional dataset with imbalanced class distribution. The *feature importance* calculation of RF enables the algorithm to perform feature selection. It measures the impact of features by checking the mean decrease accuracy and mean decrease impurity by removing a feature from the feature set. Unimportant features do not account for any significant change in the accuracy. This in-built selection mechanism allows RF to discard unnecessary features from the dataset by recursive feature elimination. In this study, we have utilized this functionality of RF to select the best features from our dataset. We performed a 10-fold cross validated recursive feature elimination using RF. The resultant feature set is then used to fit multiple classifiers.

2.2.6.2. Artificial Neural Network

ANNs [63] are extensively used in data mining problem based on organ transplant datasets. The algorithm is primarily used in classification problems however different variants of ANNs are also used alongside Cox models in survival analysis. An ANN is a computational system that consists of "a highly interconnected set of processing elements, called neurons, which process information as a response to external stimuli. An artificial neuron is a simplistic representation that emulates the signal integration and threshold firing behavior

of biological neurons by means of mathematical equations" [41]. There are several algorithms for training neural networks and the most popular one among them are multilayer perceptron. This algorithm is based on input, output, and hidden layers, where each layer includes several nodes, referred to as neurons. In the studies on organ transplants, the input layer consists of all the pre-transplant donor-recipient variables affecting the status of the graft. The response variable of graft status is represented in the output layer with two neurons depicting the possible outcomes. In this research, we experimented with different kind of ANNs. We started performing some initial experiments using Keras framework however a multilayer perceptron-based ANN was finally selected in this study as it outperformed all other ANNs.

2.2.6.3. Support Vector Machine

SVMs [64] are robust classification algorithms capable of handling linearly and non-linearly separable datasets. These algorithms are now also used in survival analysis to deal with censored data. For survival analysis, the two approaches, Ranking and Regression, have been used on high-dimensional clinical datasets to analyze their relevance against classical survival analysis techniques. SVMs generally perform by creating hyperplanes in multidimensional space to classify the class labels. A nonlinear dataset is transformed into multi-dimensional space so that the class labels become linearly separable. SVMs consist of several kernel functions. The purpose of these functions is to deal with the problem of computational inefficiency that grows immensely with the high dimensional datasets. In our study, we used SVMs with linear, polynomial, and radial basis function kernels. Our preliminary results showed that radial kernel performed best on our dataset, hence we used that kernel in further experiments.

2.2.6.4. Adaptive Boosting

AdaBoost [65] is a boosting method based on meta-learning that trains multiple weak classifiers to make one strong classifier. A pipeline of models is developed where the results of one trained model are passed onto the another model in the pipeline. The emphasis is given to the errors in the first model to be rectified in the next model. The process continues until perfect prediction is made or the desired number of models are trained. Different base classifiers can be used as weak learners in AdaBoost. A study on

breast cancer survivability prediction used the algorithm in combination with RF which resulted in improved results than standalone classifier [66]. We also emulated the strategy in this study with RF and Logistic Regression (LR). RF as base learner outperformed LR in most of the experiments.

2.2.6.5. Logistic Regression

Logistic Regression (LR) is a linear model based on the assumption that predictors do not have multi-collinearity with each other. As evident from the literature in the domain of medicine, LR has been repeatedly used in classification problems. We included this classifier due to its prominence so that a valid comparison can be made with other similar studies and baseline could be set for the other four classifiers that we have planned to use in this research. Apart from being computationally efficient, the implementation of the algorithm also provided us the probability scores within the least amount of time in comparison to other classifiers in the study.

2.2.7. Performance Metrics

2.2.7.1. Area under ROC (AUROC)

The Area Under the Curve (AUROC) is an accepted traditional performance metric for a ROC curve [67][68]. A receiver operating characteristic (ROC) graph is a technique for visualizing, organizing, and selecting classifiers based on their performance. ROC graphs are two-dimensional graphs in which the true positive rate (TPR) is plotted on the y-axis and the false positive rate (FPR) is plotted on the x-axis. AUROC describes the performance of a classifier using a single scalar value. Because both TPR and FPR are bounded in the interval [0.0,1.0], the area is also bounded between [0.0,1.0][69]. It performs a cost benefit analysis of a classifier in a graphical style by calculating the trade-off between TPR and FPR. A classifier that outputs a random label should have an AUROC value of 0.5, and therefore no functional classifier should have a lower value than that.

2.2.7.2. F1-Measure

The F1 score is the harmonic mean of precision and recall, see equation below and is bounded in the interval [0.0,1.0][70]. The score tends to be close to the minimum of both the precision and recall.

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

We also used the F1-Micro score which is the harmonic mean between micro-precision and micro-recall [71]. The F1-Micro score was used to evaluate the models. It was considered in those situations where either the class imbalance was present or the AUROC was providing the same results for all the models.

2.3. Summary

A comprehensive literature review and background of survival analysis in healthcare is provided in this chapter. We began by examining the traditional techniques of survival analysis and how machine learning has been used in the process. We then looked into a number of feature selection/reduction methodologies which are deemed as one of the preliminary steps in the process of making robust classification models for prediction. Studies showed a spectrum of techniques ranging from simple statistical analysis (such as PC) to advanced stacked auto-encoders to be used in filtering down the feature space. Although PC is rather simple and fast method of computing correlation among variables, it has a limitation to be used only on categorical variables. Same is the case with stacked auto-encoders. They generally perform better than PC, but they require large dataset and a lot of memory to provide good results.

Backward elimination and sensitivity analysis were among the most frequently used feature selection methods which were implemented either one by one on each feature or a subset of features to rank their significance. Performing them on each feature one by one is not an optimal solution therefore subsets of different features were chosen to process them in order to attain the best results. Even using the subset does not guarantee a global optimum solution therefore studies have taken a greedy approach to find a local optimum using either of the two implementations.

Among all the feature selection techniques discussed above, variable pairing technique was nowhere to be seen. A few donor and recipient variables such as donor-recipient age, sex, and race etc. have the potential to be paired together in order to reduce the redundancy however a gap seems to exist as experiments are needed to be done to explore and analyze its influence.

Censored observations in survival data is a very common phenomenon. Studies have proposed various ways to handle the survival data. The easiest way to handle is to discard it, however doing that threatens a significant portion of important information to be removed from the dataset.

Once the feature selection and censored data problems have been smoothly resolved, the next big step has been the model training and testing. The general approach has been to train multiple models simultaneously and evaluate the performance mainly using Area Under the curve of the Receiver Operating Characteristic (AUROC) or C-Index.

3. Methodology

In this research, we followed a standard data mining methodology comprising data preparation, feature selection, prediction modelling and model evaluation to fulfil our research objectives. The workflow graph in Figure 1 demonstrates the pipeline of the steps taken to solve the problem.

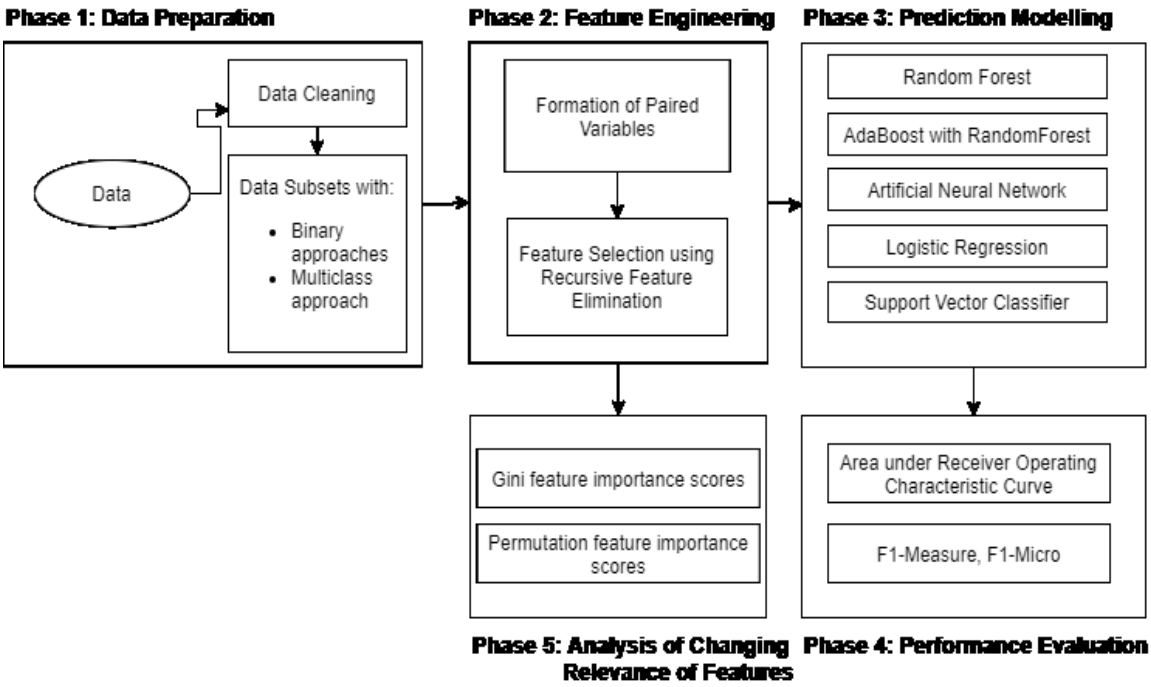


Figure 1 An overview of the methodology

The data preparation phase consists of cleaning the dataset by removing the data observations and defining the target variable for performing classification analysis. The literature has shown us two approaches to perform classification on survival data (see Binary and Multiclass Approaches). Among binary approaches, the widely adopted strategy to perform the prediction analysis is by overlapping cohort, but since overlapping cohorts cannot substantiate the importance of features in short, medium and long term, we formulated the non-overlapping cohorts for the analysis. Thus, during the data preparation phase, we performed the analysis with overlapping and non-overlapping cohorts from the perspective of binary approach along with a subsidiary multiclass approach. The overlapping and non-overlapping cohorts were divided in to three different time-cohorts, whereas, the multiclass problem had one consolidated dataset with three different classes representing the risk of graft failure in different time points. Class imbalance remained an

obvious problem during the data preparation phase irrespective of the approaches; hence, we used an oversampling technique called SMOTE to adjust the imbalance.

In the next phase, we performed feature selection using recursive feature elimination by cross validation (RFECV) and simultaneously calculated feature importance scores to analyze the changing significance of the features over time. We developed four different prediction models for each time cohort with the filtered set of features retrieved after performing feature selection. In addition to the four primary prediction models, we also trained LR because of its wide use in similar studies. The results acquired by LR model were considered as the baseline scores for the rest of the models. All the classification models incorporated in the study were evaluated with 10-fold stratified cross validation. Finally, a fraction of data that was not used in cross validation was used to make the final predictions.

3.1. Data Preparation

The dataset used in this study was taken from United Network for Organ Sharing (UNOS), which is a "private, non-profit organization that manages the nation's organ transplant system under contract with the federal government" [72]. The dataset was based on 277316 kidney transplants that took place between 1987 and 2017. The feature set primarily represented the pre-transplant attributes of donors and recipients with respect to their clinical and demographical factors. Few features (such as dates, graft outcome and patient status) were only used in the preprocessing stage to determine the outcome variable, whereas, several identifier features (such as transplant id, donor id, patient id) were immediately removed since they did not have any potential value in making predictions. There was one post-operative feature namely, Delayed Graft Function (DGF), which was recorded immediately after the transplant had been performed. Since our focus was on pre-transplant variables, we also discarded this feature from the feature set. Table 1 provides the list of all the features along with their description which were used in the training process.

Table 1 List of features

Feature Name	Description
pkpra	Peak panel Reactive Antibody – Continuous
REC_TX_PROCEDURE_TY	Type of transplant – Categorical
prevki	Any previous kidney transplant – Categorical
dage	Donor Age – Continuous
dht100	Donor Height – Continuous
rht100	Recipient Height – Continuous
dwt	Donor Weight – Continuous
rwt2	Recipient Weight – Continuous
doncreat	Donor Creatinine Level – Continuous
ecd	Expanded Criteria of Donor – Categorical
dcd	Donation after Cardiac Death – Categorical
dhtn	Donor Hypertension – Categorical
rhtn2	Recipient Hypertension – Categorical
rbmi	Recipient BMI – Continuous
dbmi	Donor BMI – Continuous
cit	Cold Ischemia Timing – Continuous
ragetx	Recipient Age – Continuous
Hlamm	Number of HLA mismatches – Categorical (Paired)
functstat	Functional Status of Recipient – Categorical
drsex	Donor-Recipient Sex – Categorical (Paired)
drrace	Donor-Recipient Race – Categorical (Paired)
drage	Donor-Recipient Age – Categorical (Paired)
rcvd	Recipient Cardiovascular Disease – Categorical
dhcv	Donor Hepatitis C Virus – Categorical
rpvd	Recipient Peripheral Vascular Disease – Categorical
dracesimp	Donor Race – Categorical
rracesimp	Recipient Race – Categorical
rmalig	Recipient Malignancy – Categorical

vintage	Years on dialysis pre-transplant – Continuous
ddm	Donor Diabetes – Categorical
preemptive	Preemptive Transplant – Categorical
rdm2	Recipient Diabetes – Categorical
rcad	Recipient Coronary Artery Disease – Categorical
esrddxsimp	Simplified ESRD diagnosis – Categorical
drcmv	Donor Recipient CMV – Categorical (Paired)
ahd1	Donor Recipient Height difference – Categorical
drwt	Donor Recipient Weight difference – Categorical

3.1.1. Data Cleaning

The first and the foremost step in data preparation was cleaning the dataset with unwanted rows and columns. We removed all the identifier attributes (e.g. donor id, patient id, transplant id, etc.) along with those attributes which had over 50% missing values. Fortunately, only one predictor (Warm Ischemia Timing) was removed for having missing values otherwise all the predictors were stay put. We further fine-tuned our dataset by excluding all the cases pertaining to: a) living kidney donors; b) the recipients below the age of 18; and, c) all the sequential and en-bloc transplants. We set this exclusion criteria based on the suggestion of domain experts and evidence of a similar approach in the following studies [73][74][32]. We further restricted our dataset to only those transplants that took place after the year 1999 because the data was relatively better recorded in terms of features after that time point.

3.1.2. Data Subsets with Overlapped Cohorts

We have already stated in the section above that the overlapped cohorts have been the mainstream binary approach to transform the survival data into a classification problem with different survival intervals. There is an extensive body of research on different organ transplants with overlapped cohorts or one single consolidated cohort with binary classes (graft failed or graft survived), however, to the best of our knowledge, UNOS kidney transplant dataset has not yet been analyzed by overlapped cohorts with a dataset over 10000 observations. Thus the first analysis that we performed in this research was guided

by overlapped cohorts in three different time points. Also, from domain expert's viewpoint it was necessary to analyze overlapped cohorts because they give a more meaningful interpretation of predicting the status of the graft in short, medium, and long term. In non-overlapped cohorts since the data is usually mutually exclusive there is a higher chance that an individual who is classified as a failure in short term would be classified as survived in medium or long term. This is semantically incorrect therefore, to reduce the probability of this happening, overlapped cohorts turn out to be a better approach than non-overlapped cohorts.

The time-points that we decided to formulate our overlapped cohorts were 0-1 years, 0-5 years, and 0-17 years. The first cohort of 0-1 years referred to short term failures, the second cohort of 0-5 years referred to medium term failures and the last cohort referred to long term failures. The outcome variable for each cohort was the status of the graft with two values, namely, graft failed and graft survived. The cases with the failed grafts includes all those transplants where the patient experienced graft failure or death during some point of time. Since these were overlapped cohorts, all the patients who failed or died in the earlier cohorts were also included as failures in the later cohorts. However, the calculation of patients with survived grafts was relatively complex because of the presence of censored data. The easy way to count survived patients could have been based on the assumption that all those patients who did not fail in a certain cohort are presumable survived. However, this assumption creates two problems: a) it does not account for censored data; b) it makes a severe class imbalance (we will discuss that in subsection below). The workaround to deal with this problem was to first remove the censored observations therefore, taking an inspiration from the study by Ali et al. [30] and Kazim et al. [14], we removed the censored observations and calculated the survived patients for each cohort using the equation below. Ali et al. [29] considered only those observations as censored which censored only during the time-cohort under analysis. All other observations were considered as survived grafts. We adopted this strategy to filter down the number of survived grafts but it still created a severe imbalance, therefore we considered the approach by Kazim et al. [13] to further refine the survived class. This study was based on a multiclass analysis on the same UNOS kidney transplant dataset, where they removed all the transplants which were considered survived for less than 7 years after the date of the

transplant. Taking inspiration from this strategy, we removed all the transplants which were considered survived for less than 8 years from the date of transplant.

$$S_i = \begin{cases} Yes\ if\ (days \geq i * 365\ \ and\ graft\ status == \text{failed}) + (days \geq 2920\ \ and\ graft\ status == \text{survived}) \\ No\ otherwise \end{cases}$$

The above equation combines the survived patients from two groups. The first group of survived patients are based on those transplants which are considered survived for all the time-cohorts. These includes patients who survived for at least 8 years (2920 days) after their transplant. The second group was based on transplants which had actually experienced graft failure or death but they did not fail in the time-cohort under analysis instead they failed in a later point of time after the end of the cohort. The i in the equation above indicates the last year of the time cohort (i.e. 1, 5, 17). The Table 2 below shows the class distribution of transplants in the three overlapped and non-overlapped time-cohorts after removing the censored data observations from the survived grafts.

Table 2 Number of failed and survived transplants in three different cohorts before oversampling.

	Overlapped Cohorts	
Time-point	**Failed**	**Survived**
1st cohort	7554	45273
2nd cohort	23475	29352
3rd cohort	37939	14888

3.1.3. Data Subsets with Non-Overlapped Cohorts

Non-overlapped cohorts were based on transplants that were restricted to a subset of an overlapped cohort, only. Unlike overlapped cohorts, the patients with the adverse event in the earlier cohorts were not reconsidered in the later cohorts for analysis in non-overlapped cohorts. However, the number of survived patients were computed using the same formula present in the equation in the section above.

There was mainly one reason to analyze the non-overlapped cohorts and that was to analyze the changing effect of features in different time periods. As mentioned in the background, the changing effect of features have been studied on organ transplant datasets, but they have been studied from the approach of overlapped cohorts. The problem of analyzing the

significance of features with overlapped cohorts is that they also factor-in the significance of the feature from other cohorts and provide a cumulative significance in the end. For example, if a certain feature 'X' has an importance of 80% in short term (0-1) years and 50% in medium term (0-5) years, the probability is high that the value of the feature in short term would be affecting the results in medium term. The degree of its impact can vary with the distribution of the transplants in each cohort and hidden interaction between different features but one can assure that the impact of the former cohort would be present to develop some kind of a bias. In order to reduce this impurity in calculation we developed the non-overlapped cohorts that would help us in bringing about a correct evaluation and discussion on the changing significance of the features.

The non-overlapped cohorts were created in exactly same manner as overlapped cohorts. There were three cohorts representing short, medium and long term transplants however the starting points of the later cohorts were the ending point of the previous cohort. The short term cohort consists of adverse events occurring in the first year of the transplant, the medium term consists of adverse events occurring between 2-5 years of the transplant and all other adverse events after 5 years of the transplant were included in the long term cohort. The distribution of the classes in non-overlapped cohorts are already provided in the Table 3 below.

Table 3 Number of failed and survived transplants in non-overlapped cohorts before oversampling.

	Non-Overlapped Cohorts	
Time-point	**Failed**	**Survived**
1st cohort	7554	45273
2nd cohort	15921	29352
3rd cohort	14464	14888

Interestingly, we did not come across any literature where non-overlapped cohorts were analyzed. The only studies where mutually exclusive cohorts were analyzed were

approached as multiclass problem. Usually the primary objective of all the classification studies on organ transplants is to develop a prediction model. The identification of features has generally been considered a secondary objective therefore researchers have simultaneously computed the significance of the features from the same prediction models that have been fundamentally developed with the aim of predicting the status of the graft in different time-cohorts. Though it is not an incorrect strategy but a better way to do it is by using non-overlapped cohorts.

3.1.4. Data as Multiclass Problem

As stated in the objectives above, this research pursues the problem of predicting the outcome of the graft from both binary as well as multiclass approach. The binary approaches are already explained in the sections above. The goal of the binary approaches has been to predict the status of the graft by the end of each time-cohort, however, the multiclass approach has a slightly different goal. The multiclass classification is performed as an exploratory study on UNOS dataset without changing the classification algorithms which were mainly used for binary approaches. We performed multiclass classification on this dataset to predict the risk of graft failure in different time intervals. The time intervals that we defined in multiclass approach are exactly the same as non-overlapped cohorts, however since the problem is multiclass we did not perform separate analysis for each cohort, instead we considered the whole dataset as one consolidated cohort for the analysis with 3 classes (high, medium and low) representing the risk of failure. The patients included in high risk class were all those who had a short term failure within 1 year from the time of the transplant, the patients included in medium risk class were those who were present in the medium term cohort and the remaining patients with low risk were those who were included in the long term cohort. We restricted the number of groups to only three because: a) similar studies have also divided the dataset in three groups; b) the number of instances in each class would be close which would help the classifiers to analyze the data with reasonable accuracy; and c) the binary cohorts were also divided into three time-points hence the results from the multiclass problem would also help in endorsing the predictions from binary models and vice-versa.

Since it was not a binary approach the interpretation of the survived patients was different. The survived patients in binary approaches were based on the combination of two type of patients: a) who survived for at least 8 years; b) who did not experience the failure in the cohort under analysis. In multiclass approach we only considered the patients who survived for at least 8 years because the other patients were already included in one of the three risk groups. The survived patients were included in the low risk group for the analysis. We took the inspiration from the study by Kazim et al. [13] who used the same approach whilst predicting the risk of graft failure. They included the survived patients in the low risk group who survived for at least 7 years after the transplant.

3.1.5. Class Imbalance

Class imbalance was one of the major challenges in almost all the experiments that we performed in our research. The classification algorithms that we used in our research require decent class balance to perform with reasonable accuracy, thus we applied two different approaches to adjust the class imbalance. Among the two classes in binary approaches, the class referring to survived grafts was extremely high (in most cases) in comparison to the failed grafts, therefore the first approach that we undertook to strike a right balance was to perform a systematic under sampling for the survived class. As per the equation of calculating survived transplants (see section Data Subsets with Overlapped Cohorts), we removed all the transplants which had actually experienced the adverse event in some point of time after the end of the analyzed time-cohort ($days \geq i * 365\ and\ graft\ status ==$ failed), but all those which survived for at least 8 years were taken into the consideration. The Table 4 below shows the distribution of the classes for the overlapped and non-overlapped cohorts after under sampling the survived class.

Table 4 Number of failed and survived transplants in three different cohorts after under sampling

	Overlapped Cohorts		Non-Overlapped Cohorts	
Time-point	**Failed**	**Survived**	**Failed**	**Survived**
1st cohort	7554	14888	7554	14888
2nd cohort	23475	14888	15921	14888
3rd cohort	37939	14888	14464	14888

Even after under sampling the majority class, the imbalance in the 1st cohorts were significantly high, therefore we implemented SMOTE which is an oversampling technique to synthetically increment the minority class. This approach provided us relatively better results in comparison to our second approach (which we will discuss later), but the problem with this approach was its weakness on the scientific front. When the same group of survived patients was used in the analysis, the feature importance scores started to show a significant bias, thus we did not rely on this approach to build the prediction models.

The second approach that we employed in adjusting the balance was alone based on the oversampling technique SMOTE. We applied SMOTE on all the cohorts which required an adjustment however the degree to which SMOTE was used to oversample the minority class was very low. Oversampling the minority class to create an equal balance with majority class meant to generate 600% additional synthetic samples (at least for the 1st cohort), which is a highly erroneous approach and is prone to overfitting. Therefore, we only doubled our minority class to make the cohorts eligible for reasonable classification. Table 5 provides the oversampled results for overlapped and non-overlapped time-cohorts.

Table 5 Class distribution after oversampling

	Overlapped Cohorts		Non-Overlapped Cohorts	
Time-point	**Failed**	**Survived**	**Failed**	**Survived**
1st cohort	**15845**	45273	**15845**	45273
2nd cohort	23475	29352	**21133**	29352
3rd cohort	37939	**27316**	14464	14888

The difference between the classes in the 1st cohort which was the most imbalanced cohort was still 1:3 however that was the best ratio that we could have come up with to prevent overfitting and enable the classifiers to perform decently.

While using SMOTE, we specified the categorical variables and continuous variables beforehand so that the library could generate the accurate samples. The algorithm required to set the number of nearest neighbors for the analysis. We tried with 5, 10 and 15 neighbors

and finally selected 10 neighbors due to a better accuracy and reasonable time that it took to generate the samples.

3.1.6. Categorical vs Continuous Features

The features in our dataset included both categorical and continuous variables, therefore classifiers such as SVM and ANN were incapable to process them in their raw form. Also, the scikit-learn library, which has been the primary tool in our experiments to perform ML, does not recognize categorical variables because of its implementation on numpy arrays [75]. Thus, we used the one-hot encoding module of scikit-learn library to transform the categorical variables into dummy variables (The description of the dummy variables is provided in the Table 21 in appendix). The transformed feature set was then used for training the classifiers mentioned, however, the original feature set was used for computing feature importance scores using H2O package which was also implemented in python.

Although, the transformation of categorical variables into dummy variables was done entirely due to a limitation of the implementation, it helped us to understand the important dummy variables in different time-cohorts when we performed the feature selection. It was an insightful analysis to look into dummy variables rather than just analyzing the integral categorical variable because many a times one dummy variable of a same categorical variable was highly important whereas another dummy variable of that same categorical variable had no importance at all. Thus, conversion into dummy variables made it possible to remove the useless units of categorical variable. It would not have been possible if the categorical variables remained integral. As a future work, the continuous variables can also be broken down into categorical variables with multiple levels where each level would be analyzed as a separate dummy variable.

3.2. Feature Engineering

This phase consists of manipulating the feature set by constructing new features and removing the redundant features for further analysis. Construction of new features was done by pairing variables together (explained below). Next, we implemented a feature selection technique called RFECV to select the significant features from all the experiments that we conducted in this study.

The total number of features that we initially used as the predictors were 45, however not all of them had an equal predictive power. Useless features tend to become a hindrance in the training process if they are not removed beforehand, therefore it was a vital process to identify and remove them from the analysis to train the classifiers without any potential confusion.

The algorithm that we used in the process of performing RFECV and computing feature importance scores was RF. Since RFECV works on coefficients or the importance of the feature, RF was the best method to use. There were three main reasons why we chose RF: a) The algorithm returned decent results in comparison to LR and other methods which had the capability to provide coefficients and feature importance. ANNs and SVMs do not have the capability to return any coefficient or feature importance except the SVM with linear kernels which were incomparable to the performance of RF on our dataset; b) RF was used in similar studies for feature selection; and c) It also has the functionality to provide feature importance scores hence it became easier to analyze whether the RFECV removed any important feature by comparing them with the resultant features after applying RFECV. It would be an interesting insight if a certain feature would be removed by recursive feature elimination but it would not appear as the least important feature on the scale of feature importance scores.

3.2.1. Paired Variables

Paired variables were an interesting addition to our feature set. After exploring the literature, we realized that the dataset has always been analyzed with individual donor and recipient variables. We constructed new variables by merging same type of donor and recipient variable into a single variable. There were around 5 variables (such as sex, age, cmv etc.) which were transformed into a pair hence we named them Paired Variables (PV). We trained classifiers with the individual donor-recipient as well as the paired variable and to our surprise the paired variables performed relatively better. Thus, we entirely removed the individual variables and restricted to only paired variables for the next stage of feature selection by cross validated recursive feature elimination.

3.2.2. Cross Validated Recursive Feature Elimination

Recursive feature elimination is a specific type of backward feature elimination which is based on the coefficients or the feature importance scores rather than p-values. In our work, we set the stopping criteria to a random number of at least 12 features out of the complete feature space, however, the optimal number of features returned by the algorithm were always more than 12 features. All the experiments (for selecting features) that were performed in this study were conducted using 10-fold stratified RFECV. One feature at a time was removed from the feature space to select the optimal set of features which provides the maximum AUROC score. The technique does not select features on the basis of individual feature importance instead it groups all those features together which performs better in consolidation. Therefore, there is a high chance that a feature with relatively high feature importance would be replaced by a less important feature merely because of its insignificance when it is used in combination with other features. We compared the features which were eliminated by cross validated recursive feature elimination with the feature importance scores to understand whether the removed features were least important or not.

3.3. Classification Methods

3.3.1. Logistic Regression

LR has been used in a number of similar studies due to its easy and fast implementation. Interestingly few studies predicting the transplant survival have even quoted this classifier as the best among all the other different classifiers considered in the study [34][30]. We fundamentally used LR to make a comparison with other similar studies and draw a baseline for the rest of the classifiers. We utilized the scikit-learn implementation as it allows to adjust several hyper parameters. We explored different penalties, tolerance levels, solvers and C values in our analysis. The class weight was also available to adjust the class imbalance. The way class weight works is by penalizing the mistakes in samples of class[i] with class_weight[i] instead of 1. For e.g. if a majority class is three times more frequent than the minority class, the class weight would be set to {majority:1, minority:3}. So higher class-weight means we want to put more emphasis on a class. The easier way to manage the weight is to set it to *balanced*. The classifier implicitly replicates the smaller class until it has as many samples as in the larger one. In several experiments that we carried out, the

minority class remained lesser than majority class even after oversampling therefore, this class weight parameter came in very handy. We set the class weight to balanced instead of setting the exact ratios of minority and majority class. The Table 6 provides the parameters for Logistic Regression below.

Table 6 Parameters for Logistic Regression

Parameter	Default Value	Optimized Values
Penalty	L2	L2
Solver	lbgfs	sag
C	1	10
Max Iteration	100	1000, 2000
Class weight	none	none, balanced

3.3.2. Random Forest

RF has been one of the most important ML algorithms throughout in our research. It was used as a standalone classifier, base learner for AdaBoost algorithm and a primary algorithm in the process of feature selection. There were two different implementations of RF that we utilized in this work. The RF implementation of scikit learn was used to build the prediction models and perform recursive feature elimination, however identifying the importance of categorical features was not possible in scikit learn implementation unless the features were transformed into dummy features. To overcome this issue, we utilized the h2o python's implementation to process and determine the importance of features in different time-cohorts. Nonetheless, we also calculated the feature importance scores of dummy features using the Gini impurity and permutation method from scikit learn before using the H2O package for original feature set.

While computing the feature importance scores for dummy features using Gini impurity, we trained the complete dataset and computed the scores without testing it, whereas the feature scores by permutation method were trained and tested in separate dataset with 80-20 split. The feature importance scores using H2O package were computed on the full

dataset (training and testing combined) like we did it for dummy features in scikit learn, but it allowed to set the number of folds for cross validation therefore we analyzed it with 10 different folds.

Both implementations provide same hyper parameters which were optimized during the analysis. The number of trees, max depth, max features and the weight of the class were exhaustively adjusted for the best results. The Table 7 provides the hyper parameters that were selected for the training.

Table 7 Parameters for Random Forest

Parameter	Default Value	Optimized Values
Number of Estimators	100	1200
Max Depth	none	9,12,14
Min Sample Split	2	2,3
Class Weight	none	none, Balanced
Max Features	auto	Sqrt, 14

3.3.3. Adaptive Boosting

Adaptive Boosting implementation by scikit learn library was tested with two weak learners: a) Random Forest, b) Logistic Regression. Logistic Regression did not perform at all and provided poor results on area under ROC, f1 and f1-micro metrics, thus we only proceed with RF in conducting the remaining analyses. RF with the optimized hyper parameters was used to train the boosting classifier. The implementation allowed to select the number of estimators, learning rate, and the algorithm apart from the base estimator. The Table 8 shows the parameters that were tested during the different experiments. Increasing the number of estimators was a memory intensive and time intensive computation. We restricted our analyses with 401 estimators which was the largest number to train the classifier in decent amount of time.

Table 8 Parameters for AdaBoost

Parameter	Default Value	Optimized Values
Number of Estimators	50	374, 401

3.3.4. Artificial Neural Network

Artificial Neural Networks were first implemented using Keras, which is a higher-level deep learning framework for neural network [76]. Keras was selected as the first choice because it abstracts away many details, making code simpler and more concise. It provides a range of hyper parameters (such as activations functions, dropout layers, loss functions etc.) which can be easily tuned to optimize each layer of the neural network. We applied a grid search to evaluate the best loss functions, activation functions, number of neurons, batch and epoch size etc., but the results remained relatively poorer than the other classifiers. AUROC was the primary performance metric to evaluate the results. Scikit-learn's implementation of ANN namely, Multi-layer Perceptron (MLP) classifier was later applied. Since scikit learn is not per se a neural network framework and is built on top of numpy library, there were several limitations in the implementation, which restricted us to change a number of hyper parameters which are usually subject to adjustment in other deep learning frameworks. Nonetheless, this basic implementation proved to be the right choice for our dataset as it returned comparative results to the other classifiers that we used in the analysis.

Neural networks require the variables to be continuous, however the dataset that we were using was mainly based on categorical variables therefore as mentioned in the sections above, we used the transformed dataset with dummy categorical variables to perform the analysis. We experimented with different normalization methods (such as MinMax, Normalizer) provided by scikit learn to normalize the continuous variables but none performed better than the original data. The hyper parameters that we changed in the analysis are given in the Table 9 below.

Table 9 Parameters for Multilayer Perceptron

Parameter	Default Value	Optimized Values
solver	adam	adam
Learning rate	0.0001	e-2,e-5
activation	2	2,3
Learning rate	constant	invscaling, adaptive
Hidden layer sizes	{100,}	{60,30,30,15}

3.3.5. Support Vector Machines

Support Vector Machines have been a useful algorithm in the prediction of survival data. We utilized the scikit implementation of SVM in our experiments. Different kernels (such as linear, radial, sigmoid and polynomial) were tested during the initial training. Linear and sigmoid kernel provided the lowest scores and thus we did not use them in our further analysis though linear kernel took the least amount of time in training the model. Polynomial kernel with degree 2 and 3 were used to train the model. The only experiment where we used polynomial kernel was based on non-overlapped cohorts with graft failures in 1st year and the restricted survived patients' dataset (those who survived for at least 8 years only, provided in Table 4). The results with degree 2 polynomial were poorer than the radial kernel, however, the results for degree 3 remained unrevealed as the memory and time constraint stood out to be a severe hindrance. Radial basis kernel provided as the best results and thus we used it in our analysis. The Table 10 below provides the hyper parameters that were selected for the support vector classifiers in our analysis.

Table 10 Parameters for Support Vector Classifier

Parameter	Default Value	Optimized Values
C	1	50,100
Kernel	rbf	rbf
Gamma	scale	auto, scale
Decision function shape	ovr	ovr, ovo
Class weight	none	none, balanced

3.4. Performance and Evaluation Metrics

3.4.1. Cross Validation

Cross validation is a standard machine learning model evaluation technique used to resample the data during the training and testing phase. The data is broken down into multiple folds where one of the old is kept for testing and the remaining folds are trained to build a ML model. The training process iterates multiple times until all number of folds

are tested one by one. The mean accuracy of all folds are computed at the end. It is a popular method because it is simple to understand and generally results in a realistic estimate of the model's performance than other methods, such as a simple train/test split. Cross validation has two advantages. Firstly, it estimates the generalizability of the algorithm and secondly it ensures that the hyper parameters of the algorithm are optimal.

We used this technique in all of our experiments with stratified 10-folds split. The stratification created each fold with same ratio of observations with a given outcome variable so that minimal biasness between the fold would be guaranteed.

3.4.2. Area Under ROC and F1 Scores

Once cross validation was performed, the results of each classifier were examined on the basis of Area Under Receiver Operating Curve (AUROC) and F1-measure. We preferred area under ROC in comparison to area under Precision Recall Curve (PRC) because of its suitability with the balanced datasets. Except the analyses where we did not oversample the minority class, all the different time-cohorts were almost balanced. We evaluated two kinds of f1 score: a) F1 for the failed grafts; b) F1-Micro which is the micro average of both failed and survived class. F1 score and F1-micro were used in two kinds of situations, respectively: a) where the classes were unbalanced and the minority class happened to be failed grafts; b) where the class were balanced but the ROC score was almost same for all the trained classifiers.

3.5. Feature Importance Scores

Feature importance scores were calculated to analyze the changing relevance of features in different time-cohorts. We calculated these scores by training a RF classifier on the complete dataset. The scores were calculated using two different techniques (See Chapter 2): a) Mean decrease in Impurity b) Permutation feature importance. The mean decrease in impurity could be calculated with two criterions: a) Gini impurity, b) information gain (entropy). We started our analysis on the basis of Gini impurity which is the default criterion for the RF implementation in scikit learn library. The information gain (entropy) criterion was tested during the initial analysis, but it did not make any difference in the performance metrics (such as Area under ROC and F1 score), thus we planned to go forward with the default settings for the remaining analysis. The problem with this

technique was the biasness towards the high cardinality features. In nearly all the experiments, the continuous features were among the leading top ten features in the list thus we applied the permutation based feature importance technique to counteract the bias.

The analysis of these feature importance scores helped us to understand the importance of individual features in each time cohort. It also revealed the changing effect of features over the three time points that we have studied. We also calculated the feature importance scores for categorical features which were already transformed into dummy features. Since there were several dummy features we picked the most important dummy features which were behaving differently in time-cohorts and presented them in the form of bar charts.

4. Results and Discussion

The study was based on a total of 52827 kidney transplants performed from year 2000 to 2017. We performed prediction analysis and identified the changing relevance of features over the period of the study from two different approaches, namely binary and multiclass (as mentioned in Chapter 3). Binary approach remained the main approach in our analysis which was further divided into overlapped cohorts and non-overlapped cohorts. Although, we built the prediction models and analyzed the features using both binary approaches, the main purpose of using overlapped cohorts remained prediction of the status of the graft, whereas, the non-overlapped cohorts were meant to analyze the changing significance of features over three different time-cohorts. Multiclass experiments were also conducted but they were relatively subsidiary to binary class experiments because a) analyzing changing significance of features was not possible through this approach, b) classifiers performed poorly in comparison to the binary approaches. The multiclass experiments were an exploratory work to develop a basic understanding with respect to the classification algorithms on this particular dataset.

We performed all the analyses on complete cases only, where all features were recorded without having any missing value. Before computing these complete cases, we removed all the unnecessary variables which were not a part of prediction model to retain the maximum possible data observations for running the experiments. The dataset consisted of 69 variables including dates, identifiers, pre-transplant, post-transplant and inter-operative variables. We removed all the identifiers and post-transplant variables and focused on the remaining variables for the analysis. One of the pre-transplant variables named as 'Warm Ischemia Timing' was also removed from the dataset for its high sparseness. There were several categorical and numerical variables for the same process. Our preliminary analysis resulted in better performance with numerical variables, so we removed the categorical variables from the feature set where their numerical counterparts (duplicates) were available.

As mentioned in Classification Methods, we applied 5 ML classifier; 2 based on black box (i.e. SVM and ANN), 2 based on ensemble methods (i.e. RF and AdaBoost), and 1 simple linear method (LR) to predict the status of the graft in three-time cohorts. These 5 classifiers

were used in two binary experiments and one multiclass experiment. LR was used to set the baseline for all the other classification methods. In addition to LR, we implemented the conventional CPH and RSF models to set the baseline for our work. The purpose of applying CPH and RSF model was to establish a difference between the performance of classification methods with the traditional and modern regression methods, whereas the baseline set by LR was used to compare the performance of the robust classification methods with the widely used LR in similar studies.

In our analysis, all the computations were carried out using Python, specifically in the SciPy environment using the scikit-learn library [77]. The significance of the features (without any transformation into dummy variables) were computed using H2O package. We ran the simulations on an Intel Core i7 -3770 CPU 3.4GHz PC, equipped with 24.00 GB of RAM Windows 10 64-bit machine.

In the next subsections 4.1, we will discuss the prediction results from the experiments then in section 4.2 we will discuss the importance of the features in the three-time cohorts and finally in section 4.3, we will have a concluding discussion on the overall findings of this research.

4.1. Analysis of prediction models

4.1.1. Baseline Results

Two different baselines were considered in this work to evaluate our solution approach. Our first baseline scores were based on the regression analyses that were performed with one statistical and one ML based regression technique. The second baseline scores were calculated using the LR classifier which is a linear model and has been widely used in similar studies. The rationale to develop these two baselines was to make a comparison with the traditional approaches of survival analysis and understand the degree of improvement in our ML models by comparing them with the LR model.

The baseline results are provided in the sections below. The first two sections show the results for the regression techniques whereas the last section provides the results for the LR classifier applied on overlapped cohorts before and after feature selection.

4.1.1.1. Cox Proportional Hazards Model

We implemented the Cox Proportional Hazards (CPH) Model using 'survival' package in R [78]. The complete dataset - which was considered for the classification approaches - was trained and evaluated on the basis of concordance index (equivalent to the area under the Receiver Operating Characteristic curve [79]). The p-values for the log likelihood test, Wald test, log rank test were also calculated to interpret the significance of the model. The summary of the results including the coefficients, hazard ratios and p-values for each variable are given in the Table 11 below.

Table 11 Results of Cox Proportional Hazards Model

Feature Name	coef	exp(coef)	se(coef)	z	p
drrace2	0.24	1.27	0.02	11.11	2E-16
drrace3	-0.12	0.88	0.06	-1.99	0.046182
drrace4	0.10	1.10	0.01	6.93	4.24E-12
drrace5	-0.13	0.88	0.03	-5.15	2.57E-07
drrace6	0.17	1.19	0.02	7.53	4.98E-14
drrace7	0.02	1.02	0.07	0.29	0.771198
drrace8	0.15	1.16	0.04	3.62	0.000296
drrace9	0.19	1.21	0.06	3.41	0.000658
drsex2	-0.06	0.94	0.02	-3.02	0.002524
drsex3	-0.02	0.98	0.02	-1.08	0.279864
drsex4	0.01	1.01	0.02	0.77	0.442771
ahd11	-0.01	0.99	0.03	-0.51	0.613348
ahd12	-0.03	0.97	0.02	-1.51	0.130171
ahd13	0.03	1.03	0.02	1.46	0.14368
ahd14	0.02	1.02	0.03	0.65	0.515553
drwt2	0.03	1.03	0.02	1.77	0.076369
drwt3	0.07	1.07	0.03	2.62	0.008918
drwt4	0.02	1.02	0.02	1.07	0.286778
drwt5	0.03	1.03	0.03	1.17	0.24051
drage2	-0.01	0.99	0.02	-0.37	0.715183
drage3	0.35	1.42	0.04	9.98	2E-16
drcmv2	-0.02	0.98	0.02	-1.11	0.267119

drcmv3	0.02	1.02	0.02	1.23	0.220213
drcmv4	0.05	1.05	0.02	2.52	0.011894
rpvd	0.19	1.21	0.02	8.88	2E-16
pkpra	0.00	1.00	0.00	13.02	2E-16
REC_TX_PROCEDURE_TY102	-0.02	0.99	0.01	-1.46	0.144671
prevki	0.12	1.13	0.02	6.70	2.09E-11
dage	0.01	1.01	0.00	11.09	2E-16
dht100	0.00	1.00	0.00	-6.10	1.06E-09
dwt	0.00	1.00	0.00	-1.38	0.167742
doncreat	0.02	1.02	0.01	4.51	6.62E-06
ecd1	0.07	1.07	0.02	3.89	0.000102
rht2100	0.00	1.00	0.00	2.90	0.003741
rwt2	0.00	1.00	0.00	1.80	0.072567
rbmi2	0.00	1.00	0.00	1.55	0.121616
cit	0.00	1.00	0.00	7.12	1.09E-12
ragetx	0.01	1.01	0.00	19.90	2E-16
hlamm1	0.03	1.03	0.03	0.78	0.436263
hlamm2	0.17	1.19	0.03	5.98	2.30E-09
hlamm3	0.18	1.20	0.02	7.98	1.42E-15
hlamm4	0.19	1.21	0.02	9.10	2E-16
hlamm5	0.22	1.25	0.02	10.76	2E-16
hlamm6	0.23	1.26	0.02	10.42	2E-16
dbmi	0.00	1.00	0.00	1.16	0.245633
functstat2	0.08	1.08	0.02	4.77	1.85E-06
functstat3	0.16	1.18	0.02	10.75	2E-16
functstat4	0.26	1.30	0.02	14.81	2E-16
functstat5	0.25	1.28	0.02	12.95	2E-16
functstat6	0.45	1.56	0.04	11.90	2E-16
functstat7	0.24	1.27	0.05	4.42	1.01E-05
functstat8	0.53	1.70	0.13	4.03	5.69E-05
functstat9	0.90	2.46	0.08	11.52	2E-16
functstat10	0.86	2.37	0.12	7.47	7.90E-14
rhtn2	-0.02	0.98	0.02	-0.92	0.358774
rcvd	0.04	1.04	0.03	1.58	0.114486

rmalig	0.15	1.16	0.02	6.61	3.98E-11
dhtn21	0.08	1.08	0.01	5.65	1.59E-08
ddm1	0.18	1.20	0.02	8.79	2E-16
dhcv	0.35	1.42	0.03	11.70	2E-16
dcd1	0.12	1.12	0.02	6.41	1.46E-10
preemptive2	0.24	1.28	0.02	11.95	2E-16
esrddxsimp2	0.14	1.15	0.02	5.93	3.03E-09
esrddxsimp3	-0.29	0.75	0.02	-12.48	2E-16
esrddxsimp4	0.10	1.10	0.02	6.01	1.88E-09
esrddxsimp5	0.08	1.08	0.02	4.26	2.08E-05
rdm21	0.18	1.19	0.02	8.92	2E-16
rcad2	0.05	1.05	0.02	3.15	0.001619
vintage	0.03	1.03	0.00	16.84	2E-16

Concordance= 0.623 (se = 0.002)

Likelihood ratio test= 5245 on 69 df, p=<2e-16

Wald test = 5237 on 69 df, p=<2e-16

Score (logrank) test = 5351 on 69 df, p=<2e-16

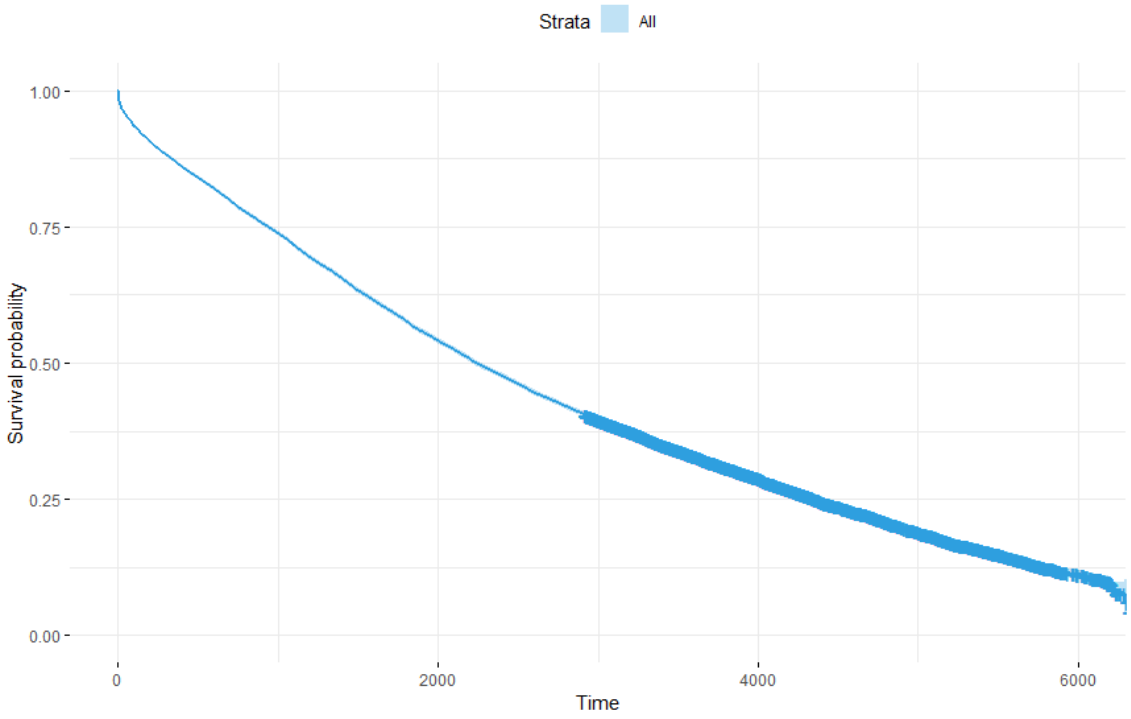


Figure 2 Probability of Survival by Cox Proportional Hazards Model

The concordance index that we received after running cox proportional hazards model was 0.623. The p-values for all three overall tests (likelihood, Wald, and score) were significant, indicating that the model was significant. In addition to that, half of the covariates as shown in the Table 11 were also significant ($p < 0.05$), however, half of them were insignificant ($p>0.05$). The study by Yoo et al. [80] received the score between 0.6-0.63 for their conventional cox model. Although, their dataset was relatively smaller than ours, it gave us a superficial understanding of an average strength of cox models on transplant datasets. The probability of survival was also generated by the trained CPH model as shown in the Figure 2 above. The diminishing graph shows the decreasing probability of survival with time.

4.1.1.2. Random Survival Forests

We implemented Random Survival Forest (RSF) using RandomForestSRC package in R. The summary of the developed model is given below.

Training Dataset Summary	
Sample size	36979

Number of deaths	26576
Number of trees	1200
Forest terminal node size	15
Average no. of terminal nodes	941.8933
No. of variables tried at each split	7
Total no. of variables	37
Resampling used to grow trees	swor
Resample size used to grow trees	23371
Analysis	RSF
Family	surv
Splitting rule	logrank *random*
Number of random split points	3
Error rate	37.88%
Testing Dataset Summary	
Sample size of test (predict) data	15848
Number of deaths in test data	11363
Number of grow trees	1200
Average no. of grow terminal nodes	941.8933
Total no. of grow variables	37
Resampling used to grow trees	swor
Resample size used to grow trees	10016
Analysis	RSF
Family	surv
Test set error rate	37.98%

The number of trees and node depth which were two of the hyper parameters were set to 1200 and 14, respectively. The values were selected based upon the best c-index. The concordance index that we received after making the prediction on the test set was 0.6202 which is approximately similar to the c-index that we got from CPH model.

4.1.1.3. Logistic Regression

The LR score before and feature selection are provided in the Table 12 and

Table 13, respectively. Interestingly, the pre-feature selection scores for LR were relatively better than the scores after feature selection. It proves that the feature selection by RF did not affect the LR models, positively. The cross validated ROC score for 1st cohort dropped down from 67% to 61% and the score for 3rd cohort came down from 76% to 71%. However, all scores for 2nd cohort remained intact after removing the unwanted features with no changes at all. We have compared the performance of LR with other classifiers in much more detail in the sections to follow.

Table 12 Logistic Regression Scores before Feature Selection

Model	CV F1	CV F1_Micro	CV ROC	Test F1	Test F1_Micro	Test ROC
1st cohort						
LR	0.47	0.62	0.67	0.47	0.62	0.63
2nd cohort						
LR	0.58	0.61	0.65	0.58	0.61	0.61
3rd Cohort						
LR	0.72	0.69	0.76	0.71	0.69	0.7

Table 13 Logistic Regression Scores after Feature Selection

Model	CV F1	CV F1_Micro	CV ROC	Test F1	Test F1_Micro	Test ROC
1st cohort						
LR	0.41	0.57	0.61	0.58	0.41	0.575
2nd cohort						
LR	0.58	0.61	0.65	0.58	0.61	0.61
3rd Cohort						
LR	0.71	0.68	0.76	0.71	0.69	0.68

4.1.2. Overlapped Cohorts

We used overlapped cohorts to predict the status of the graft in the three defined time-cohorts. One of the objectives of this research was to build the prediction models using

supervised ML algorithms; we reached this objective by preparing the dataset in the form of overlapped cohorts. Multiple classifiers were used in our experiments. The best four classifiers that stood out during the initial runs were later used throughout the research. LR which is a simple linear classifier was considered as the baseline because of its extensive use in the literature. The results that we received from overlapped cohorts are explained under three subheadings, below. We first provided the preliminary results based on the complete dataset without optimizing the hyper parameters followed by the final prediction models. The final models were developed by optimizing the hyper parameters and selecting the important features using RFECV. The last subsection discusses the changing effects of the features. This analysis was particularly performed to understand the effect of overlapped cohorts on the significance of the features and why non-overlapped cohorts were a better way to analyze the significance.

4.1.2.1. Preliminary Results

The preliminary results were computed with complete feature set after minor parameters' tweaking. Except LR, the default settings of the classifiers were resulting in very poor scores, thus few parameters such as number of trees in RF, number of layers and neurons in ANN and the type of kernel for SVC were fixed in preliminary analysis and they remain unchanged till the development of final prediction models. According to Table 2 which shows the distribution of the classes in three time-cohorts, the first cohort and third cohort needed a class adjustment but the second cohort had an adequate balance. Hence we used SMOTE to oversample the *graft failed* class in first cohort and *survived* class in third cohort to bring the class distribution into a state of consistency for classifiers to perform decently. Since the *failed grafts* in 1st cohort were still three times less than *survived* class even after oversampling, we utilized the class weight feature in scikit-learn library during the process of model training.

The main evaluation metrics to select the winning model in our analysis was Area under ROC, f1 and f1 micro scores. The results provided in the Table 14 below are generated after performing 10-fold stratified cross validation on 80% of the dataset. The remaining 20% of the dataset was used in predictions on the unseen test set. The results from the test set are also provided in Table 14. We have included the results of LR in these tables again

to make an easy comparison in the discussion ahead. As evident from the results, oversampling made a remarkable difference in both cohorts, therefore we preferred oversampled cohorts for developing final prediction models.

Among all the classifiers, AdaBoost and SVC were most benefitted with oversampling; as a clear spike in AUROC score can be seen in the results in both first and third cohort, whereas LR remained the least affected classifier with oversampling. The best results in the 1st cohort were produced by Support Vector Machine with 85% on AUROC followed by AdaBoost with 82%. Other than AUROC, the metrics provided close scores with not much of any interesting insight. The second cohort was the largest cohort based on entirely original data without any synthetic oversampling. Since the dataset was decently balanced and AUROC scores were almost same across all the developed models, the metric of interest becomes f1 and f1-micro scores. In 2nd cohort, the LR and AdaBoost competed to be the winning models as they both have the best f1 and f1-micro scores. The f1-micro score was 61% for LR but a percent more (62%) for AdaBoost, whereas the F1 score was 58% for LR and 53% for AdaBoost.

The third cohort provided the best results with AdaBoost with 84% on AUROC as well as all the other metrics followed by ANN with 80% on AUROC.

The testing set as shown in Table 14 below justified the results from the cross validated training set: a) SVC provided the maximum AUROC score in the 1st cohort after oversampling followed by AdaBoost; b) LR performed the worst among all the classifiers; c) the performance of the all the classifiers in 2nd cohort was almost same.

Table 14 Overlapped Cohorts Baseline Cross Validation Scores

Model	CV F1	CV F1_Micro	CV ROC	T F1	T F1_Micro	T ROC
1st Cohort without oversampling						
Random Forest	0.07	0.85	0.62	0.11	0.85	0.53
ADA with RF	0.09	0.85	0.57	0.12	0.85	0.53
SVC	0.04	0.85	0.55	0.04	0.85	0.51
LR	0.31	0.61	0.65	0.3	0.61	0.6
ANN	0.00	0.86	0.65	0	0.86	0.5

1st cohort with oversampling						
Random Forest	0.53	0.77	0.76	0.53	0.77	0.69
ADA with RF	0.67	0.87	0.82	0.69	0.88	0.76
SVC	0.7	0.87	0.85	0.73	0.88	0.8
LR	0.47	0.62	0.67	0.47	0.62	0.63
ANN	0.42	0.79	0.75	0.41	0.78	0.62
2nd cohort (oversampling not required)						
Random Forest	0.48	0.62	0.66	0.49	0.63	0.61
ADA with RF	0.53	0.62	0.65	0.53	0.62	0.61
SVC	0.47	0.62	0.66	0.48	0.63	0.61
LR	0.58	0.61	0.65	0.58	0.61	0.61
ANN	0.54	0.62	0.66	0.48	0.62	0.59
3rd Cohort without oversampling						
Random Forest	0.84	0.73	0.7	0.84	0.73	0.54
ADA with RF	0.84	0.74	0.7	0.84	0.74	0.57
SVC	0.84	0.72	0.67	0.84	0.72	0.5
LR	0.71	0.63	0.68	0.71	0.63	0.63
ANN	0.84	0.73	0.7	0.83	0.73	0.6
3rd Cohort with oversampling						
Random Forest	0.75	0.71	0.78	0.75	0.72	0.71
ADA with RF	0.82	0.78	0.84	0.82	0.78	0.77
SVC	0.73	0.71	0.78	0.73	0.71	0.71
LR	0.72	0.69	0.76	0.71	0.69	0.7
ANN	0.78	0.73	0.8	0.71	0.69	0.7

4.1.2.2. Results after Feature Selection

We performed RFECV on each of the three cohorts to select the best set of features. Since the technique was implemented in scikit-learn, we were bound to only use it on the dataset based on dummy variables. The training and testing scores that we received after the process of feature selection are shown in the Table 15 below. The scores of the features which were selected after feature elimination process in all three cohorts are shown in

Figure 3, Figure 4 and Figure 5 below. We will discuss these scores during the analysis of the changing effects of features in section Analysis of changing effects of features.

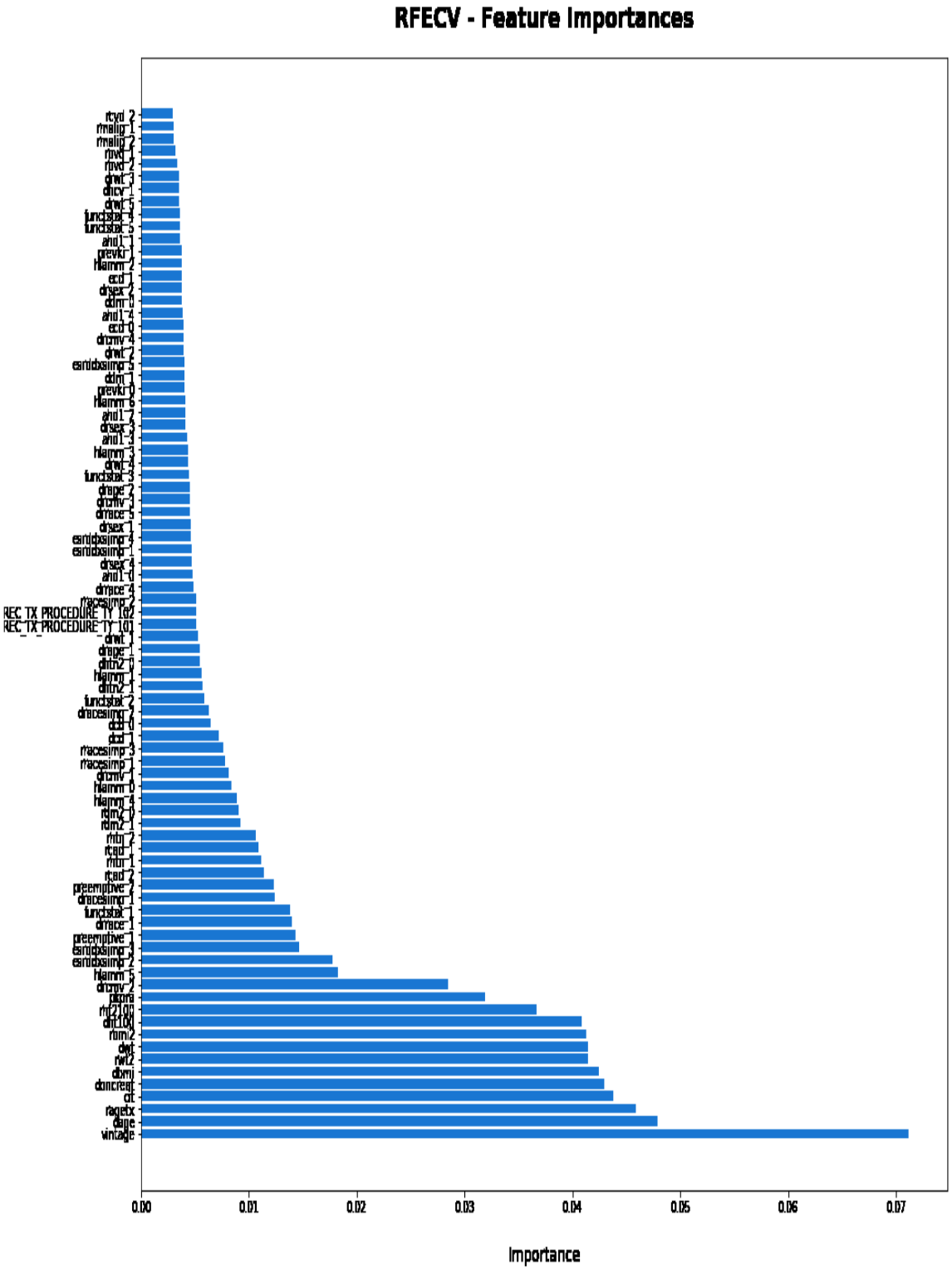


Figure 3 1st cohort scores after recursive feature elimination

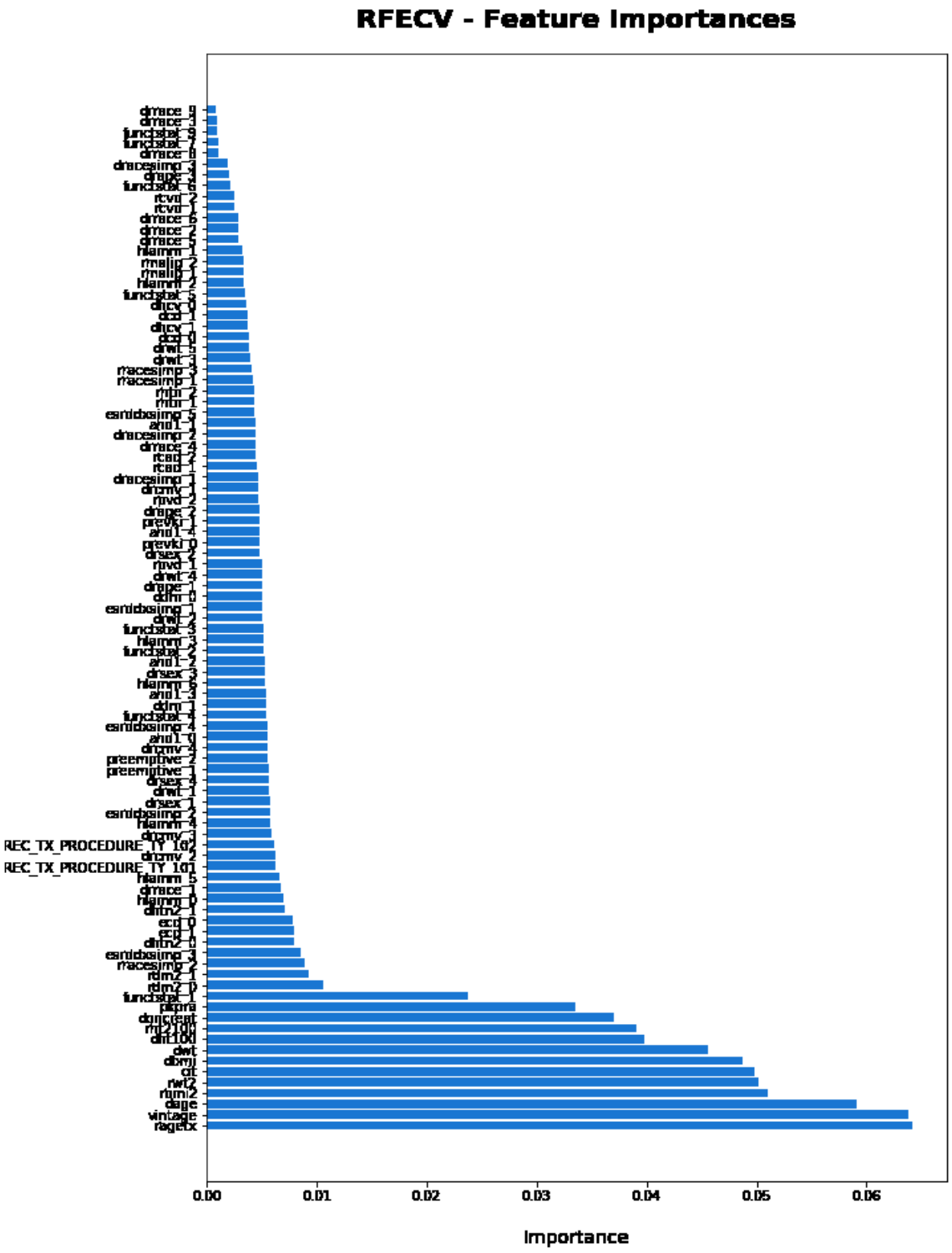


Figure 4 2nd cohort scores after recursive feature elimination

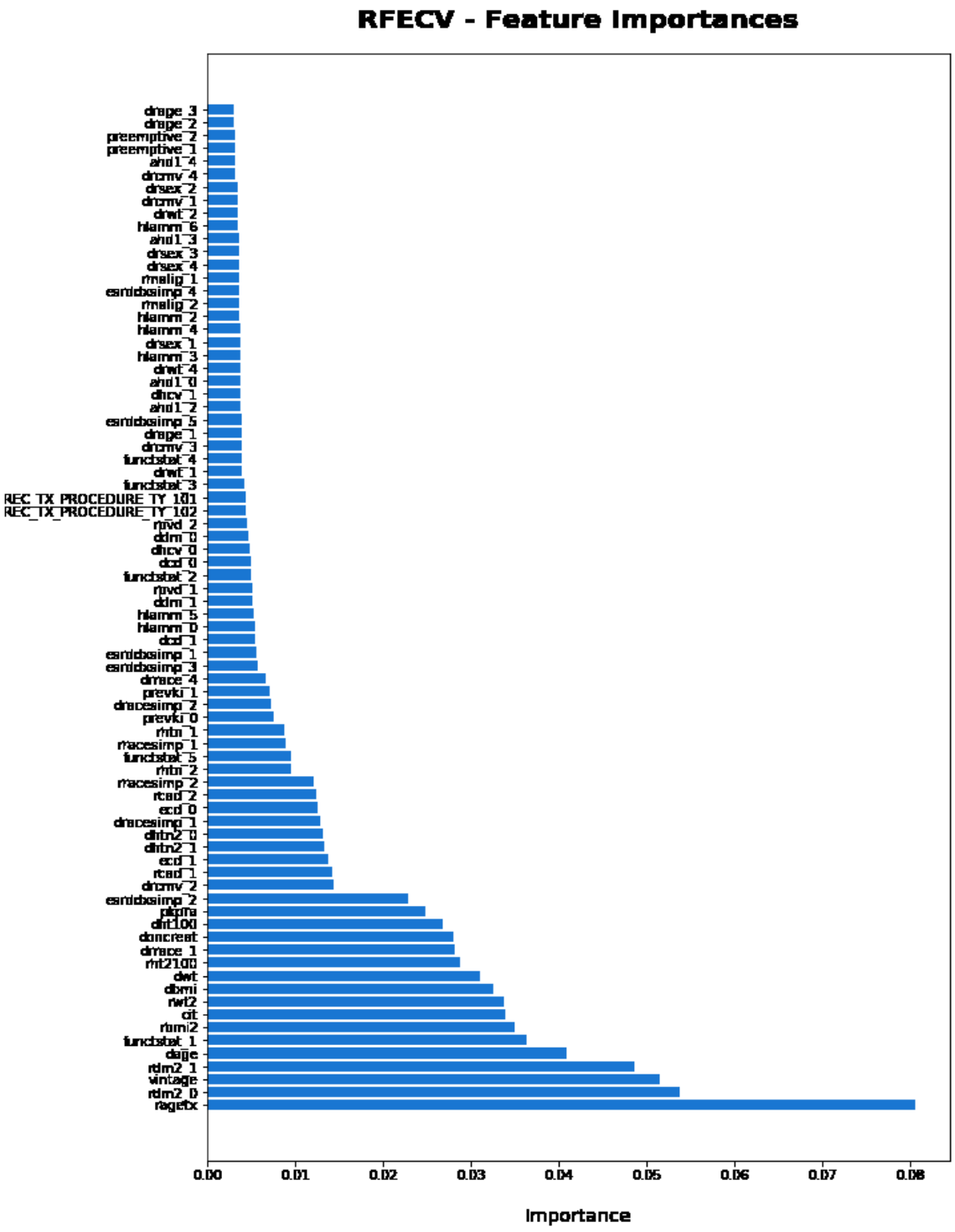


Figure 5 3rd cohort scores after recursive feature elimination

Among the five classifiers, we focused on the AUROC scores for the 1st cohort and noticed that RF showed an improvement of 1% (from 76% to 77%) on cross validated training set whereas AdaBoost showed an improvement of 2% (75% to 77%) on testing set. SVC did

not exhibit any significant change in both the datasets and remained the winner for the first cohort.

The second cohort did not show any significant improvements in scores except that several dummy variables were removed after the process of feature selection. SVC and AdaBoost remained the leading classifiers with 66% AUROC score on training dataset, however, LR performed the best on the test set with 61% AUROC score.

The third cohort showed an improvement in nearly all the classifiers except LR. The leading classifier still remained AdaBoost with 84% AUROC score followed by ANN with 81% AUROC on training dataset.

Table 15 Scores after oversampling the overlapped cohorts

Model	CV F1	CV F1_Micro	CV ROC	T. F1	T. F1_Micro	T. ROC
1st cohort						
Random Forest	0.614	0.79	0.774	0.534	0.733	0.686
ADA with RF	0.68	0.87	0.824	0.7	0.87	0.77
SVC	0.68	0.87	0.851	0.72	0.87	0.794
LR	0.41	0.57	0.61	0.58	0.41	0.575
ANN	0.45	0.78	0.74	0.41	0.78	0.62
2nd cohort						
Random Forest	0.49	0.62	0.66	0.47	0.62	0.60
ADA with RF	0.53	0.62	0.65	0.52	0.62	0.60
SVC	0.47	0.62	0.66	0.47	0.62	0.60
LR	0.58	0.61	0.65	0.58	0.61	0.61
ANN	0.51	0.62	0.65	0.52	0.62	0.60
3rd Cohort						
Random Forest	0.76	0.72	0.79	0.78	0.72	0.70
ADA with RF	0.82	0.78	0.84	0.83	0.79	0.78
SVC	0.75	0.72	0.79	0.76	0.72	0.72
LR	0.71	0.68	0.76	0.71	0.69	0.68
ANN	0.78	0.74	0.81	0.77	0.74	0.73

4.1.3. Non-Overlapped Cohorts

4.1.3.1. Preliminary Results

Baseline for the non-overlapped cohorts were drawn in the exact same manner like overlapped cohorts, but since the aim of non-overlapped cohorts were not the development of prediction models, the baseline scores were not separately provided in the section 4.1.1.3. Logistic Regression. The scores for the training and testing datasets are provided in the Table 16, below. As the distribution of transplants in 1st cohorts of both overlapped and non-overlapped approaches was same, we did not explain the results in this section again. Contrary to 2nd cohort in overlapped cohorts, the 2nd cohort in non-overlapped cohorts needed a class adjustment. Interestingly, all the classifiers showed an improvement after oversampling except LR, which deteriorated by 2% on training set, however, a significant improvement was observed on the testing dataset. AdaBoost remained the winning model in 2nd cohort with 74% AUROC on cross validated training set and 67% AUROC on testing set, whereas all other models remained in the range of 67%-69%. The third cohort had a close competition among the models on both training and testing sets. The ANN model provided the best score 69% AUROC on training and 64% AUROC on testing. All the remaining models were stopped with 68% AUROC on training sets and 62-64% AUROC on testing sets.

Table 16 Preliminary results for non-overlapped cohorts

Model	CV F1	CV F1_Micro	CV ROC	T. F1	T. F1-Micro	T. ROC
1st Cohort without oversampling						
Random Forest	0.07	0.85	0.62	0.11	0.85	0.53
ADA with RF	0.09	0.85	0.57	0.12	0.85	0.53
SVC	0.04	0.85	0.55	0.04	0.85	0.51
LR	0.31	0.61	0.65	0.3	0.61	0.6
ANN	0.00	0.86	0.65	0	0.86	0.5
1st cohort with oversampling						
Random Forest	0.53	0.77	0.76	0.53	0.77	0.69
ADA with RF	0.67	0.87	0.82	0.69	0.88	0.76
SVC	0.7	0.87	0.85	0.73	0.88	0.8

LR	0.47	0.62	0.67	0.47	0.62	0.63
ANN	0.42	0.79	0.75	0.41	0.78	0.62
2nd Cohort without oversampling						
Random Forest	0.41	0.65	0.64	0.41	0.65	0.58
ADA with RF	0.27	0.67	0.64	0.27	0.66	0.55
SVC	0	0.65	0.64	0.00	0.65	0.50
LR	0.51	0.61	0.64	0.61	0.59	0.45
ANN	0.32	0.67	0.65	0.32	0.66	0.56
2nd Cohort with oversampling						
Random Forest	0.57	0.64	0.68	0.58	0.64	0.64
ADA with RF	0.56	0.7	0.74	0.57	0.70	0.67
SVC	0.54	0.66	0.69	0.46	0.69	0.64
LR	0.55	0.59	0.62	0.55	0.59	0.59
ANN	0.48	0.63	0.67	0.55	0.59	0.59
3rd Cohort (oversampling not required)						
Random Forest	0.61	0.63	0.68	0.61	0.63	0.63
ADA with RF	0.62	0.63	0.68	0.62	0.63	0.63
SVC	0.61	0.64	0.68	0.61	0.64	0.64
LR	0.63	0.64	0.68	0.61	0.62	0.62
ANN	0.63	0.63	0.69	0.61	0.64	0.64

4.1.3.2. Results after Feature Selection

As mentioned in the previous section, the results for the 1st cohort were same as the overlapped cohorts hence they are not discussed in this section again. The RFECV discarded around 14 features from 1st cohort, 26 features from the 2nd cohort and 18 features from the 3rd cohort out of 98 features, hence the number of features in the 1st, 2nd cohort and 3rd cohort were 84, 72 and 80, respectively. The Figure 6 and Figure 7 shows the importance of the features which remained in the feature set after performing recursive feature elimination. It is interesting to see that the continuous features have been the most important features in all three cohorts.

The categorical features which were entirely removed from the 2nd cohort were *rmalig, rpvd,dhcv,* whereas, the number of dummy features which were removed after the feature

selection process were 23. In 3rd cohort, no categorical feature was completely removed, however 18 dummy features which were a part of one of the categorical variables were eliminated after recursive feature elimination. Detailed discussion about the importance of the features with respect to their changing relevance and the implication of potential bias coming from continuous features is done in next section.

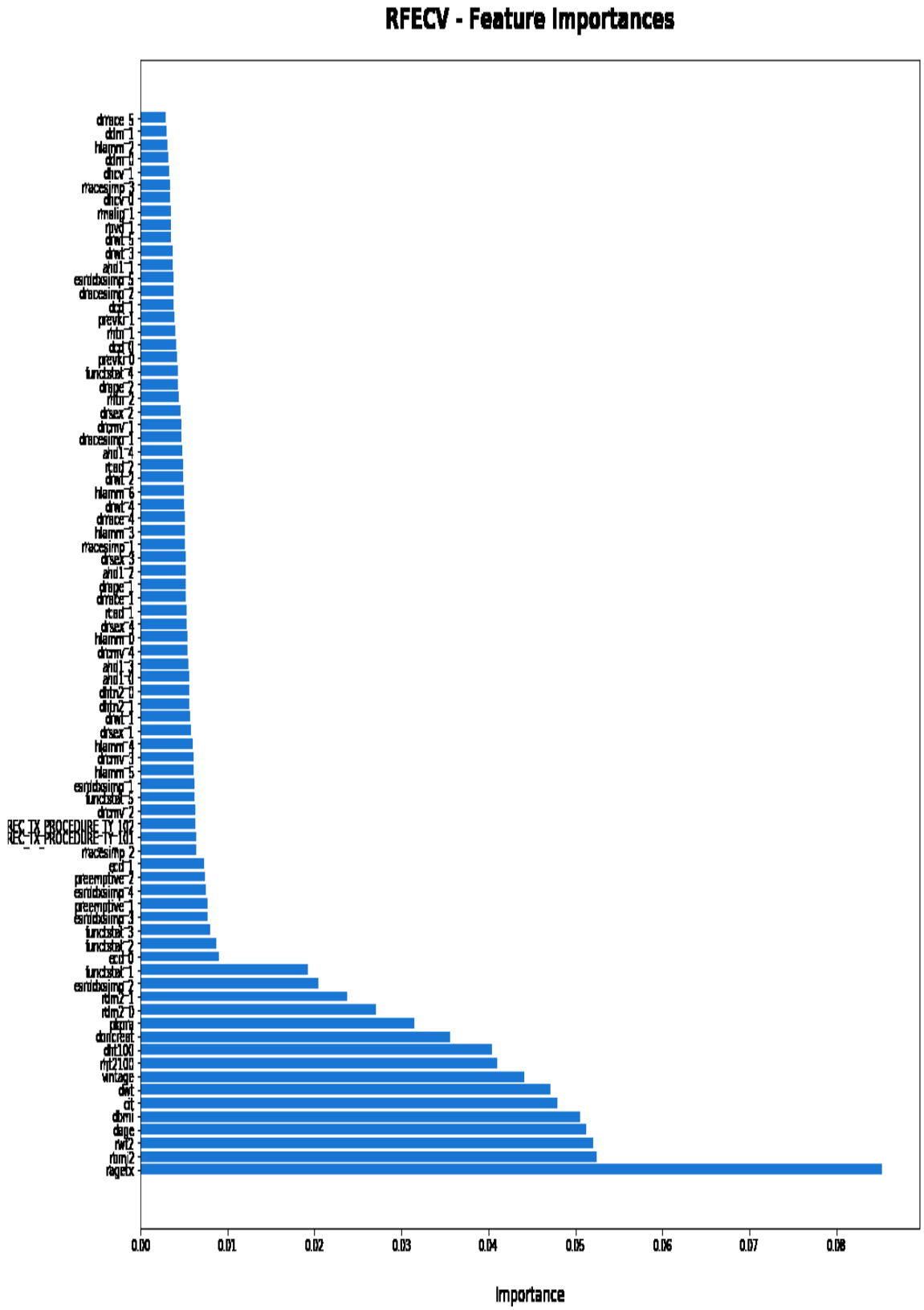


Figure 6 Recursive Feature Elimination for 2nd Cohort

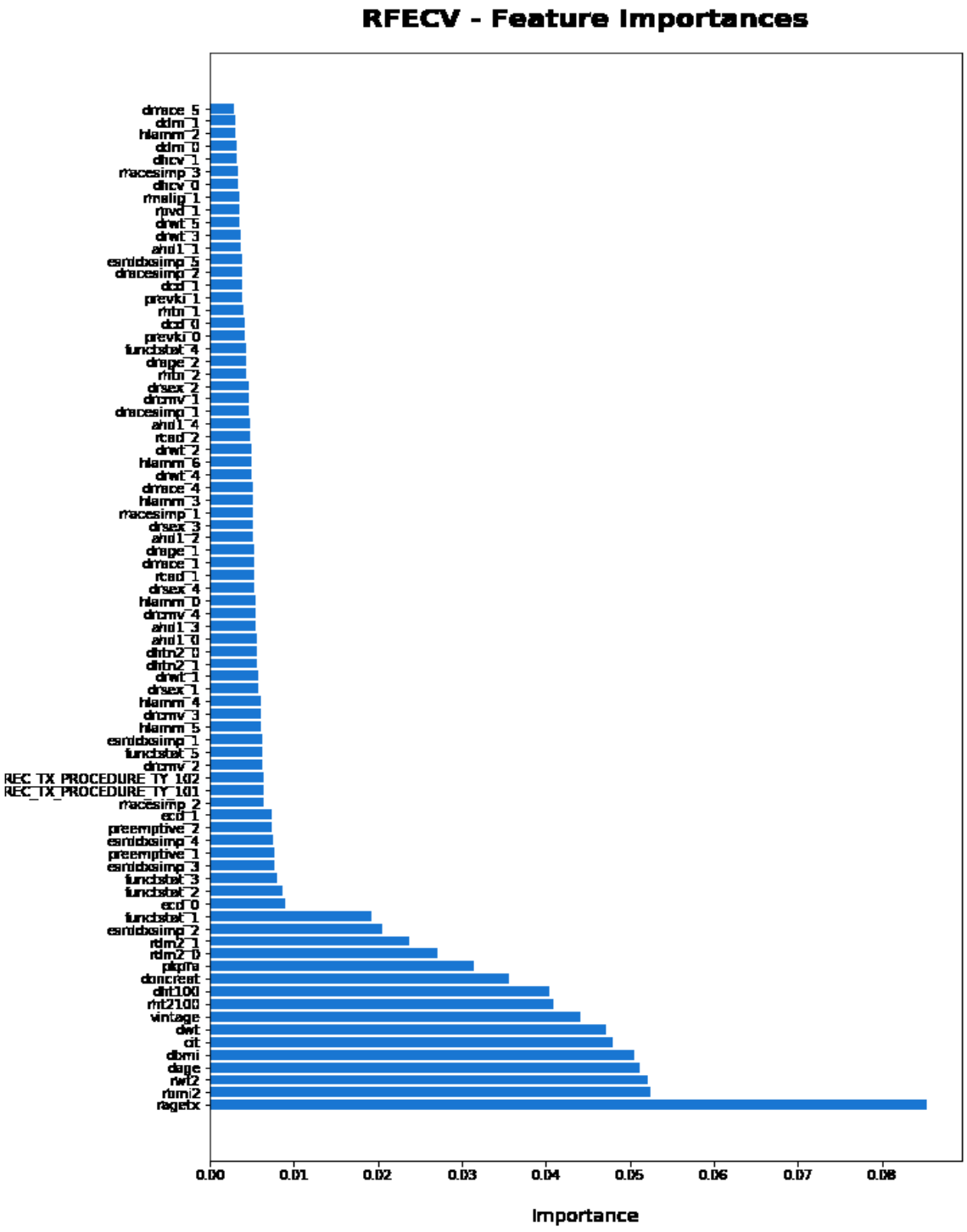


Figure 7 Recursive Feature Elimination for 3rd Cohort

In the 2nd cohort, RF and AdaBoost performed best with 73% AUROC on training dataset, whereas Adaboost provided 67% AUROC on test dataset which was the best amongst all

the classifier. The 3rd cohort provided relatively close results with almost all the classifier performing similarly with 68% AUROC on training set and 64% AUROC on test set.

Table 17 Scores after feature selection for non-overlapped cohorts

Model	CV F1	CV F1_Micro	CV ROC	T. F1	T. F1-Micro	T. ROC
1st cohort						
Random Forest	0.614	0.79	0.774	0.534	0.733	0.686
ADA with RF	0.68	0.87	0.824	0.7	0.87	0.77
SVC	0.68	0.87	0.851	0.72	0.87	0.794
LR	0.41	0.57	0.61	0.58	0.41	0.575
ANN	0.45	0.78	0.74	0.41	0.78	0.62
2nd cohort						
Random Forest	0.55	0.70	0.73	0.58	0.64	0.64
ADA with RF	0.56	0.70	0.73	0.57	0.70	0.67
SVC	0.44	0.68	0.68	0.46	0.69	0.64
LR	0.55	0.59	0.62	0.55	0.59	0.59
ANN	0.55	0.59	0.62	0.55	0.59	0.59
3rd Cohort						
Random Forest	0.61	0.63	0.68	0.61	0.64	0.64
ADA with RF	0.62	0.63	0.68	0.61	0.65	0.64
SVC	0.59	0.63	0.68	0.61	0.64	0.64
LR	0.62	0.62	0.66	0.61	0.62	0.63
ANN	0.60	0.63	0.68	0.57	0.64	0.64

4.1.4. Multiclass Results

The multiclass approach did not return any interesting results. Since oversampling had already been established as the right technique for our dataset, we applied the same 5 classifiers that we have used in all other experiments to build the prediction models using the oversampled dataset. The Table 18 below shows the number of instances in each class before and after oversampling. We oversampled each minority cohort class with a certain number of samples that would not generate too much noise in the data. The high risk class

was almost doubled, whereas, the medium risk class had an additional 4000 samples which are nearly one-quarter of original samples in the class. The idea was to generate the minimal amount of synthetic samples by which the classifiers would be able to provide decent scores.

Table 18 Class distribution in multiclass approach

Classes	Without Oversampling	With Oversampling
High Risk	7554	15000
Medium Risk	15921	20000
Low Risk	29352	29352

Unfortunately, none of the classifiers were even able to score at par with the baseline set by regression methods. Recursive feature elimination happened to improve the results but the improvement was not in terms of an increment in scores rather the feature set experienced a reduction of 52 features. However, after feature selection only RF showed an improvement of 4% from 49% to 53%. For all other classifiers the performance remained nearly indifferent.

The Figure 8 below shows the importance of remaining features after performing RFECV. The top 10 features in the list belongs to the type of continuous features in our feature set. This high importance of continuous features over categorical features was caused by an intrinsic bias of RF algorithm. We have discussed this in detail in the next section.

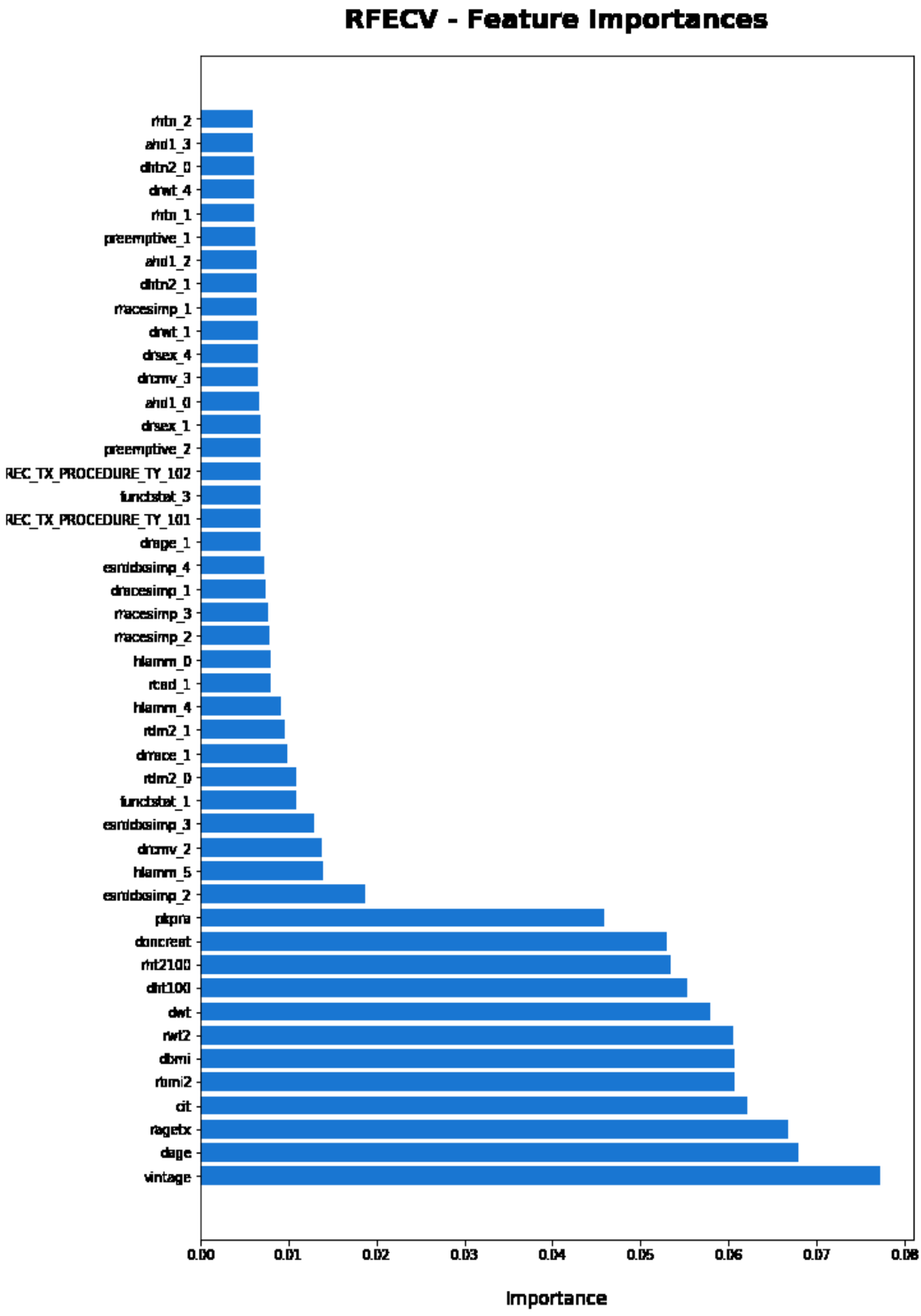


Figure 8 Scores after Feature Selection for multiclass approach

The Table 19 below shows the results that were acquired through multiclass approach. Since there were more than two classes, the AUROC score could not be calculated, we evaluated the models on the basis of f1-micro scores only.

Table 19 Results for multiclass approach

Model	CV F1-Micro	Test F1-micro
Before Feature Selection		
Random Forest	0.49	0.49
ADA with RF	0.61	0.62
SVC	0.61	0.63
LR	0.46	0.47
ANN	0.51	0.5
After Feature Selection		
Random Forest	0.53	0.54
ADA with RF	0.61	0.62
SVC	0.62	0.63
LR	0.46	0.46
ANN	0.49	0.5

The highest score was produced by SVC with 63% on testing set and 62% on training dataset. AdaBoost also provided 61% on training dataset however it lagged behind 1% on testing set with 62% from SVC. Throughout our experiments in this work, the training and testing scores were different with training scores overshadowing the testing scores however, whilst performing a multiclass classification the difference shrunk drastically. In fact, the testing scores for LR, AdaBoost and SVC exceeded the training score which was surprisingly a unique phenomenon for our dataset.

4.2. Analysis of changing effects of features

The second objective of this research was to analyze the relevance of the features over the period of time. The main approach that was used for the analysis was based on non-overlapped cohorts. However, we also reported the results with overlapped cohorts in order to analyze the difference between the two approaches. We first calculated the feature

importance scores based on mean decrease impurity (Gini) using the H2O implementation of Random Forest for both binary approaches.

The Figure 9 and Figure 10 below shows the importance of the features in overlapped cohorts and non-overlapped cohorts.

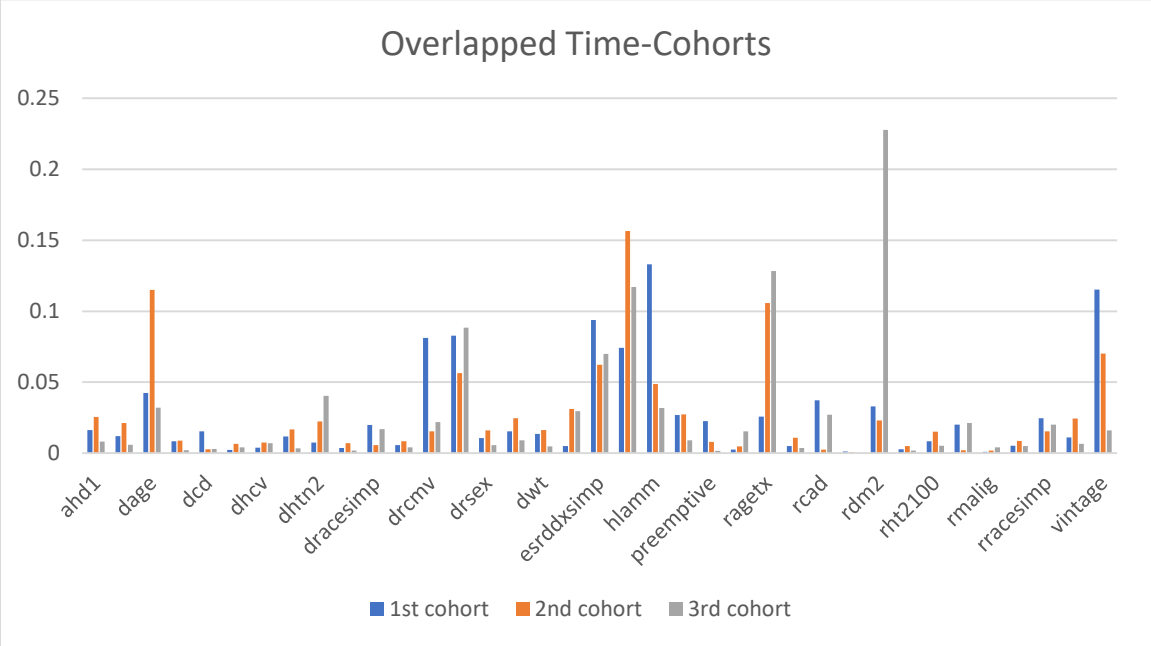


Figure 9 Changing relevance of features based on overlapped time-cohorts

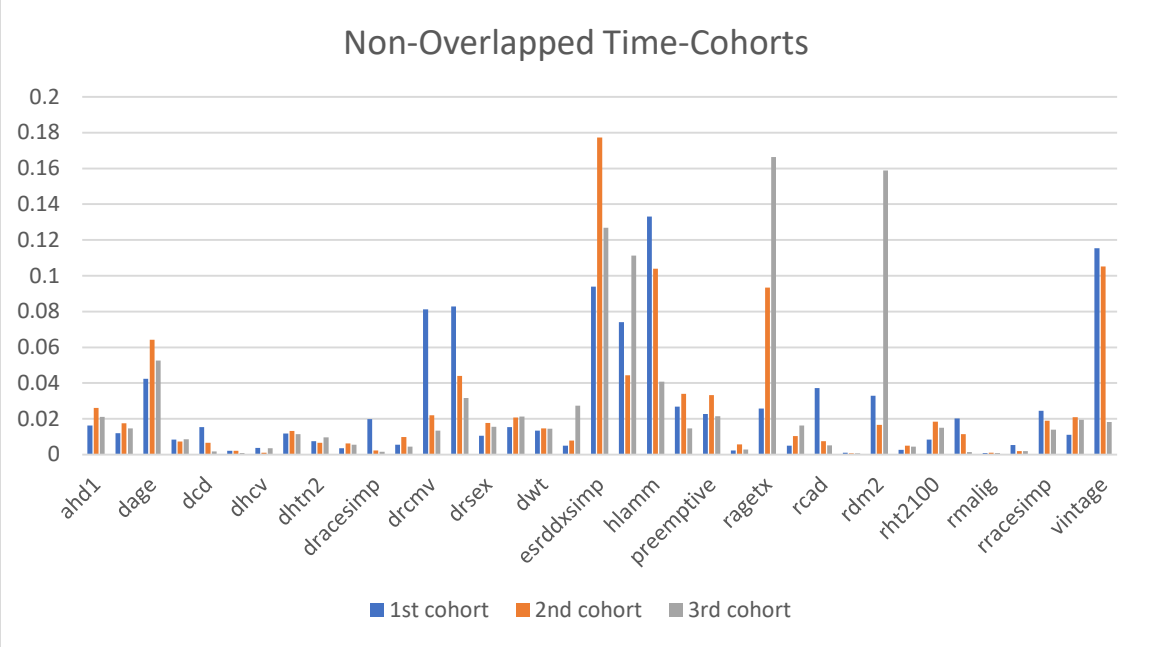


Figure 10 Changing relevance of features based on non-overlapped time-cohorts

We categorized the features into three groups based on their importance (i.e. high, medium and low). The high group includes all the features which have the importance over 10% (>0.1), the medium group has all the feature which have importance between 5% to 10% (0.05-0.1) and the remaining falls in the last group which denotes the features with low importance. Since the data observations in 1st cohort of both approaches was same, the importance of the features also remained indifferent. According to our definition of highly important features, HLA mismatches (HLAMM) and the Years-On-Dialysis-Before-Transplant (VINTAGE) were the most important features for 1st cohort. The importance of both variables decreased with the passage of time in both overlapped and non-overlapped cohorts. The features having medium importance in 1st cohort were donor-recipient CMV status (DRCMV), donor-recipient race (DRRACE), end stage renal disease (ESRDDXSIMP) and functional status of the recipient (FUNCTSTAT). In both overlapped and non-overlapped cohorts, the importance of DRCMV dropped from medium to low significance group, however DRRACE showed a downward trend in non-overlapped cohorts but remained in the medium group in overlapped cohorts. On the other hand, ESRDDXSIMP remained in the medium group in overlapped cohorts but in non-overlapped cohorts, it joined the high importance group for 2nd cohort and then again

showed a downward trend for 3rd cohort but still remained a highly important feature. FUNCTSTAT showed an erratic behavior. The importance of the feature increased to 15.6% in 2nd cohort of overlapped cohorts, whereas it decreased to 1% in 2nd cohort of non-overlapped cohorts. Though most features have shown a same trend in both overlapped and non-overlapped cohorts, we did not have any obvious interpretation for this change.

There were several features that have low importance throughout the cohorts. However, few features were insignificant in the 1st cohort, but came in lime light in the later cohorts. Amongst them were Donor's Age (DAGE), Recipient's Age (RAGETX) and recipient's diabetes status (rdm2). They became one of the most important features in 2nd and 3rd cohorts of both overlapped and non-overlapped. Based on the scores of non-overlapped cohorts, we tried to summarize the changing importance of the features (especially those features which fall in high and medium group) in the Table 20 below.

Table 20 Categorical importance of features in three time-cohorts

Time-Cohorts	High	Medium	Low
1st	esrddxsimp, hlamm, vintage	drcmv, drrace, functstat	remaining
2nd	esrddxsimp,hlamm,vintage	dage,ragetx	remaining
3rd	esrddxsimp,functstat,ragetx,rdm2	dage,ragetx	remaining

We already stated the features that moved from low importance group to medium or high importance group, but, there were several features that did not show any significant change in importance scores which would lead to shifting them from one group to another. Features such as Recipient's Coronary Disease (RCAD), Recipient's Hypertension (RHTN), Status of Dialysis before Transplant (PREEMPTIVE) and Donor's Donation after Cardiac Death (DCD) were relatively more important in 1st cohort than 2nd and 3rd cohort, whereas Expanded Criteria of Donor (ecd) was more important in 3rd cohort. Though, these features have differing importance in time-cohorts, they were all part of low importance group.

Several features (such as DHCV, PREVKI, RCVD, REC_TX_PROCEDURE_TY, RMALIG etc.) had negligible importance throughout the period of the study. Also, whilst performing RFECV these features were usually removed from the feature set. Interestingly both overlapped and non-overlapped cohorts provided the scores for them alike.

The analysis that we did above was mainly based on complete features. The categorical features were based on multiple levels, hence we performed the same analysis that we performed above with the dummy variables which were developed by transforming the categorical variables using one-hot encoding module of scikit-learn library. The figures below show the changing relevance of the categorical variables in overlapped and non-overlapped cohorts. Since the number of dummy variables were huge (98), it was difficult to visualize all of them in a single figure whilst maintaining the visual clarity. Therefore, we filtered out all those variables which have the importance of less than 1% from all the three time-cohorts.

The Figure 11 and Figure 12 below shows the feature importance scores in overlapped and non-overlapped cohorts.

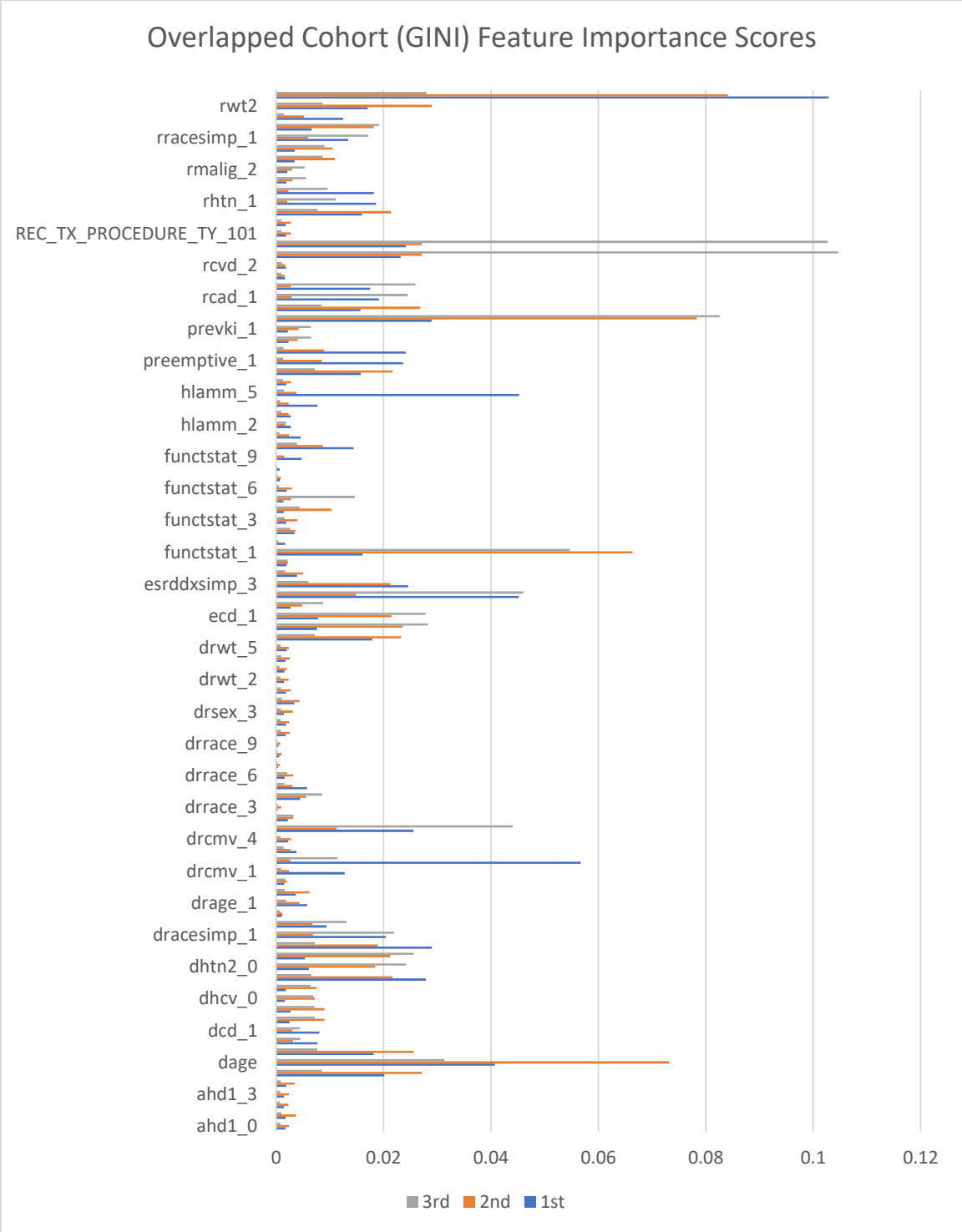


Figure 11 Dummy feature importance scores based on overlapped cohorts

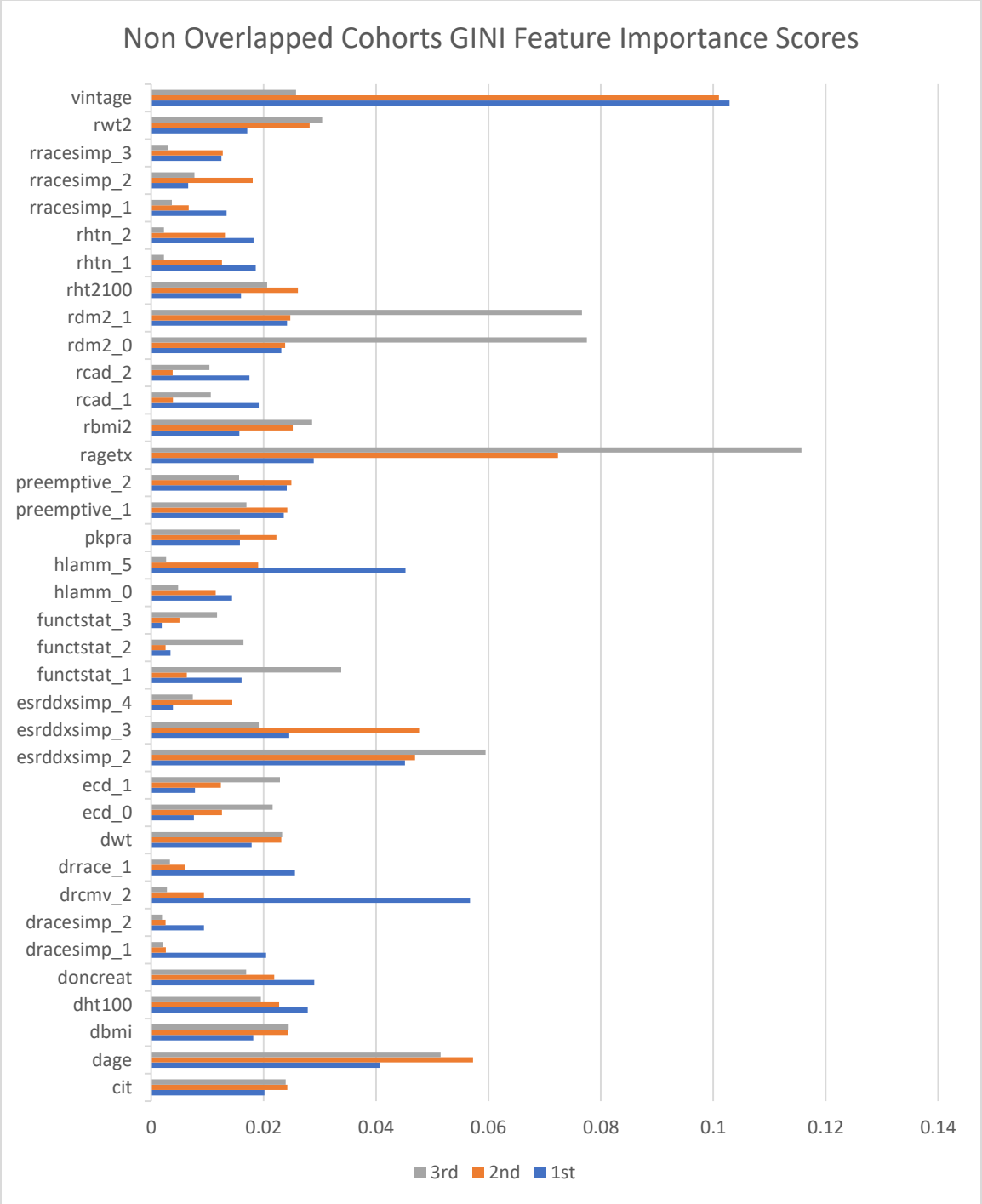


Figure 12 Dummy feature importance scores based on non-overlapped cohorts

The important continuous features (such as time on dialysis before transplant (vintage), Donor' Age (DAGE), Recipient's Age (RAGETX) etc.) showed the same trend in comparison to the analysis done above, but this analysis unsurfaced few more continuous

features which had very low importance in the former analysis. These features include Cold Ischemia Timing (CIT), Donor's BMI (DBMI), Recipient's BMI (RBMI) and Donor's Creatinine Level (DONCREAT). As per overlapped cohorts CIT was more important in 1st and 2nd cohort in comparison to the 3rd cohort, however, non-overlapped cohorts did not show any significant distinction in the importance of the variables. DONCREAT showed a decreasing importance over the period of study in both overlapped and non-overlapped cohorts, whereas RBMI and DBMI showed an increasing importance over the time. There was one more feature, namely Peak Panel Reactive Antibody (PKPRA), which showed an increasing importance in the beginning but stooped down in the 3rd cohort. This feature was relatively more important to predict the graft status in the medium term (2nd cohort) of both non-overlapped and overlapped cohorts.

The dummy categorical features provided many interesting insights in this analysis. We already knew which features were important in the different time-cohorts, however, it was still not known that which value of the feature (dummy variable) is responsible for that prediction. This analysis helped us in overcoming that problem. The previous analysis on whole categorical features revealed us that HLAMM, ESRDDXSIMP etc. were important features in the 1st cohort. This analysis further explained that HLA with 5 mismatches (HLAMM_5) is the actual value that determine the status of the graft in 1st cohort. Similarly, ESRDDXSIMP was based on 5 different values. The distinguishing values for this feature are ESRDDXSIMP_2, ESRDDXSIMP_3 and ESRDDXSIMP_4. ESRDDXSIMP_2 and ESRDDXSIMP_3 possessed relatively more importance than ESRDDXSIMP_4. These dummy values were important in all three cohorts, however, ESRDDXSIMP_2 was more important in 3rd cohort whereas, ESRDDXSIMP_3 and ESRDDXSIMP_4 were more important in 2nd cohort. Amongst the paired variables, DRCMV_2 and DRRACE_1 were important in 1st cohort, whereas their importance was not notable in other cohorts. These two variables were also mentioned in Table 20 where they were included in medium importance group as integral categorical variables. The Functional Status of Recipient (FUNCTSTAT) was more important in long term (3rd cohort) with FUNCTSTAT values FUNCTSTAT _1, FUNCTSTAT _2, FUNCTSTAT _3, and FUNCTSTAT _5. Few features (such as ECD, RHTN, PREEMPTIVE, RCAD, RPVD

etc.) showed changing importance during the time-cohorts, but their dummy values had the same trend thus we did not specifically include them in this analysis.

All the feature importance scores that we have discussed above were calculated on the basis of mean decreased impurity (Gini). The main problem with these scores was the bias towards high cardinality features. Interestingly most of our top scoring features, whether based on Gini index or RFECV, were continuous features. In order to make sure whether these features were actually important or were affected by the bias, we calculated the permutation based feature importance scores for both overlapped and non-overlapped cohorts. Our focus was particularly on the importance of the continuous features to analyze any potential difference between the new and the old feature importance scores. The Figure 13 and Figure 14 below shows the importance of the features which made a difference after applying permutation technique. Most of the features that were shown as important by Gini impurity (See Figure 11 and Figure 12) above were also endorsed as important by permutation technique. RAGETX, DAGE, VINTAGE were continuous features that were stated as highly important by Gini method, but since these features have an equal significance by permutation method, we can deduce that the biasness of Gini method has not affected these features. Apart from this the dummy features (such as HLAMM_5, ESRDDXSIMP_2 etc.) also remained same in this analysis, however, few features (such as DONCREAT, CIT, DBMI etc.) were important in the previous analysis in both overlapped and non-overlapped cohorts, but they had very little permutation based score for non-overlapped cohorts. We only stated these features in overlapped cohorts, as they had shown some degree of importance in comparison to non-overlapped cohorts.

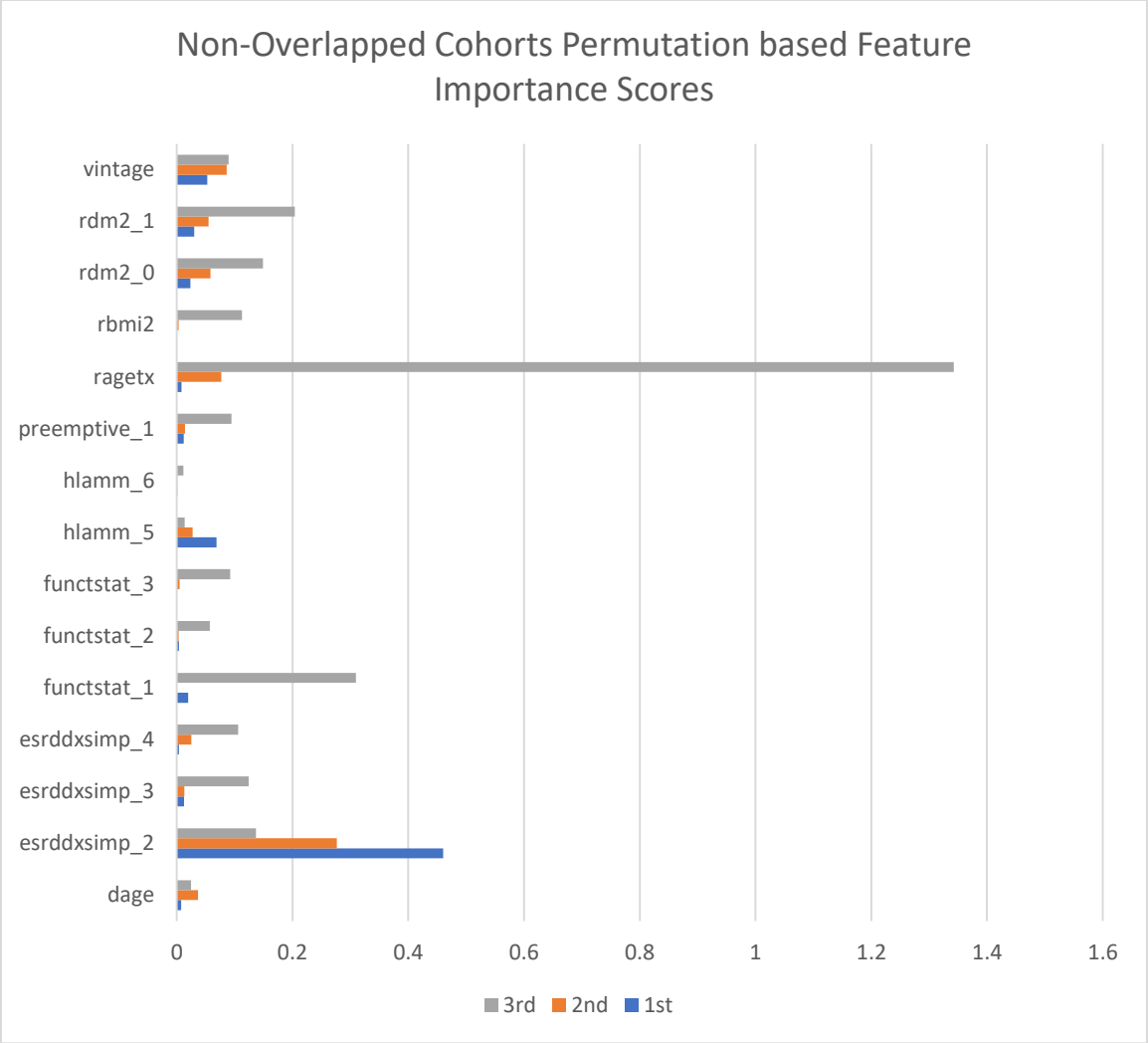


Figure 13 Permutation feature scores based on non-overlapped cohorts

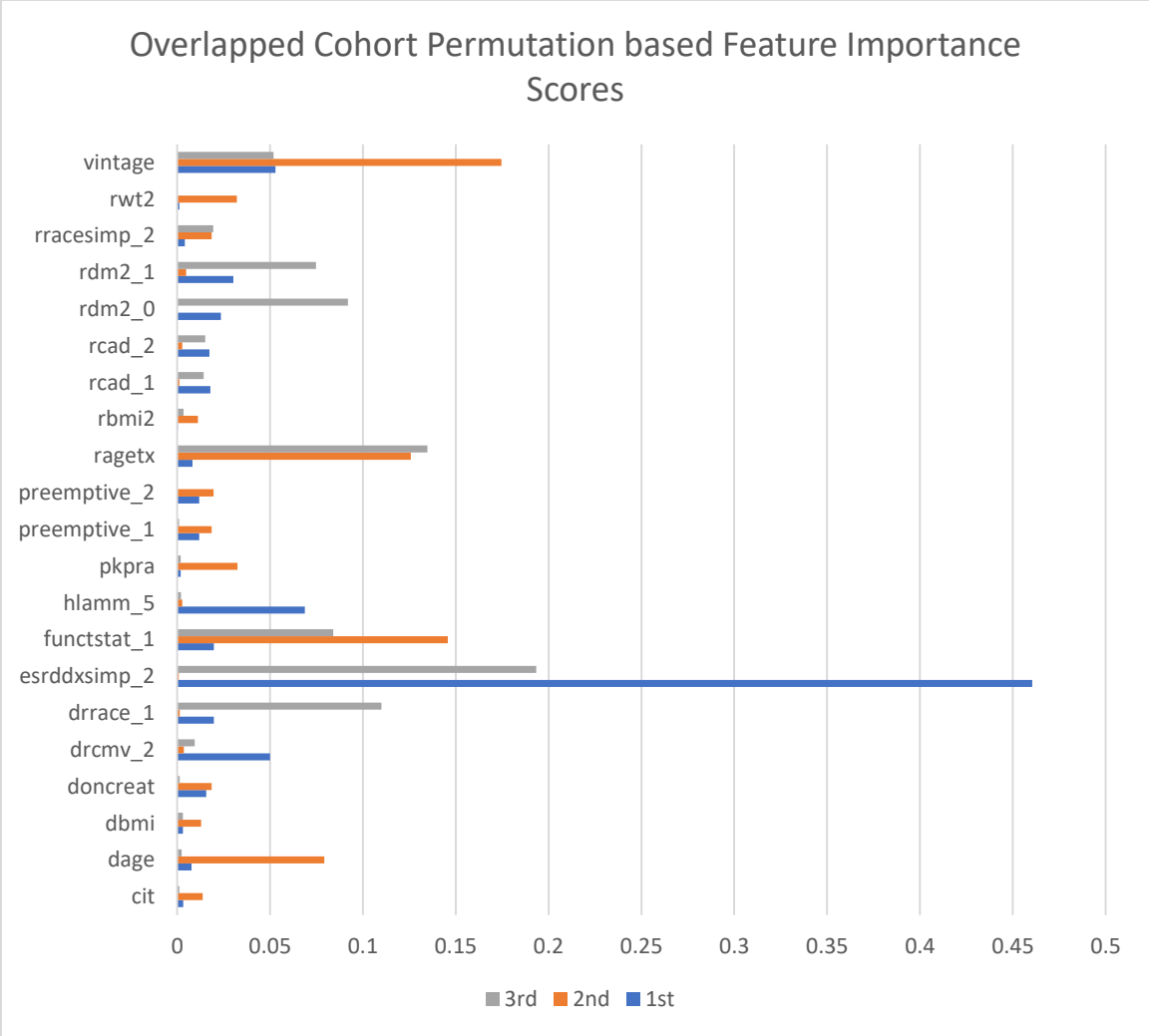


Figure 14 Permutation feature scores for overlapped cohorts

4.3. Discussion

In this research, we predicted the status of the graft and the risk of graft survival using binary and multiclass classification approaches. The binary approaches were dedicated to the prediction of graft status, whereas, the multiclass approach was employed in the prediction of survival risk. Simultaneously, the binary approaches (particularly the non-overlapped approach) were also utilized in the analysis of changing significance of features over the period of three time intervals.

In the discussion that follows, we first analyzed our results for the status of the graft. As shown in the results above, overlapped cohorts were used to make prediction models to predict the graft status in short, medium and long term. Amongst all the models developed

for short term, SVM turned out to be the best classifier providing an AUROC score of 85%. We compared these results with three different studies on organ transplants which involves similar datasets with nearly same amount of data observations. The study by Ray S. Lin et al. [18] applied ANN and LR on their combined dataset from UNOS and United States Renal Data System (USRDS) for predicting the status of kidney grafts. The LR returned 71%, whereas, the ANN returned 73% AUROC for the prediction models developed for 1st year. The study by A.Dag et al. [30] predicted the outcome of heart transplants on a similar dataset by UNOS consisting over 30000 data observations. Their findings for the 1st year were considerably low. The best classifier for them was LR providing 63% followed by SVM providing 62% AUROC score. There were several other studies which predicted short term graft status of different organ transplants (such as [37][33][45][62][46][81]), but due to their small size of datasets we did not make any comparison with them.

Results for all the models which were trained in 2nd cohort did not show any marginal difference on AUROC metric, however, the f1-score for the failed grafts was highest for LR, therefore, LR was the winning model for 2nd cohort. The AUROC score was around 66% on training dataset and 60% on testing dataset. The study by Tiong, H. Y. et al. [82] analyzed 20085 living donor transplant cases from UNOS for 5-year graft survival using nomograms and predicted the concordance index of 0.71. The predictors included in the study were pre-transplant features as well as a post-transplant feature delayed graft function to develop the nomograms. The study by Ray S. Lin et al. [18], - which is also referred for comparing the results of short term graft status above – also predicted the graft status for medium term 5-years. They provided a score of 77% for artificial neural network. Though their score was much higher than our results but their dataset was based 10641 survivals and 7215 failures, whereas, we analyzed 23475 failures and 29352 survivals (See Table 2). The study by A. Dag et al. [30] provided an analysis on 5- years heart graft status using UNOS dataset. They applied SVM and LR on their dataset after oversampling and received 67% from both classifiers. The findings were closest to our results although the organ under analysis was different.

It was surprising to see that LR became the winning classifier among all the other robust classifiers used in this cohort. The only factor that could have resulted in this anomaly is the nature of data observations used in this particular cohort. In both 1st and 3rd cohort we made use of oversampling technique to generate synthetic samples however, in 2nd cohort the data was completely original with no use of oversampling. Hence it can be deduced that the other classifiers were able to make better prediction models in presence of oversampled data but LR came out to be the best algorithm in absence of oversampled data.

The results for the 3rd cohort were better than 2nd cohort. Unlike 2nd cohort, the classifiers did not return same AUROC scores. AdaBoost and ANNs performed the best on the dataset providing 84% and 81% AUROC scores on the cross validated dataset. The testing dataset also showed AdaBoost and ANNs as the best classifiers with 78% and 73% AUROC scores. The study by Ray S. Lin et al. [18] predicted 82% AUROC score for 7-years graft survival using ANNs whereas a similar study by M. Luck et al.[29], based on nearly same number of transplants predicted the c-index between the range of 0.63-0.66 for 14-years graft survival. Our findings were relatively better than the results of both studies.

The risk of graft failure was approached as a multiclass ML problem. We calculated the F1-micro scores for predicting the risk groups (i.e. high, medium and low) with our five classifiers. The SVM and AdaBoost provided us with 62% and 61% on cross validated dataset, respectively, whereas, the scores were slightly higher for the testing dataset with 63% and 62% for SVM and AdaBoost, respectively. A.Kazim et al. [14] also performed a multiclass classification on the same dataset that we have analyzed in this research, however, the number of observations were slightly lesser than ours because they considered the transplants after June, 2004 whereas the our dataset was based on transplants between years 2000 to 2017. The F1-score for the 5-fold cross validation that they received was 60.2% which is lower than our findings of 62%. Another study by Jiakai et al. [45] also performed the prediction of the period of the graft survival based on a multiclass approach. Although, their dataset was very small, the prediction results were between the range of 30% to 68% with a variety of different Bayesian models. The best model which provided 68% prediction score was Hill Climber (-P 3 –N –S BAYES).

We will further extend our discussion by comparing the changing significance of the features with respect to the findings of the similar studies. We will first look into the factors which were relatively more important for the short term (1st cohort). As stated in section Analysis of changing effects of features above, the most important features for the 1st cohort were VINTAGE, HLAMM, DRRACE, and DRCMV. HLA mismatches were an important predictor for the first cohort but not for the other two cohorts and this was close to the findings of Goldfarb-Rumyantzev et al. [83], who found HLA mismatches to be an important factor in first three years of the transplant. VINTAGE showed a diminishing significance with a very high significance in the first cohort. In the third cohort, VINTAGE was placed in the least important group. These findings were confirmed in other studies [33][84]. The paired variables such as Donor-Recipient Race and CMV (DRRACE and DRCMV) also showed its distinctiveness in short term. They had more predictive strength for the 1st cohort than the other two cohorts. Since the paired variables were not analyzed in the studies before, the research literature did not provide us its significance in the short, medium or long term. The recipient race was however individually recognized as an important variable for short term in the following study by T. Brown et al. [81]. Donation after cardiac death (DCD) and extending criteria of donor (ECD) were among the least important predictors; however, they had a relatively higher predictive power for the first cohort and third cohorts, respectively. DCD is linked to higher chances of Delayed Graft Function (DGF), which itself is an important predictor for short term failure, therefore the results for DCD are not surprising [85][86]. Since DGF is a post-transplant variable, we have not taken that variable into consideration.

The most deterministic features for long-term (3rd cohort) were recipient's diabetes status and recipient's age. The findings were consistent with the results of some other studies [18][87].

To conclude the discussion above was fundamentally based on the results of the prediction models pertaining to different cohorts and the changing significance of the features. We have not considered the aspect of explaining and interpreting our machine learning models in this research work. Our explanation was only restricted to finding the most significant features in our prediction models using Gini impurity and permutation feature importance technique.

5. Conclusion

5.1. Summary

It is important to identify potential kidney failure as early as possible to help increase positive outcomes for the patient and to help reduce costs associated with end stage renal disease. In this research, we predicted the outcome of kidney transplants using the UNOS dataset based on 52827 kidney transplant cases from year 2000- 2017. The outcome of the transplant was considered as the status of the graft and risk of graft failure. We measured the status of the graft in three different time-periods as a binary class problem, whereas, the risk of the graft failure was approached as a multiclass problem. The three time-cohorts were formed to predict the status of the graft using overlapped and non-overlapped cohorts. The non-overlapped cohorts were based on cases which experienced an adverse event during 0-1 years, between 2-5 years and more than 5 years following the transplant, whereas the overlapped cohorts were based on cases which experienced an adverse event during 0-1 year, 0-5 years, and 0-17 years following a transplant. The risk of graft failure was calculated by categorizing the non-overlapping time cohorts into three categories representing the risk of failure i.e. high, medium and low. The 1st time-cohort represented a high risk of failure whereas the later time-cohorts represented likewise. We experimented with 5 classification algorithms (i.e. random forest, adaptive boosting, artificial neural network, logistic regression and support vector machine). In addition to developing the prediction models, we also analyzed the changes in the significance of the features over the period of the study. Our results indicate that support vector machine and adaptive boosting combined with SMOTE provided the best area-under-the-receiver-operating-characteristic-curve (AUROC). The cross-validated AUROC scores for predicting the graft status were 85%, 66%, and 84% in 1st and 2nd and 3rd cohort, respectively, whereas the F1-Micro score for the risk of graft failure was 62%. The feature importance scores were calculated using Gini impurity and permutation based techniques to identify the important predictors and analyze their changing contribution in predicting the results for the three different time-cohorts; we noted a change in the significance of attributes across the three different time cohorts (e.g. the number of years on dialysis before transplant was an important attribute in only 1st and 2nd time-cohorts, whereas, the recipient's age and recipient's diabetes status were important in only 3rd cohort).

5.2. Limitations

One of the main limitations of this research was the computational constraint of applying ML algorithms. Since the dataset was extremely large, the adequate computational resources needed to train the models became a hindrance. Except the ANN and LR classifiers, all the algorithms took a considerably time to train. The bagging classifier of scikit-learn library was implemented numerous times during the model training process but despite waiting for considerably long time, we were not able to get the completely trained models entirely due to the lack of computational power. The same issue was present while applying survival SVM during the development of baseline for our work. The training crashed several times due to the unavailability of the Ram. We skipped survival SVM and only relied on RSF method for generating the baseline scores.

In addition to the computational limitation, a functional limitation of the research was realized whilst performing the feature selection with recursive feature elimination. Since the scikit-learn library does not allow categorical variables without transforming them into dummy variables, we were not able to calculate how many complete features were supposed to be eliminated. To measure the number of features, which were eliminated after performing RFECV, we manually checked the dummy features belonging to a certain categorical feature. If all dummy features of that categorical features were removed after the process, we considered the whole categorical feature as eliminated.

Another limitation of our research lies in the assumption that all those transplants where graft status was missing and the patient had not died during the study were actually considered survived. In order to deal with this limitation, we removed all those survived cases which were censored before 8-years following their transplant. A better analysis could have been done if the status of the transplant survival would have been provided by the dataset providers rather than the assuming it by heuristics.

5.3. Future Work

The results presented in this thesis open a number of possibilities for future research. We developed the prediction models and analyzed the changing significance of features using a combination of categorical and continuous features. Most of the continuous features were found to be important throughout our experiments. There is a potential to further extend

the set of predictors by categorizing the continuous features into different categorical features. We believe the domain experts can assess this information and their advice can help us in efficient binning of these continuous features. The recipient's age which is one the important features in 3rd cohort is a continuous feature. By categorizing this feature into different bins, we can identify which age group of recipients is more significant in the 3rd cohort. We can implement this on all the continuous features to get further insights.

The multiclass experiments that we performed to predict the risk of graft failure were rather exploratory in nature. We implemented the same five classifiers that were used in binary approaches without exploring separately for the multiclass problem. Although, our results were relatively better than the contemporary studies, there is an immense potential to perform future research in this area.

In addition to improving the accuracy of prediction models, the research work needs to be converted into a useful tool for the nephrology community. The iBox [88] tool is an interesting example which uses the patient's follow-up reports to predict the probabilities of graft loss up to 10 years after patient evaluation. Our tool can be developed in the similar manner but it will provide the classification of the patient survival and failure in the three different time-periods.

References

[1] H. Sasaki, “Kidney transplantation,” *Japanese J. Clin. Urol.*, vol. 66, no. 10, pp. 753–758, 2012.

[2] Yoshio N Hall; Glenn M Chertow, “Kidney disorders End stage renal disease Search date April 2007 Kidney disorders End stage renal disease,” no. April, pp. 1–15, 2007.

[3] C. E. Mcculloch, D. Ph, and C. Hsu, “Chronic Kidney Disease and the Risks of Death, Cardiovascular Events, and Hospitalization,” pp. 1296–1305, 2004.

[4] P. Rashidi Khazaee, J. Bagherzadeh, Z. Niazkhani, and H. Pirnejad, “A dynamic model for predicting graft function in kidney recipients’ upcoming follow up visits: A clinical application of artificial neural network,” *Int. J. Med. Inform.*, vol. 119, no. August, pp. 125–133, 2018, doi: 10.1016/j.ijmedinf.2018.09.012.

[5] “About Chronic Kidney Disease | National Kidney Foundation.” [Online]. Available: https://www.kidney.org/atoz/content/about-chronic-kidney-disease. [Accessed: 15-Nov-2019].

[6] J. D. Schold and D. L. Segev, “Increasing the pool of deceased donor organs for kidney transplantation,” *Nat. Rev. Nephrol.*, vol. 8, no. 6, pp. 325–331, 2012, doi: 10.1038/nrneph.2012.60.

[7] “Conservative Care - The Kidney Foundation of Canada | La Fondation canadienne du rein.” [Online]. Available: https://www.kidney.ca/conservative-care. [Accessed: 15-Nov-2019].

[8] J. Perl, “Kidney Transplant Failure: Failing Kidneys, Failing Care?,” *Clin. J. Am. Soc. Nephrol.*, vol. 9, no. 7, pp. 1153–1155, Jul. 2014, doi: 10.2215/CJN.04670514.

[9] N. Tangri *et al.*, “A predictive model for progression of chronic kidney disease to kidney failure,” *JAMA - J. Am. Med. Assoc.*, vol. 305, no. 15, pp. 1553–1559, 2011, doi: 10.1001/jama.2011.451.

[10] T. G. Clark, M. J. Bradburn, S. B. Love, and D. G. Altman, “Survival Analysis Part I: Basic concepts and first analyses,” *Br. J. Cancer*, vol. 89, no. 2, pp. 232–238, Jul. 2003, doi: 10.1038/sj.bjc.6601118.

[11] D. M. Vock *et al.*, “Adapting machine learning techniques to censored time-to-event health record data: A general-purpose approach using inverse probability of censoring weighting,” *J. Biomed. Inform.*, vol. 61, pp. 119–131, 2016, doi: 10.1016/j.jbi.2016.03.009.

[12] K.-M. Leung, R. M. Elashoff, and A. A. Afifi, “Censoring Issues in Survival Analysis,” *Annu. Rev. Public Health*, vol. 18, no. 1, pp. 83–104, 1997, doi: 10.1146/annurev.publhealth.18.1.83.

[13] J. Yoon, W. R. Zame, A. Banerjee, M. Cadeiras, A. M. Alaa, and M. van der Schaar, “Personalized survival predictions via Trees of Predictors: An application to cardiac transplantation,” *PLoS One*, vol. 13, no. 3, pp. 1–19, 2018, doi: 10.1371/journal.pone.0194985.

[14] K. Topuz, F. D. Zengul, A. Dag, A. Almehmi, and M. B. Yildirim, “Predicting graft survival among kidney transplant recipients: A Bayesian decision support model,” *Decis. Support Syst.*, vol. 106, pp. 97–109, 2018, doi: 10.1016/j.dss.2017.12.004.

[15] M. Kleinbaum, David G., Klein, *Survival Analysis: A Self-Learning Text*, Third Edit. New York: Springer-Verlag, 2012.

[16] J. Kishore, M. Goel, and P. Khanna, “Understanding survival analysis: Kaplan-Meier estimate,” *Int. J. Ayurveda Res.*, vol. 1, no. 4, p. 274, 2010, doi: 10.4103/0974-7788.76794.

[17] N. Mantel, “Evaluation of survival data and two new rank order statistics arising in its consideration.,” *Cancer Chemother. reports*, vol. 50, no. 3, pp. 163–70, Mar. 1966.

[18] R. S. Lin, S. D. Horn, J. F. Hurdle, and A. S. Goldfarb-Rumyantzev, “Single and multiple time-point prediction models in kidney transplant outcomes,” *J. Biomed. Inform.*, vol. 41, no. 6, pp. 944–952, 2008, doi: 10.1016/j.jbi.2008.03.005.

[19] A. Akl, A. M. Ismail, and M. Ghoneim, “Prediction of Graft Survival of Living-Donor Kidney Transplantation: Nomograms or Artificial Neural Networks?,” *Transplantation*, vol. 86, no. 10, pp. 1401–1406, 2008, doi: 10.1097/TP.0b013e31818b221f.

[20] L. Ohno-Machado, “Modeling medical prognosis: Survival analysis techniques,” *J. Biomed. Inform.*, vol. 34, no. 6, pp. 428–439, 2001, doi: 10.1006/jbin.2002.1038.

[21] V. Van Belle, K. Pelckmans, S. Van Huffel, and J. A. K. Suykens, “Support vector methods for survival analysis: A comparison between ranking and regression approaches,” *Artif. Intell. Med.*, vol. 53, no. 2, pp. 107–118, 2011, doi: 10.1016/j.artmed.2011.06.006.

[22] H. Wang and L. Zhou, “Random survival forest with space extensions for censored data,” *Artif. Intell. Med.*, vol. 79, pp. 52–61, 2017, doi: 10.1016/j.artmed.2017.06.005.

[23] C. N. Yu, R. Greiner, H. C. Lin, and V. Baracos, “Learning patient-specific cancer survival distributions as a sequence of dependent regressors,” *Adv. Neural Inf. Process. Syst. 24 25th Annu. Conf. Neural Inf. Process. Syst. 2011, NIPS 2011*, pp. 1–9, 2011.

[24] S. Pölsterl, S. Conjeti, N. Navab, and A. Katouzian, “Survival analysis for high-dimensional, heterogeneous medical data: Exploring feature extraction as an alternative to feature selection,” *Artif. Intell. Med.*, vol. 72, pp. 1–11, 2016, doi: 10.1016/j.artmed.2016.07.004.

[25] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, “Random survival forests,” *Ann. Appl. Stat.*, vol. 2, no. 3, pp. 841–860, 2008, doi: 10.1214/08-AOAS169.

[26] S. Goli, H. Mahjub, J. Faradmal, H. Mashayekhi, and A.-R. Soltanian, “Survival Prediction and Feature Selection in Patients with Breast Cancer Using Support Vector Regression,” *Comput. Math. Methods Med.*, vol. 2016, pp. 1–12, 2016, doi: 10.1155/2016/2157984.

[27] D. Faraggi and R. Simon, “A neural network model for survival data,” *Stat. Med.*, vol. 14, no. 1, pp. 73–82, Jan. 1995, doi: 10.1002/sim.4780140108.

[28] K. Matsuo *et al.*, “Survival outcome prediction in cervical cancer: Cox models vs deep-learning model,” *Am. J. Obstet. Gynecol.*, vol. 220, no. 4, pp. 381.e1-381.e14, 2019, doi: 10.1016/j.ajog.2018.12.030.

[29] M. Luck, T. Sylvain, H. Cardinal, A. Lodi, and Y. Bengio, “Deep Learning for Patient-Specific Kidney Graft Survival Analysis,” no. Nips 2017, 2017.

[30] A. Dag, A. Oztekin, A. Yucel, S. Bulur, and F. M. Megahed, “Predicting heart transplantation outcomes through data analytics,” *Decis. Support Syst.*, vol. 94, pp. 42–52, 2017, doi: 10.1016/j.dss.2016.10.005.

[31] R. Greco, T. Papalia, D. Lofaro, S. Maestripieri, D. Mancuso, and R. Bonofiglio, “Decisional Trees in Renal Transplant Follow-up,” *Transplant. Proc.*, vol. 42, no. 4, pp. 1134–1136, 2010, doi: 10.1016/j.transproceed.2010.03.061.

[32] D. Lofaro *et al.*, “Prediction of Chronic Allograft Nephropathy Using Classification Trees,” *Transplant. Proc.*, vol. 42, no. 4, pp. 1130–1133, 2010, doi: 10.1016/j.transproceed.2010.03.062.

[33] L. Shahmoradi, M. Langarizadeh, G. Pourmand, Z. A. Fard, and A. Borhani, “Comparing Three Data Mining Methods to Predict Kidney Transplant Survival,” *Acta Inform. Medica*, vol. 24, no. 5, pp. 322–327, 2016, doi: 10.5455/aim.2016.24.322-327.

[34] P. Piros *et al.*, “Comparing machine learning and regression models for mortality prediction based on the Hungarian Myocardial Infarction Registry,” *Knowledge-Based Syst.*, vol. 179, pp. 1–7, 2019, doi: 10.1016/j.knosys.2019.04.027.

[35] T. Shaikhina, D. Lowe, S. Daga, D. Briggs, R. Higgins, and N. Khovanova, "Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation," *Biomed. Signal Process. Control*, vol. 52, pp. 456–462, 2019, doi: 10.1016/j.bspc.2017.01.012.

[36] C. G. Raji and S. S. Vinod Chandra, "Long-Term Forecasting the Survival in Liver Transplantation Using Multilayer Perceptron Networks," *IEEE Trans. Syst. Man, Cybern. Syst.*, vol. 47, no. 8, pp. 2318–2329, 2017, doi: 10.1109/TSMC.2017.2661996.

[37] L. Tapak, O. Hamidi, P. Amini, and J. Poorolajal, "Prediction of kidney graft rejection using artificial neural network," *Healthc. Inform. Res.*, vol. 23, no. 4, pp. 277–284, 2017, doi: 10.4258/hir.2017.23.4.277.

[38] U. R. Acharya *et al.*, "Deep convolutional neural network for the automated diagnosis of congestive heart failure using ECG signals," pp. 16–27, 2019.

[39] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," 2015, doi: 10.1038/nature14539.

[40] D. Medved, P. Nugues, and J. Nilsson, "Predicting the outcome for patients in a heart transplantation queue using deep learning," *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, pp. 74–77, 2017, doi: 10.1109/EMBC.2017.8036766.

[41] C. Lee, W. R. Zame, J. Yoon, and M. Van Der Schaar, "DeepHit: A deep learning approach to survival analysis with competing risks," *32nd AAAI Conf. Artif. Intell. AAAI 2018*, pp. 2314–2321, 2018.

[42] R. Blagus and L. Lusa, "SMOTE for high-dimensional class-imbalanced data," *BMC Bioinformatics*, vol. 14, 2013, doi: 10.1186/1471-2105-14-106.

[43] L. Al-ebbini, "Recipients : A Hybrid Genetic Algorithms-based Methodology," 2017.

[44] G. Weiss, K. McCarthy, and B. Zabar, "Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs?," *Dmin*, pp. 1–7, 2007.

[45] J. Li, G. Serpen, S. Selman, M. Franchetti, M. Riesen, and C. Schneider, “Bayes net classifiers for prediction of renal graft status and survival period,” *World Acad. Sci. Eng. Technol.*, vol. 63, pp. 144–150, 2010.

[46] A. Agrawal, R. Al-Bahrani, M. J. Russo, J. Raman, and A. Choudhary, “Lung transplant outcome prediction using UNOS data,” *Proc. - 2013 IEEE Int. Conf. Big Data, Big Data 2013*, pp. 1–8, 2013, doi: 10.1109/BigData.2013.6691751.

[47] A. Saltelli, M. Ratto, S. Tarantola, and F. Campolongo, “Sensitivity analysis practices: Strategies for model-based inference,” *Reliab. Eng. Syst. Saf.*, vol. 91, no. 10–11, pp. 1109–1125, Oct. 2006, doi: 10.1016/j.ress.2005.11.014.

[48] *Analog Design and Simulation Using OrCAD Capture and PSpice*. Elsevier, 2018.

[49] *Advances in System Reliability Engineering*. Elsevier, 2019.

[50] *LEED v4 Practices, Certification, and Accreditation Handbook*. Elsevier, 2016.

[51] D. Delen, R. Sharda, and P. Kumar, “Movie forecast Guru: A Web-based DSS for Hollywood managers,” *Decis. Support Syst.*, vol. 43, no. 4, pp. 1151–1170, 2007, doi: 10.1016/j.dss.2005.07.005.

[52] H.-C. Lee *et al.*, “Derivation and Validation of Machine Learning Approaches to Predict Acute Kidney Injury after Cardiac Surgery,” *J. Clin. Med.*, vol. 7, no. 10, p. 322, 2018, doi: 10.3390/jcm7100322.

[53] N. S. Escanilla, L. Hellerstein, R. Kleiman, Z. Kuang, J. Shull, and D. Page, “Recursive Feature Elimination by Sensitivity Testing,” in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2018, pp. 40–47, doi: 10.1109/ICMLA.2018.00014.

[54] M. S. Lee, T. Q. Huang, J. S. Seo, and W. Y. Park, “Prediction of the Exposure to 1763MHz Radiofrequency Radiation Based on Gene Expression Patterns,” *Genomics Inform.*, vol. 5, no. 3, pp. 102–106, 2007.

[55] E. B. Huerta, R. M. Caporal, M. A. Arjona, and J. C. H. Hernández, “Recursive Feature Elimination Based on Linear Discriminant Analysis for Molecular Selection and Classification of Diseases,” 2013, pp. 244–251.

[56] A. Decruyenaere, P. Decruyenaere, P. Peeters, F. Vermassen, T. Dhaene, and I. Couckuyt, “Prediction of delayed graft function after kidney transplantation: Comparison between logistic regression and machine learning methods Standards, technology, and modeling,” *BMC Med. Inform. Decis. Mak.*, vol. 15, no. 1, pp. 1–10, 2015, doi: 10.1186/s12911-015-0206-y.

[57] I. Guyon, “Gene Selection for Cancer Classification,” pp. 389–422, 2002.

[58] D. Medved, P. Nugues, and J. Nilsson, “Selection of an optimal feature set to predict heart transplantation outcomes,” *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, vol. 2016-Octob, pp. 3290–3293, 2016, doi: 10.1109/EMBC.2016.7591431.

[59] L. Breiman, “Random Forests,” *J. Mach. Learn.*, vol. 45, pp. 5– 32, 2001.

[60] A. J. Aljaaf *et al.*, “Early Prediction of Chronic Kidney Disease Using Machine Learning Supported by Predictive Analytics,” *2018 IEEE Congr. Evol. Comput. CEC 2018 - Proc.*, pp. 1–9, 2018, doi: 10.1109/CEC.2018.8477876.

[61] M. Zafar, D. Zhu, X. Li, K. Yang, and P. Levy, “SAFS : A Deep Feature Selection Approach for Precision Medicine.”

[62] J. Lasserre, S. Arnold, M. Vingron, P. Reinke, and C. Hinrichs, “Predicting the outcome of renal transplantation,” *J. Am. Med. Informatics Assoc.*, vol. 19, no. 2, pp. 255–262, 2012, doi: 10.1136/amiajnl-2010-000004.

[63] B. C., *Neural networks for pattern recognition*, 1st editio. New York, NY, USA: Oxford University Press, 1995.

[64] R. G. Brereton and G. R. Lloyd, “Support Vector Machines for classification and regression,” *Analyst*, vol. 135, no. 2, pp. 230–267, 2010, doi: 10.1039/b918972f.

[65] Y. Freund and R. Schapire, “A decision-theoretic generalization of online learning and application to boosting,” *J. Comput. Syst. Sci.*, vol. 55, pp. 119–139, 1997.

[66] J. Thongkam, G. Xu, and Y. Zhang, “6-AdaBoost, random forests, ABRF and.pdf,” pp. 3062–3069, 2008.

[67] R. O. Duda, P. E. Hart, and D. G. Stork, “Pattern Classification (2nd ed .),” *Comput. Complex.*, 1998.

[68] A. P. Bradley, “The use of the area under the ROC curve in the evaluation of machine learning algorithms,” *Pattern Recognit.*, vol. 30, no. 7, pp. 1145–1159, Jul. 1997, doi: 10.1016/S0031-3203(96)00142-2.

[69] T. Fawcett, “An introduction to ROC analysis,” *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006, doi: 10.1016/j.patrec.2005.10.010.

[70] D. M. W. Powers, “Evaluation: From Precision, Recall and F-Measure To Roc, Informedness, Markedness & Correlation,” *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, 2011.

[71] P. ElKafrawy, A. Mausad, and H. Esmail, “Experimental Comparison of Methods for Multi-label Classification in different Application Domains,” *Int. J. Comput. Appl.*, vol. 114, no. 19, pp. 1–9, 2015, doi: 10.5120/20083-1666.

[72] “What is UNOS? | About United Network for Organ Sharing.” [Online]. Available: https://unos.org/about/. [Accessed: 06-Nov-2019].

[73] M. Dorado-Moreno, M. Pérez-Ortiz, P. A. Gutiérrez, R. Ciria, J. Briceño, and C. Hervás-Martínez, “Dynamically weighted evolutionary ordinal neural network for solving an imbalanced liver transplantation problem,” *Artif. Intell. Med.*, vol. 77, pp. 1–11, 2017, doi: 10.1016/j.artmed.2017.02.004.

[74] K. N. Hong *et al.*, “Who is the high-risk recipient? Predicting mortality after heart transplant using pretransplant donor and recipient risk factors,” *Ann. Thorac. Surg.*, vol. 92, no. 2, pp. 520–527, 2011, doi: 10.1016/j.athoracsur.2011.02.086.

[75] S. van der Walt, S. C. Colbert, and G. Varoquaux, “The NumPy Array: A Structure for Efficient Numerical Computation,” *Comput. Sci. Eng.*, vol. 13, no. 2, pp. 22–30, Mar. 2011, doi: 10.1109/MCSE.2011.37.

[76] F. and others Chollet, “Keras.” GitHub, 2015.

[77] T. E. Oliphant, “Python for Scientific Computing,” *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 10–20, 2007, doi: 10.1109/MCSE.2007.58.

[78] T. M. Therneau, “A package for Survival Analysis in S.” 2015.

[79] P. C. Austin and E. W. Steyerberg, “Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable,” *BMC Med. Res. Methodol.*, vol. 12, no. 1, p. 82, Dec. 2012, doi: 10.1186/1471-2288-12-82.

[80] K. D. Yoo *et al.*, “A Machine Learning Approach Using Survival Statistics to Predict Graft Survival in Kidney Transplant Recipients: A Multicenter Cohort Study,” *Sci. Rep.*, vol. 7, no. 1, pp. 1–12, 2017, doi: 10.1038/s41598-017-08008-8.

[81] T. S. Brown *et al.*, “Bayesian modeling of pretransplant variables accurately predicts kidney graft survival,” *Am. J. Nephrol.*, vol. 36, no. 6, pp. 561–569, 2012, doi: 10.1159/000345552.

[82] H. Y. Tiong *et al.*, “Nomograms for Predicting Graft Function and Survival in Living Donor Kidney Transplantation Based on the UNOS Registry,” *J. Urol.*, vol. 181, no. 3, pp. 1248–1255, 2009, doi: 10.1016/j.juro.2008.10.164.

[83] A. S. Goldfarb-Rumyantzev, J. D. Scandling, L. Pappas, R. J. Smout, and S. Horn, “Prediction of 3-yr cadaveric graft survival based on pre-transplant variables in a large national dataset,” *Clin. Transplant.*, vol. 17, no. 6, pp. 485–497, 2003, doi: 10.1046/j.0902-0063.2003.00051.x.

[84] S. Design, “The Ne w E n g l a nd Jo u r n a l o f Me d ic i ne EFFECT OF THE USE OR NONUSE OF LONG-TERM DIALYSIS ON THE SUBSEQUENT SURVIVAL OF RENAL TRANSPLANTS FROM LIVING DONORS E F F EC T O F U S E O R N O N U S E O F L O N G -T E R M D IA LYS IS O N T H E S U B ,” vol. 344, no. 10, pp. 726–731, 2001.

[85] A. C. Wells *et al.*, “Donor kidney disease and transplant outcome for kidneys donated after cardiac death,” *Br. J. Surg.*, vol. 96, no. 3, pp. 299–304, Mar. 2009, doi: 10.1002/bjs.6485.

[86] D. Bahl, Z. Haddad, A. Datoo, and Y. A. Qazi, “Delayed graft function in kidney transplantation,” *Curr. Opin. Organ Transplant.*, vol. 24, no. 1, pp. 82–86, 2019, doi: 10.1097/MOT.0000000000000604.

[87] C. Legendre, G. Canaud, and F. Martinez, “Factors influencing long-term outcome after kidney transplantation,” *Transpl. Int.*, vol. 27, no. 1, pp. 19–27, Jan. 2014, doi: 10.1111/tri.12217.

[88] A. Loupy *et al.*, “Prediction system for risk of allograft loss in patients receiving kidney transplants: International derivation and validation study,” *BMJ*, vol. 366, pp. 1–12, 2019, doi: 10.1136/bmj.l4923.

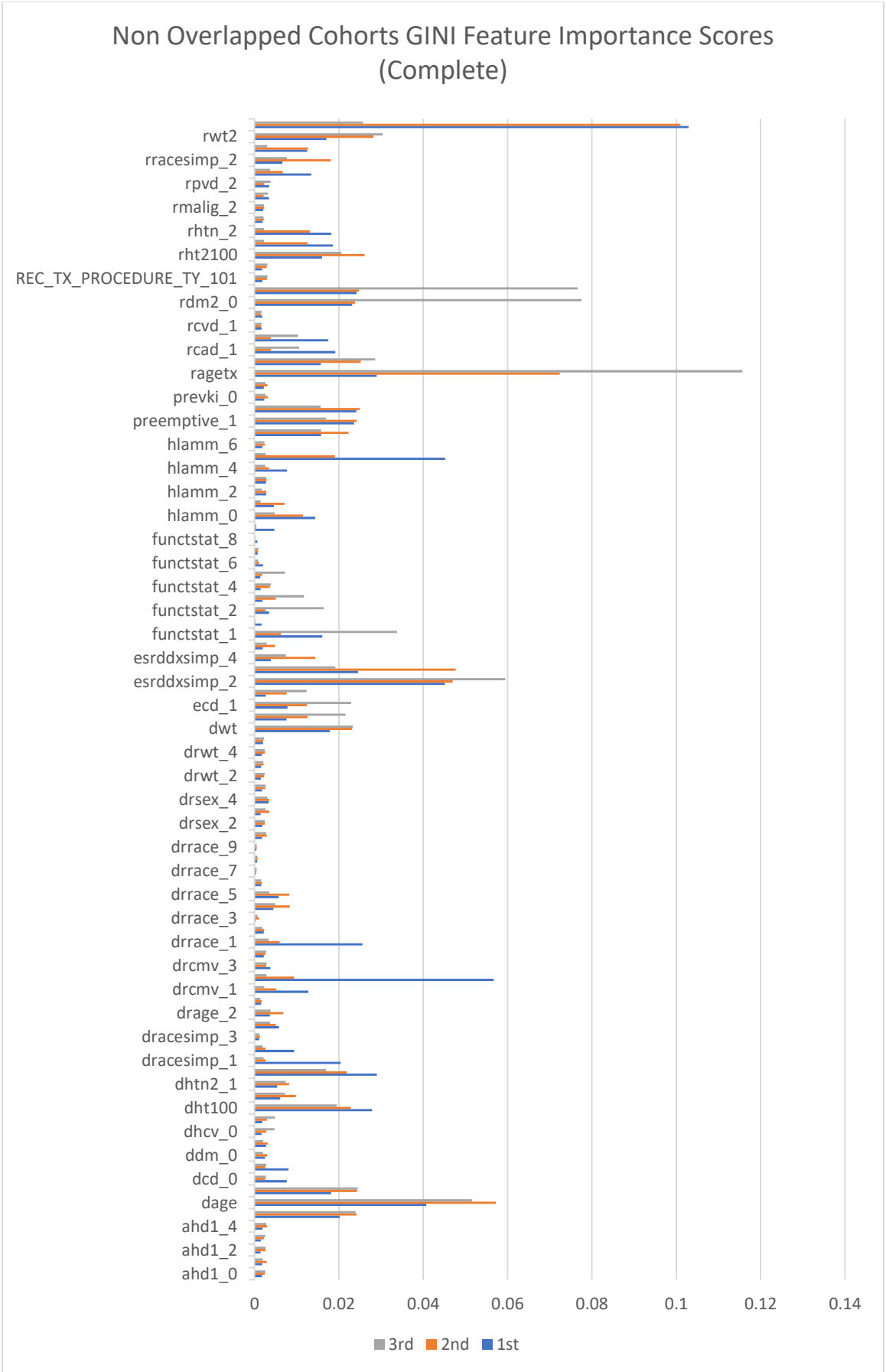
Appendix

Table 21 Description of Dummy Variables

Dummy Variables	Description
ahd1_0	minus 5 to pos 5 (D=R)
ahd1_1	less than minus 15 (D<R)
ahd1_2	minus 15 to minus 5 (D<R)
ahd1_3	post 5 to pos 15 (D>R)
ahd1_4	> pos 15 (D>R)
dbmisimp_0	<18.5
dbmisimp_1	18.5-<25
dbmisimp_2	25-<30
dbmisimp_3	30-<35
dbmisimp_4	>35
dcd_0	No
dcd_1	yes
ddm_0	No
ddm_1	Yes
dhcv_0	No
dhcv_1	Yes
dhtn_0	No
dhtn_1	Yes
dracesimp_1	White
dracesimp_2	Black
dracesimp_3	Other
drage_1	minus 20 to pos 20 (D=R)
drage_2	less than minus 20 (D<R)
drage_3	greater than pos 20 (D>R)
drcmv_1	Donor neg, recipient neg
drcmv_2	donor pos, recipient pos
drcmv_3	donor neg, recipient pos
drcmv_4	donor pos, recipient neg
drrace_1	donor white-recipient white (DWRW)
drrace_2	donor black-recipient black (DBRB)

drrace_3	donor other-recipient other (DORO)
drrace_4	DWRB
drrace_5	DWRO
drrace_6	DBRW
drrace_7	DBRO
drrace_8	DORW
drrace_9	DORB
drsex_1	male donor-male recipient (MDMR)
drsex_2	female donor-female recipient (FDFR)
drsex_3	MDFR
drsex_4	FDMR
drwt_1	minus 10 to pos 10 (D=R)
drwt_2	10 to 30 (D>R)
drwt_3	>30 (D>R)
drwt_4	minus 10 to minus 30 (D<R)
drwt_5	less than minus 30 (D<R)
ecd_0	No
ecd_1	Yes
esrddxsimp_1	GN
esrddxsimp_2	DM
esrddxsimp_3	PCKD
esrddxsimp_4	HTN
esrddxsimp_5	Other
functstat_1	1-100% no complaints
functstat_2	90%-minor sx
functstat_3	80%-some sx
functstat_4	70%-unable to do normal activities
functstat_5	60%-req assistance
functstat_6	40%-disabled
functstat_7	30%-severely disabled
functstat_8	20%-very sick
functstat_9	10%-moribund
preemptive_1	Yes

preemptive_2	No
prevki_0	No
prevki_1	Yes
rcad_1	No
rcad_2	Yes
rcvd_1	No
rcvd_2	Yes
rdm2_0	No
rdm2_1	Yes
REC_TX_PROCEDURE_TY _101	Left Kidney
REC_TX_PROCEDURE_TY _102	Right Kidney
rhtn_1	No
rhtn_2	Yes
rmalig_1	No
rmalig_2	Yes
rpvd_1	No
rpvd_2	Yes
rracesimp_1	White
rracesimp_2	Black
rracesimp_3	Other


Non Overlapped Cohorts GINI Feature Importance Scores (Complete)
rwt2
rracesimp_2
rpvd_2
rmalig_2
rhtn_2
rht2100
REC_TX_PROCEDURE_TY_101
rdm2_0
rcvd_1
rcad_1
ragetx
prevki_0
preemptive_1
hlamm_6
hlamm_4
hlamm_2
hlamm_0
functstat_8
functstat_6
functstat_4
functstat_2
functstat_1
esrddxsimp_4
esrddxsimp_2
ecd_1
dwt
drwt_4
drwt_2
drsex_4
drsex_2
drrace_9
drrace_7
drrace_5
drrace_3
drrace_1
drcmv_3
drcmv_1
drage_2
dracesimp_3
dracesimp_1
dhtn2_1
dht100
dhcv_0
ddm_0
dcd_0
dage
ahd1_4
ahd1_2
ahd1_0
0
0.02
0.04
0.06
0.08
0.1
0.12
0.14
3rd
2nd
1st