The Risk of Re-intervention after Endovascular Aortic Aneurysm Repair

OMNEYA AMR ATTALLAH

Doctor of Philosophy

ASTON UNIVERSITY

June 2015

ASTON UNIVERSITY

The Risk of Re-intervention after Endovascular Aortic Aneurysm Repair

OMNEYA AMR ATTALLAH

Doctor of Philosophy, 2015

Thesis Summary

This thesis studies survival analysis techniques dealing with censoring to produce predictive tools that predict the risk of endovascular aortic aneurysm repair (EVAR) re-intervention. Censoring indicates that some patients do not continue follow up, so their outcome class is unknown. Methods dealing with censoring have drawbacks and cannot handle the high censoring of the two EVAR datasets collected. Therefore, this thesis presents a new solution to high censoring by modifying an approach that was incapable of differentiating between risks groups of aortic complications.

Feature selection (FS) becomes complicated with censoring. Most survival FS methods depends on Cox's model, however machine learning classifiers (MLC) are preferred. Few methods adopted MLC to perform survival FS, but they cannot be used with high censoring. This thesis proposes two FS methods which use MLC to evaluate features. The two FS methods use the new solution to deal with censoring. They combine factor analysis with greedy stepwise FS search which allows eliminated features to enter the FS process. The first FS method searches for the best neural networks' configuration and subset of features. The second approach combines support vector machines, neural networks, and K nearest neighbor classifiers using simple and weighted majority voting to construct a multiple classifier system (MCS) for improving the performance of individual classifiers. It presents a new hybrid FS process by using MCS as a wrapper method and merging it with the iterated feature ranking filter method to further reduce the features.

The proposed techniques outperformed FS methods based on Cox's model such as; Akaike and Bayesian information criteria, and least absolute shrinkage and selector operator in the log-rank test's p-values, sensitivity, and concordance. This proves that the proposed techniques are more powerful in correctly predicting the risk of re-intervention. Consequently, they enable doctors to set patients' appropriate future observation plan.

Keywords: Endovascular Aortic Aneurysm Repair, Survival Analysis, Censoring, Feature Selection, Model Selection, Artificial Neural Networks.

Dedication

I would like to dedicate this thesis to my dear parents, husband, and daughter and son.

Acknowledgements

First of all, I would like to express my gratitude to my supervisor, Dr. Xianghong Ma, for her guidance, patience and understanding. Dr. Ma has been a great supervisor and has given me a lot of helpful advice throughout my PhD studies. Her support was fundamental to the accomplishment of this research. I would also like to thank Alan Karthikesalingam and his team at St George's and Leicester Vascular Institute for supplying us with the two datasets.

I would like to express my sincere gratitude to my lovely husband, Dr. Amr El-Helw for his patience and support throughout my study period. He has been very understanding and supportive; especially in my down time. Amr was the one who found me the opportunity to do a PhD at Aston University. He helped me a lot by travelling with me to the UK to meet my supervisor, and taking care of my daughter while I was busy.

Importantly, I would like to express my warm appreciation to my dear parents. Without you both I would not have achieved this. Thanks Mum and Dad for all your prayers, love, and support. I would like to acknowledge my mum, Dr. Hala Mandour for her helpful discussions and advice. You have always been there for me. Thank you Dad, Captain Amr Attallah, without you I would not be where I am today, having completed my Masters and PhD studies.

Last but not least, I would like to acknowledge my dearest friend Dr. Shaza Zaghlool for helping me with the R software, proof reading my thesis, and numerous useful discussions. I would also like to express my gratitude to my closet friend Eng. Amina El-Zein for proof reading my thesis. Thanks to my friend Dr. Heba Shabaan for her advice and help. Finally, I would like to thank my sister, brother and all my friends for their love and support.

List of Contents

List of Abbreviations

AAA	Abdominal Aortic Aneurysm
AdaBoost	Adaptive Boosting
AFT	Accelerated Failure Time
AIC	Akaike Information Criteria
ANN	Artificial Neural Network
AUROC	Area Under Receiver Operating Characteristics curve
BD	Bayesian Dirichlet
BDe	Bayesian Dirichlet score Equivalence
BDeu	Bayesian Dirichlet score Equivalence and Uniform priors
BIC	Bayesian information Criteria
BN	Bayesian Network
BP	Back Propagation
CI	Concordance Index
CT	Computed Tomography
DAG	Direct Acyclic Graph
EVAR	Endovascular Aortic Repair
FA	Factor Analysis
FR	Feature Ranking
FS	Feature Selection
FSFS	First Stage Feature Selection
GA	Genetic Algorithm
HR	Hazard Ratio
IFR	Iterated Feature Ranking
LASSO	Least Absolute Shrinkage and Selector Operator
KNN	K-Nearest Neighbors
KM	Kaplan Meier
MCS	Multiple Classifiers System
MDL	Minimum Description Length
MLP	Multilayer Perceptions
MPL	Maximum Partial Likelihood
PCA	Principal Component Analysis
PH	Proportional Hazard
PLANN	Partial Logistic regression Artificial Neural Network
REINT	Risk of Endovascular aortic repair Intervention
ROC	Receiver Operating Characteristics Curve

SVM	Support Vector Machine
SWFMS	Stepwise Feature Model Selection
TAN	Tree Augmented Naïve Bayesian

List of Tables

List of Figures

Chapter 1: Introduction

This chapter gives a brief description of the endovascular aortic repair surgery. It also highlights the thesis's motivation and rational, aims and objectives, main contributions and organization. This chapter starts with an introduction about abdomen aortic aneurysm and how it can be treated through an endovascular aortic repair operation instead of an open surgery. Then, it discusses the motivation and rational of this study. Next, the aims and objectives of the thesis are followed by the main contributions of this work. Finally, the thesis organization is given with a brief discussion of the contents of each chapter.

1.1 Endovascular aortic aneurysm repair

The aorta is the largest blood vessel in the body. It transfers blood from the heart to the chest and abdomen and then it splits to supply the legs. Sometimes, part of the aorta may balloon and form an aneurysm as shown in Figure 1.1. The size of the aneurysm may grow increasing the risk of its breach which could lead to death due to internal bleeding. For this reason, treatments should be considered to avoid rupture. Two popular solutions to this problem are endovascular aortic aneurysm repair (EVAR) surgery and open repair surgery.

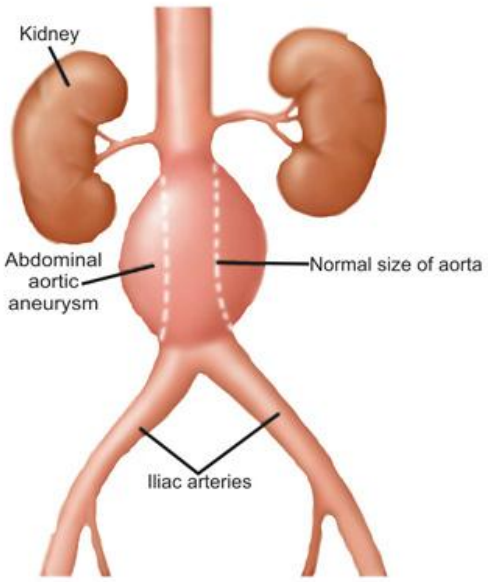


Figure 1.1 The aortic expanding over its normal size which is known as abodominal aortic aneurysm. The figure has been adopted from [1].

Open surgery is done by making a large cut in the abdomen to visualise the aorta and remove the portion of the blood vessel containing the aneurysm and replacing it by a synthetic graft material. Alternatively, EVAR puts extendable stent grafts in the aorta through a small incision in the groin. Stent grafts are tubes with metal frames coated with fabric. They are packed firmly in a thin plastic tube and inserted through the arteries. Afterwards, they are released and left to

expand inside the blood vessel. It works as a channel in which all blood passes through and not the aneurysm [2] as shown in Figure 1.2. EVAR carries significantly lower preoperative morbidity and mortality risks, enables shorter length of hospital stay and faster patient recovery rates compared to open repair surgery. Therefore, it is preferred by patients and recommended in medical guidelines as the better choice required for treating abdominal aortic aneurysm AAA [3].

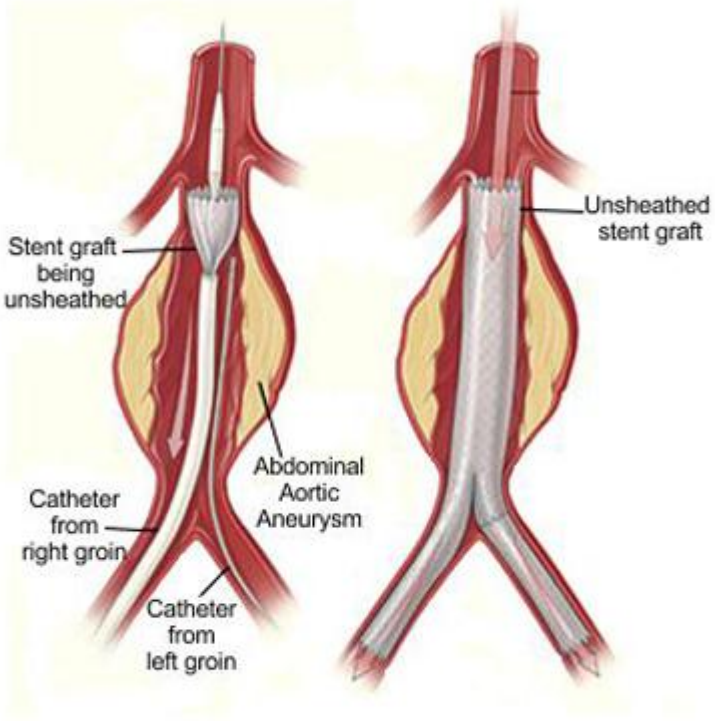


Figure 1.2 The insertion of a stent graft during the endovascular aortic aneurysm repair surgery. The figure has been adopted from [4].

1.2 Motivation and rational

The two EVAR datasets collected in this study included survival data. Survival data contains a variable called survival time, which indicates the time until an event of interest occurs for patients included in a study (in this case, the event of interest is the EVAR re-intervention). However, for some patients the time until the event of interest occurs is unknown. This is caused when these patients do not complete their follow up observations due to various reasons such as death, feeling better, or changing their residence location. This property is known as censoring and these patients are called censored patients. Censoring is the key characteristic that differentiates standard supervised data from survival data. Ignoring or deleting censored patients when constructing a predictive survival model may lead to prediction bias which increases when

the level of censoring increases (the percentage of censored patients to whole patients in the datasets). The EVAR datasets available in this study contain almost 91% censored patients which is a very high level of censoring making the survival model construction and classification difficult tasks. It also makes the feature selection process more complicated.

Survival analysis is the type of the analysis performed for this type of data. It deals with censoring without ignoring or deleting censored patients in order to construct a predictive model. Survival analysis methods are classified into standard statistical survival models and machine learning survival models. Standard statistical models include the non-parametric Kaplan Meier, the semi-parametric Cox's model, and the parametric accelerated time failure model. These methods take into consideration censored patients; however, they have several drawbacks. For example; they cannot capture complex and nonlinear relationships between variables of a dataset such as that existing in the EVAR datasets. The Kaplan Meier non-parametric method does not use all variables of a dataset to build the survival model. Although the parametric accelerated failure time model makes use of all the variables in constructing the model, it assumes a survival distribution for patient survival time. Choosing the appropriate survival distribution is a difficult task and a test must be performed to check if the family of the distribution that was chosen is suitable or not. The semi-parametric distribution does not make assumptions about the distribution and uses all variables available to construct the model. However, it assumes a proportional hazard which means that the hazard function of one patient is proportional to the hazard function of another one. Hazard function is the risk of event of interest occurrence for any patient at a given time *t*. To overcome these entire set of limitations, machine learning survival models are preferred for usage with model survival data. These methods have been shown to have superior performance compared to standard statistical survival models [5].

Standard machine learning methods cannot be used directly with survival data such as our EVAR datasets. The reason for that is that they are not designed to deal with situations where the outcome is censored [5]. Therefore, censoring should be handled to construct a predictive model using machine learning methods. The literature has shown several scenarios of dealing with censoring such as deleting, ignoring, weighting, repeating, or considering censored patients as event free [5-7]. These methods have several drawbacks and cannot be used to deal with high level of censoring. Therefore, they cannot be used with our EVAR datasets. For this reason, a new approach has been proposed in Chapter 3 to deal with high censoring in order to construct a survival predictive model using machine learning techniques. This model was constructed to predict aortic complications and classify patient risk of EVAR re-intervention. The proposed method is a modification of an approach proposed by Štajduhar and Dalbelo [6] which was incapable of differentiating between risks groups of aortic complications.

Feature selection is essential in the medical field as it reduces the time needed and the effort made by physicians to measure irrelevant and redundant features. It could avoid the over-fitting that might occur during the learning process of the predictive model. It also lowers its complexity and speeds up the prediction process [8]. Feature selection is widely used for standard medical data, though this task becomes more complex for survival data type due to the presence of censoring [8]. The literature has shown that a large number of papers employ feature selection methods for survival data based on Cox's proportional hazard model [9]. It is the most common statistical survival analysis model for medical applications; however, machine learning techniques are preferred for their capability in dealing with complex relations and nonlinearities between covariates which will improve predictions accordingly [10]. However, they are less used in survival analysis due to the fact that censoring makes them less capable to be directly used for survival data [11, 12].

Among the feature selection methods that use machine learning classifiers for survival data, are the popular partial logistic artificial neural network (PLANN) with backward elimination search strategy and its extension PLANN with automatic relevance detection (PLANN-ARD). These methods depend on repeating patients which leads to unbalanced and biased predictive models especially with censoring of high level [13]. It also increases the complexity and training time of these models and may increase the noise level existing in the datasets. Other methods include wrapped feature selection with other classifiers such as support vector machines, Bayesian classifiers and K-nearest neighbors. The limitation of these methods is in the way they handle censoring which is not appropriate for dealing with the high levels of censoring that exist in our EVAR datasets. Therefore, in Chapter 4 a new survival feature selection was proposed based on artificial neural network machine learning technique instead of standard statistical survival models. This approach is capable of dealing with the high level of censoring by embedding the previously proposed method dealing with censoring through the feature selection process. The proposed approach selects the best combination of hidden neurons and reduced number of features that yields accurate prediction of EVAR re-intervention while avoiding over-fitting. The new approach also uses iterated nested cross validation to overcome over-fitting and over-optimistic results that may occur during model construction and feature selection. Finally, it uses feature transformation and reduction methods, such as factor analysis, as an initial step to reduce features in order to lower the computational cost that might occur during the feature selection process.

Recently, multiple classifier systems have been preferred by researchers over individual classifiers, as they combine the strengths and output prediction of multiple predictive models which may consequently enhance classification performance [14]. In health studies, it

corresponds to taking several doctors' diagnoses in order to come up with a more confident and accurate decision. The literature has shown that those techniques have been used with standard statistical models such as Cox's proportional hazard model and accelerated failure time model, but they have several drawbacks. Others used machine learning methods; however, they either dealt with censoring by ignoring, deleting, or labeling according to observation time threshold. Again, these methods have several drawbacks and cannot be used with the high level of censoring found in the EVAR datasets. Therefore, in Chapter 5, a hybrid feature selection approach was proposed. It combined three well known classifiers namely support vector machines, neural networks, and K nearest neighbors. Predictions were combined using simple (unweighted) and weighted majority based on a survival metric which makes it applicable for usage with survival data. The approach also merges filter and wrapper feature selection approaches which usually improve prediction results. The wrapper method is used as it is known to yield better performance than the filter type since it uses a classifier to evaluate features. On the other hand, the filter method is a simpler and faster method compared to the wrapper method.

In order to produce a clinically validated model to be used in medical prediction, it should achieve acceptable performance on a dataset that was not used in the training process. Most of previous studies dealing with survival data employ only one dataset for FS and model construction and validate their results using K-fold cross validation, leave one out cross validation, or bootstrapping methods. However external validation is essential. It tests the capability of the predictive model in establishing the corresponding performance on other patients from another center who have dissimilar population from the one used to construct the model. The opportunity to cross check the model prediction upon different centers strengthens the results and study outcomes[15, 16]. Therefore, in this study two EVAR datasets were collected from two vascular centers in the UK (St George and Leicester vascular institutes respectively). Center 1 data was used for model construction, uncensoring, and feature selection processes, and Center 2 data for its validation.

1.3 Thesis aims and objectives

There is an obligatory need for lifelong monitoring after an EVAR operation which is considered expensive, varied, and poorly-calibrated [17]. However, monitoring procedures vary extensively [18] and there is a limitation in the choice of the best timing or modality. Patients may be exposed to radiation and contrast nephropathy as a result of frequent monitoring [19, 20]; however, for some patients complications required for treatment might be lost between monitoring [21]. For that reason, optimizing surveillance is very important [22, 23]. It is

considered as an important issue in clinics and affects the long-term cost-effectiveness of EVAR.

Usually, some patients need an EVAR re-intervention after a certain period of time. By specifying which patients are more likely to require re-intervention after 5 years (high risk patients) and which are less likely (low risk patients), a cost-effective and risk-stratified surveillance system could be achieved. This study focuses on developing and validating a predictive model for aortic complications after EVAR and selecting a reduced number of features that affect the prediction process. The model will help doctors to make decisions about patients' re-intervention risk and specify their monitoring plan. Those that are more likely to be re-intervened (high risk patients) will have a more frequent monitoring schedule. Others with lower likelihood will be monitored less regularly (low risk patients). Also, a lower number of features is selected leading to better prediction results compared to the full model and this may reduce the cost, effort and time for collecting additional unimportant features. Moreover, feature reduction reduces the model's complexity, making it faster and more efficient.

This thesis has several objectives which can be summarized as follows:

- Carry out a detailed review on the existing survival analysis techniques used to deal with censoring including standard statistical methods, machine learning techniques and multiple classifiers system.

- Examine the use of Bayesian networks (BN) and artificial neural networks (ANN) to handle the high level of censoring present in the EVAR datasets and use them to construct a risk stratified surveillance predictive model capable of determining the risk of EVAR re-intervention after 5 years.

- Investigate current feature selection and model selection approaches dealing with censored survival data.

- Combining feature selection, model selection, and variable transformation and reduction approaches with the construction of individual artificial neural network classifiers and multiple classifier systems to construct a reduced surveillance predictive model and compare their performance with common exiting methods.

1.4 Main contributions

This thesis presents major contributions which have resulted in original work. Part of this work was already published in two journals [24, 25] and the other parts are submitted in another

journal and still under review. This study includes three key areas of contribution (uncensoring, feature selection, and classification). The block diagram in Figure 1.3 highlights these three key areas and their inter-relationships.

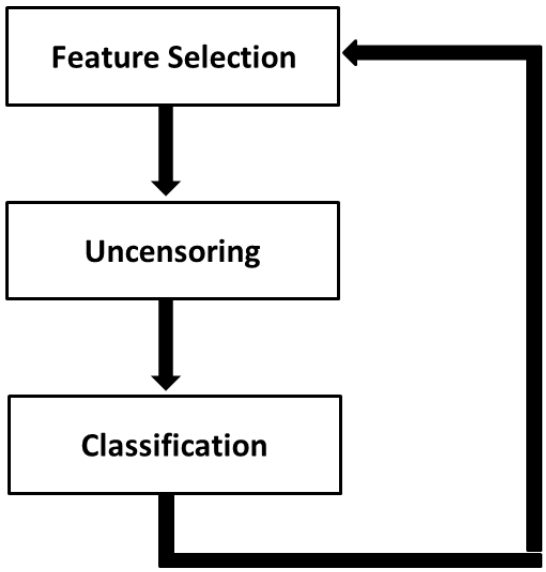


Figure 1.3 The Block diagram of the thesis's three key areas of contribution and their interrelations.

The main contributions discussed in this thesis can be summarized as follows.

1. Propose a new approach based on machine learning techniques such as Bayesian networks and artificial neural networks instead of the standard statistical survival models to avoid the drawbacks of current methods and deal with the censoring issue especially for high levels included in the EVAR datasets.

2. Propose a new survival feature selection approach based on individual artificial neural networks to overcome the limitations of current methods and use cross-validation and iterated nested cross-validation to produce a stable reduced surveillance predictive model.

3. Present a new hybrid feature selection approach which uses a multiple classifier system as a wrapper method to evaluate features, and combine it with a feature ranking process as a filter method to further reduce features.

4. Construct a multiple classifier system by combining predictions of support vector machine, artificial neural networks, and K-nearest neighbor classifiers using simple

(un-weighted) majority and weighted majority voting approaches that can be used for survival data to enhance prediction of individual classifiers.

1.5 Thesis organization

This thesis is divided into 6 Chapters.

Chapter 1 gives an introduction about endovascular repair surgery. It also explains the main contribution, motivation and rational, aims and objectives of this work.

Chapter 2 gives an introduction to survival analysis and describes the standard statistical survival models and machine learning methods specifically, the artificial neural networks and Bayesian networks used to construct models that are capable of predicting survivability. This chapter discusses techniques that handle the censoring issue, which is the main characteristic differentiating standard data from survival data. In addition, a survey is done on feature selection and model selection approaches that were employed in survival data in order to reduce the complexity of the predictive model and avoid over-fitting.

Chapter 3 discusses a modified Bayesian neural networks approach to deal with censoring. It first illustrates the difference between the original technique that was previously proposed by Štajduhar and Dalbelo-Basic in [6] and the new modified approach. Chapter 3 also describes the challenges faced by the approach. The Bayesian networks were used to deal with the censoring issue, while the neural networks was employed to construct a predictive model that was capable of predicting the risk of aortic re-intervention after 5 years of the endovascular aortic repair surgery. Moreover, the chapter explains the results of both techniques and shows the improvement of the new modified algorithm.

Chapter 4 describes a proposed feature selection approach in order to find a reduced number of features that are able to predict the risk of re-intervention. It uses factor analysis to first reduce the dimension of the data, and then embeds the proposed approach dealing with censoring in the feature selection process. The approach uses a stepwise greedy search to produce a final stable predictive model through cross validation and iterated nested cross validation. It used Center 1 for feature selection and the predictive model construction and Center 2 for its validation. This model was constructed using individual artificial neural networks to predict the risk of re-intervention. This chapter also discusses the results of the proposed approach and compares them with other popular variable selection methods constructed with the Cox's proportional hazard model, which is the most common survival model.

Chapter 5 illustrates a hybrid feature selection approach that combines the wrapper and filter methods. Variable selection is combined with a multiple classifier system instead of individual classifiers to predict the risk of aortic re-intervention after 5 years of endovascular aortic repair operation. The feature selection process is done in two phases; the first phase is feature selection and the second is iterated filter ranking. The first phase is a wrapper approach similar to the one used in the previous chapter. However, neural networks, support vector machines, and K-nearest neighbours were used to construct a multiple classifier system to predict the risk of re-intervention instead of individual classifiers. The second phase is a filter method that further reduces the number of selected variables. This chapter explains the performance of the proposed approach and compares it with three popular variable selections that use Cox's regression survival model.

Finally, Chapter 6 concludes this thesis and describes several aspects that can be investigated as future work in order to enhance the performance of the proposed approaches.

Chapter 2: Survival Analysis in Medical Applications

2.1 Introduction

Survival analysis methods are needed to deal with a specific type of data called survival data. They are used in many fields and can also be referred to as lifetime data analysis, reliability analysis, time to event analysis, and event history analysis [10]. Lately, they have been widely used in medical applications in order to model prognosis, determine the probability of patients' survival, or determine the time until an event of interest occurs. Generally, medical survival data has a time variable, which can indicate either the time when a certain event of interest occurs, such as mortality, or the time when a disease recurs. It can also describe the time until the last follow up, for a patient who did not complete his/her observations due to various reasons. Survival techniques analyse this unique data without ignoring any instance in the dataset even if the event had not occurred [26]. In this chapter, several standard statistical survival methods such as Kaplan-Meier, Cox proportional hazard model and the accelerated failure model will be reviewed. Moreover, machine learning methods that have been used for survival analysis will be reviewed along with multiple classifier system methods. Finally, the feature and model selection methods used in survival analysis will be discussed.

2.2 Censoring

2.2.1 Censoring Types

Generally, it is difficult to anticipate survivability after a surgical intervention. Some patients could not be observed in the whole period of the survival study, leading to a type of missing data called censored data. Censoring means that the outcome (ending point) for censored patients is missing. The only information available for censored patients is the time until the last follow up or death (if it is not the event of interest). Therefore, the time until the event of interest is unknown. For some patients, the event of interest did not happen during the study period and might happen before the study starts or after the study ends. They are considered as censored as well [27-29]. Survival analysis is the solution to this type of problem as it constructs a model for predicting survivability even if the data is censored.

2.2.1.1 Point Censoring

It is the type of censoring that occurs for patients that are monitored regularly from the beginning of the study until a specific censoring time T_c. The two types of point censoring are right and left censoring [30] .

- **Left censoring** arises when the event of interest occurs before the start of the study period. It rarely appears in medical survival data.
- **Right censoring** is the more common type of censoring in medical survival data. It can be decomposed into two types. Type I is when patients die or leave the follow up observation during the study due to certain reasons such as feeling better or moving to another place which makes it hard for them to complete the follow ups**.** In this situation T_c is less than the study period. Type II is when patients complete their follow up observations until the end of the study but the event of interest does not occur. In this type of censoring, the event of interest actually takes place after the study ends. In this case T_c is greater than the whole duration of the study [27]**.** Figure 2.1 shows the difference between them. The symbols $*$ and $\circ$ represent the time the event of interest took place or if it did not occur respectively.

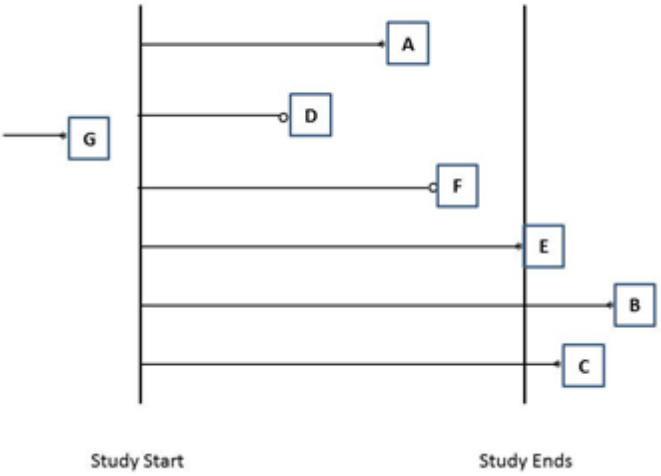


Figure 2.1 The difference between the types of point censoring (right and left point censoring). The letters refer to distinct patients. Patient G is left censored. Patients B, C, D, F are right censored. The symbols $*$ and $\circ$ represent the time when the event of interest took place or if it did not occur respectively.

For example, in patient A, the monitoring started and the event of interest occurred during the study period. Similarly, the event occurred at the end of the study in patient E. Both patients are not censored cases. On the other hand, the event of interest did not occur for patients D and F. They might have died or dropped out of the study due to any reason not relevant to the case study. In this case these patients are considered right censored. In the case of patients B and C, the event of interest occurred after the end of study period, so they are considered as right censored as well. Finally, Patient G is considered left censored because the event of interest occurred before the study even started [10].

2.2.1.2 Interval censoring

Interval censoring is a type of censoring in which monitoring is performed periodically. In other words, there is a schedule for the follow up observations to be taken. The only information available is the time of the first follow up and the time interval at which the censoring occurred, not its specific time. This case is known as single interval censoring. Sometimes, the only knowledge in hand is the time interval during which the follow up started and the other interval during which the censoring occurred which is called double interval censoring [31]. Figure 2.2 shows the difference between the two types of censoring. For the case of single interval censoring, the follow up start time is T_2 while the censoring occurs in the time interval between (T_3-T_4). While for double interval censoring, both the starting of the observation and the censoring occurrence lie between the time intervals (T_1-T_2) and (T_3-T_4) respectively.

Figure 2.2 The difference between double and single interval censoring. In single interval censoring, the censoring occurs in the time interval between (T_3-T_4). For double interval censoring, both the starting of the observation and the censoring occurrence lie between the time intervals (T_1-T_2) and (T_3-T_4).

2.2.2 Handling Censoring

The term censoring means that the outcome for some patients is unknown. Censored patients cannot be considered as unlabeled or labeled as they contain valuable partial information about the outcome [5]. In other words, the censoring time of these patients indicates that the event of interest definitely did not occur until the censoring time. Therefore, censoring is the main reason why standard machine learning techniques cannot be used directly with survival data. There are different approaches to handle censoring. The literature shows that the author in [32] creates an outcome variable called null martingale residual (NMR) which replaces the censor indicator and the observation time variables. This variable represents the difference between the observed and expected number of events that might have occurred till time *T*. The main limitation of this

method is that the output does not directly correspond to the time or probability of survival. Some authors [33-35] deal with censoring by deleting censored patients or considering that the event of interest will definitely not happen to those patients (zero target output). However, this leads to a biased predictive model [26], and this becomes apparent when the data is highly censored (most of the patients are censored, and only a few patients experienced the event). Others [36, 37] used only patients with complete follow up observations to build the predictive model. This approach is only applicable for slightly censored datasets. When the dataset has larger amount of censored data samples, this approach will result in a predictive model constructed with only small number of examples. The prediction will not be accurate enough (not sufficient). Another solution suggested by authors in [5, 6, 11, 38] is to use a Kaplan Meier (KM) survival method to weigh the output probability estimates of censored patients predicted by their proposed predictive models. This would work for low to medium censoring levels, for higher censoring levels KM analysis can result in overoptimistic estimated survival probabilities [6]. This is due to the small number of patients experiencing the event of interest and used to construct the predictive model. Moreover, KM does not consider the effect of prognostic variables for estimation. It also assumes that the censoring time is independent of the event time. De Laurentiis and Ravdin [39] used the Cox's model to impute the survival time. However Cox's model does not consider the time varying effects. Finally, some authors [7, 28, 40-43] recommended splitting the observation time into a number of intervals and repeating censored patients for all time intervals. This repetition may lead to an imbalanced and biased predictive model and it will also increase its complexity [13]. In addition, repetition cannot be used for noisy datasets containing unexplained variation in data samples.

2.3 Common Terms of Survival Analysis

2.3.1 Survival function

The survival function is the function *S(t)* used to calculate the probability of survival of a patient at a time *T* greater than a particular time *t*. It can be calculated from the cumulative distribution function *F(t)* as shown in Equations (2.1) and (2.2). Kaplan and Meier in [44] plot Kaplan Meier (KM) curves which is an estimation of the survival function and describes the probability that an event of interest does not occur at any time during observation with the presence of censored patients. The procedure of plotting the curve will be discussed later.

$$F(t) = P(T < t) = \int_0^t f(t)dt, \tag{2.1}$$

$$S(t) = P(T > t) = \int_T^\infty f(t)dt = 1 - F(t), \tag{2.2}$$

where; *F(t)* is the distribution function of *T* , *S(t)* is the survival function, and *f(t)* is the probability density function .

2.3.2 Hazard Function

The hazard function is the function *h(t)* that determines the instantaneous risk of the event of interest for any patient at a given time *t*. It is also known as the "instantaneous failure rate", "force of mortality", and "age-specific failure rate" [31]. It calculates the probability that for a given patient, the event of interest occurs for a short interval of time from *T* until $T + \delta t$, given that it did not happen earlier [45]. As the hazard function increases, the risk of the event of interest increases and vice versa [29]. It can be calculated from Equation (2.3).

$$h(t) = \lim_{\delta.t \to 0} \frac{P(t \le T \le t + \delta.t / T \ge t)}{\delta.t}. \tag{2.3}$$

According to Bayesian theory,

$$P(t \le T \le t + dt / T \ge t) = \frac{P(t \le T \le t + \delta.t)}{P(T \ge t)}, \tag{2.4}$$

Substituting equations (2.1) and (2.2) in (2.4),

$$\frac{P(t \le T \le t + \delta.t)}{P(T \ge t)} = \frac{P(T \le t + \delta.t) - P(t \le T)}{P(T \ge t)} = \frac{F(t + \delta.t) - F(t)}{S(t)}, \tag{2.5}$$

Then,

$$h(t) = \lim_{\delta.t \to 0} \left\{ \frac{F(t + \delta.t) - F(t)}{\delta.t} \right\}. \frac{1}{S(t)}. \tag{2.6}$$

Since,

$$f(t) = \lim_{\delta.t \to 0} \left\{ \frac{F(t + \delta.t) - F(t)}{\delta.t} \right\}, \tag{2.7}$$

where; *f(t)* is the derivative of *F(t)* with respect to t.

Then,

$$h(t) = \frac{f(t)}{S(t)}. \tag{2.8}$$

The survival function can be related to the cumulative hazard function *H(t)* [46] by Equations (2.9) and (2.10) as follows:

$$h(t) = -\frac{d}{dt}\{\log S(t)\}, \tag{2.9}$$

and so,

$$S(t) = \exp\{-H(t)\}, \tag{2.10}$$

where;

$$H(t) = \int_0^t h(t)dt. \tag{2.11}$$

2.3.3 Hazard Ratio

This is often used in therapeutic trials, when patients are divided into two groups (*A* and *B*) and are given different treatments to determine which group will have more risk of illness relapse. The hazard function is calculated in each group separately, then the Hazard Ratio *HR* is calculated from Equation (2.12) and used to compare the probabilities of illness relapse between the two groups at any given time to show which treatment is better. For example, if *HR* = 5, this means that the risk of relapse in group *B* is 5 times that of group *A*, and that the treatment used in group *A* is better [47].

$$HR = \frac{h_B(t)}{h_A(t)}. \tag{2.12}$$

2.4 Survival Rate measures

Survival rate measures are used to measure survivability after a certain disease or its progression after treatment. They can be used to determine the effect of two different treatments on separate groups of patients, to decide which is better. Different types of survival rates are discussed below.

- **Overall Survival** is the time from randomisation until death due to any reason even if it is not related to the case study [48].
- **Median survival** is the time at which 50% of the patients are alive and the other 50% are dead, or in other words half of them experienced the event while the rest did not [46].
- **Mean survival** is the area under the survival curve in the presence of censoring [31, 44]. If the largest time is the time to an event, the survival function will go to zero, and then the area under the curve can be calculated, otherwise it cannot. Hence, it is rarely used.

- **5 year survival** is the proportion of patients that have not experienced the event yet after 5 years of the study period [49].

2.5 Statistical Survival Analysis Methods

There are several techniques that can be used in survival analysis dealing with censoring. They can be classified into standard statistical and machine learning survival analysis techniques. Standard statistical methods can be divided into three groups. The first one is the nonparametric (Kaplan Meier product limit method). The second is parametric (Accelerated failure time model and parametric proportional hazard model). The third is a semi-parametric method (Cox's-proportional hazards method).

2.5.1 Non Parametric Survival Analysis (Kaplan Meier Method)

Kaplan and Meier [44] proposed a method known as Kaplan Meier or product limit estimate of the survival function *S(t).* It gives an estimation of the probability of patient's survival at any time for the whole dataset even if it is censored. It is widely used in clinical trials such as in the determination of the effectiveness of a specific treatment on illness relapse by calculating the number of patients at risk. Another example, is estimating the risk of re-intervention after a surgical intervention [50].

It is a very common method and used extensively in survival analysis due to its various advantages, such as its simplicity when calculating the observed probability of survival at any time. It does not assume any mathematical hypothesis such as a hazard distribution for the survival probability estimation. Hazard distribution means how the risk of the event of interest occurrence changes with time. If the sample size of the data is large enough, its estimated survival function would describe the real survivability of the population. It is widely used for comparing the survival functions between two groups of patients. This is helpful when the target is to differentiate between the survivals rates of two risk groups, or the effect of two treatments on different groups of patients. However, its limitation is when the level of censoring increases (more than 80 % of patients in a dataset are censored), as the estimated survival probabilities become overly optimistic [6]. Moreover, it does not take into consideration the covariates effect when calculating the survival function.

Three assumptions are made when plotting the Kaplan- Meier (KM) curve. First, the probability of survival for a censored patient at a specific time *t* is the same as the patient still observed until this time *t*. Second, survival probabilities for patients that entered the study at different times are equal. Third, the time assigned to the model is the time when the event occurs or is censored [50]. The KM curve $\hat{S}_o(t)$ is plotted using Equation (2.13). The estimation of the probability of a patient's survival to any point *t* is calculated from the cumulative probability of

the survival of each of the previous time intervals (calculated as the product of previous probabilities) [45].

$$\hat{S}_o(t) = \prod_{t_i \langle t} \frac{n_i - d_i}{n_i} = \hat{S}_o(t-1).(1 - \frac{d_t}{n_t}), \tag{2.13}$$

where; d_i is the number of the patients experienced the event of interest, n_i is the number of patients that are still being observed (including those who are censored and have not experienced the event of interest yet).

The following example will help to understand how KM works [50]. Suppose the dataset given is for patients receiving Ayurvedic therapy for HIV infection and the event of interest is death. Patients' observation times are 6, 12, 21, 27, 32,39, 43, 43, 46+, 89, 115+, 139+, 181+, 211+, 217+, 261, 263, 270, 295+, 311, 335+, 346+, 365+ (+ is the censoring time). Note that, the survival function is 1 at the beginning as no event has occurred yet. Table 2.1 shows the KM estimates of this example. Figure 2.3 is the KM curve of the previous dataset.

Table 2.1 Kaplan Meier non parametric survival method to calculate the survival functions of for patients receiving Ayurvedic therapy for HIV infection and the event of interest is death.

Event Time (t)	**Number of dead patients (d)**	**Number of patients still alive (n)**	**Estimated Probability of death (d/n)**	**Estimated Probability of survival (1-d/n)**	**Probability of survival at the end of time (L)**
6	1	23	0.0435	0.9565	0.9565
12	1	22	0.0455	0.9545	0.9565*0.9545=0.9130
21	1	21	0.0476	0.9524	0.9130*09524=0.8695
27	1	20	0.05	0.95	0.8695*0.95=0.8260
32	1	19	0.0526	0.9474	0.826*0.9474=0.7826
39	1	18	0.0556	0.9444	0.7826*0.9444=0.7391
43	2	17	0.1176	0.8824	0.7391*0.8824=0.6522
89	1	14	0.0714	0.9286	0.6522*0.9286=0.6056
261	1	8	0.125	0.875	0.6056*0.875=0.5299
263	1	7	0.1429	0.8571	0.5693*0.8571=0.4542
270	1	6	0.1667	0.8333	0.4959*0.8333=0.3785
311	1	4	0.25	0.75	0.4132*0.75=0.2839

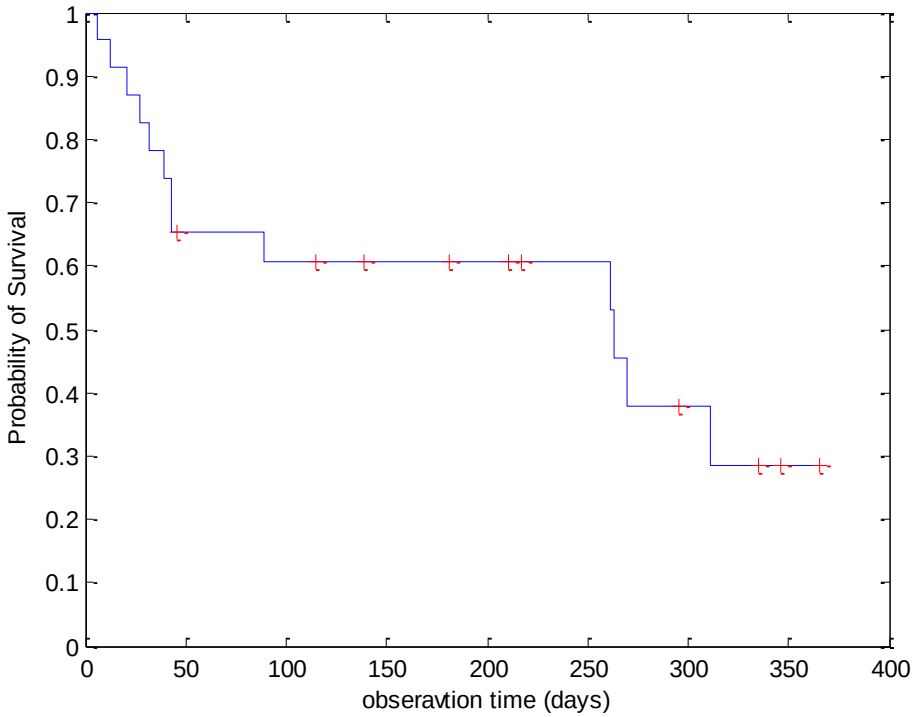


Figure 2.3 Kaplan Meier curve indicating the survival probability for patients receiving Ayurvedic therapy for HIV infection and the event of interest is death.

The literature shows that most authors use the KM curve for the purpose of either survival prediction or differentiating between risk groups. Authors in [51, 52] used it to predict the survivability and the prognosis of Head and Neck Squamous Cell Carcinoma and compared the results with the prediction of other survival approaches. In the papers [53-56], it was employed to estimate the overall survival of patients with cervical cancer, brain tumor, breast cancer, and multiple myeloma respectively. Farber et al used it in [57] to predict the survival in pulmonary arterial hypertension. In [58-60] the authors used it to differentiate between survival risk group recurrence predictions of large-B-cell lymphoma, ovarian cancer, and breast cancer.

2.5.2 Semi-parametric Survival Analysis (Cox-proportional hazards method)

Semi-parametric methods are like the Kaplan Meir non-parametric models in terms of not assuming a hazard distribution for making a prediction. However, they are like parametric methods that will be discussed in the next section in using the variables information in constructing the predictive model. This is the reason why it is called a semi-parametric approach. It is commonly used in medical applications, for example comparing the effect of a new treatment with an old one. It is preferred by clinicians over parametric models due to the difficulty in choosing an appropriate survival distribution and the strong assumptions made for

the latter model [61]. However, there is no need to assume a hazard distribution for semi-parametric models [30].

A Cox-proportional hazard or Cox regression is a popular semi-parametric model that builds a hazard model using covariate values by assuming a proportional hazard which means that the hazard function of a patient is proportional to the hazard function of another one [62]. In other words, the hazard ratio is constant over time (time independent as shown in Equation (2.14). Cox is used to estimate the risk of occurrence of an event given the prognostic variables at a time *t*. The model's output will be a hazard as a function of time and specific covariates. It is calculated using Equation (2.14) and the hazard ratio is calculated from Equation (2.15).

$$h(t, x_i, ...x_k) = h_o(t)\exp(\beta_1 x_1 + \beta_2 x_2 + ...\beta_k x_k) = h_o(t)\exp(\mathrm{B}X), \tag{2.14}$$

$$HR_{cox} = \frac{h_o(t)\exp(\beta_1 x_{1,i} + \beta_2 x_{2,i} + ...\beta_k x_{k,i})}{h_o(t)\exp(\beta_1 x_{1,j} + \beta_2 x_{2,j} + ...\beta_k x_{k,j})} = \exp(\mathrm{B}(X_i - X_j)), \tag{2.15}$$

where; *k* is the number of covariates of *x*, $X = (x_1, x_2, x_3, ...x_k)$ is a vector of covariates values, X_i is a vector of covariates values for patient *i*, *j* is a vector of covariate values for patient *j*, B = $(\beta_1, \beta_2, \beta_3, ...\beta_k)$ is a vector of regression coefficients, and $h_o(t)$ is the baseline hazard which is the hazard function for a patient with all his covariates available in the model equal to zero. No assumption is made about the baseline hazard. HR_{cox} is the hazard ratio of two different patients calculated with Cox's model. In order to fit the semi-parametric proportional hazard model, β is needed to be estimated. The maximum partial likelihood (MPL) approach is used which maximizes the partial likelihood estimate for B denoted as $\hat{\mathrm{B}}$. MPL does not depend on $h_o(t)$ for its estimation. It does not assume a prior for $h_o(t)$ [45, 63]. This is done using Equations (2.16) and (2.17).

The partial likelihood function for Cox PH model is given by

$$L(\hat{\mathrm{B}}) = \prod_{i=1}^{n}\left(\frac{\exp(\mathrm{B}X_i)}{\sum_{j\in R(t_i)}\exp(\mathrm{B}X_j)}\right)^{\delta_i}, \tag{2.16}$$

where; $\hat{\mathrm{B}}$ represents the maximum partial likelihood estimate of B, which is calculated by maximizing the partial log likelihood function $l(\hat{\mathrm{B}}) = \ln(L(\hat{\mathrm{B}}))$,

$$l(\hat{\mathrm{B}}) = \sum_{i=1}^{n}\delta_i(\mathrm{B}X_i) - \sum_{i=1}^{n}\delta_i \ln\left(\sum_{j\in R(t_i)}\exp(\mathrm{B}X_j)\right), \tag{2.17}$$

where; $R(t_i) = \{j : t_j \geq t_i\}$ represents the risk set at time t_i, and δ_i is the censor indicator.

The Cox proportional hazard model appeared extensively in the survival analysis literature especially with medical data. It was used in [51, 64-67] to predict survival of gastric cancer, head and neck cancer, hepatocellular carcinoma, coronary disease, and heart failure respectively. MacKenzie, et al in [68] applied it to predict survival after coronary revascularization. Bambha, et al used it in [69] for assessing mortality in women with hepatitis C virus and HIV. In addition, [70-73] used it to predict survivability of breast, gallbladder and lung cancer patients. Cui, et al in [74] used it to test for recurrence of cardiovascular events. Sheets, et al in [75] adopted it to predict the recurrence of major depressive disorder in emerging adults. Emura, et al [76] employed it for predicting survival for liver and lung cancer patients. Zhang, et al [77] used it to estimate both survivability and recurrence of ovarian cancer.

2.5.3 Parametric Survival Analysis (parametric proportional hazard model and Accelerated Failure time model)

Even though non-parametric methods are commonly used in survival analysis, parametric estimates are also employed in cases that assume a hazard distribution [78]. The first difference between the two approaches is that the latter estimates the instantaneous risk of an event with the presence of censored data under the assumption of the survival distribution. The second difference is that it takes into account the information given by the covariates which may affect the hazard. In other words, estimates are done by assuming the hazard is a function of time and particular covariates. The most popular distributions are Weibull and Exponential distributions [30].

The Weibull distribution estimates the hazard function with the assumption that it varies monotonically with time. Table 2.2 shows how the Weibull method calculates the hazard, where; λ is the scale parameter, p is the shape parameter. If p is greater than 1, then this indicates that the hazard increases with time [78]. While if p is less than 1, the hazard will decrease with respect to time.

The exponential distribution is a special case of the Weibull distribution in which p is equal to 1. This means that the instantaneous hazard is constant; it does not change with time. Constant hazard leads to unchangeable probability of failure within the study period [30]. The Weibull and exponential distribution shapes are shown in Figure 2.4. Other distribution functions can be found Table 2.2.

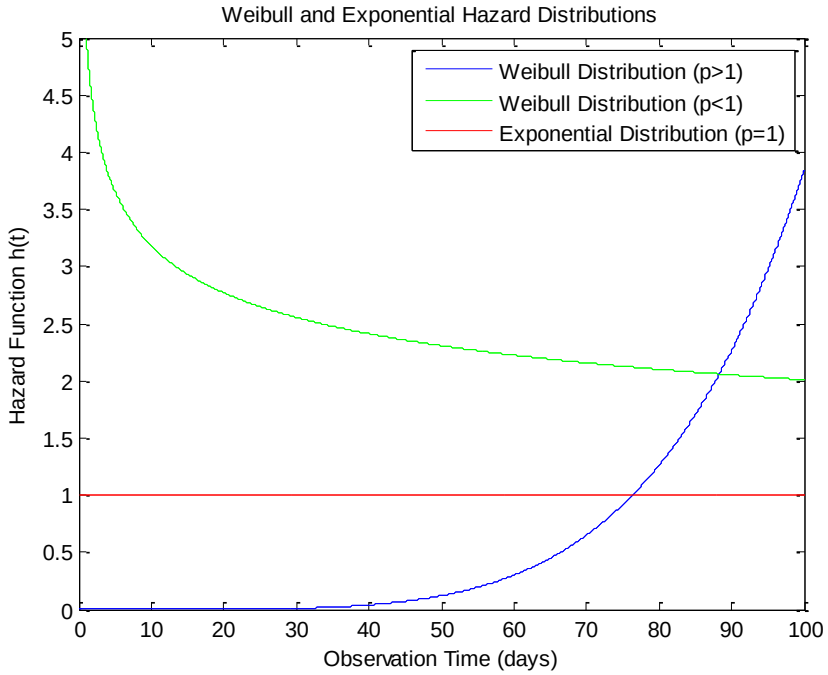


Figure 2.4 The weibull and exponential hazard distributions used to estimate the hazard function for parametric survival models.

Table 2.2 The common hazard distributions in survival analysis.

Survival Analysis distribution	Probability density function	Hazard function	Survival function
Weibull distribution	$f(t)=\lambda p(\lambda t)^{p-1}e^{-(\lambda t)p}$	$h(t)=\lambda p(\lambda t)^{p-1}$	$S(t)=e^{-(\lambda t)p}$
Exponential distribution	$f(t)=\lambda e^{-\lambda p}$	$h(t)=\lambda$	$S(t)=e^{-(\lambda t)}$
Gamma distribution Γ is the gamma function	$f(t)=\frac{\lambda(\lambda t)^{p-1}e^{-\lambda t}}{\Gamma(p)}$ $\Gamma(p)=\int_0^{\infty} t^{p-1}e^{-t}$	$h(t)=\frac{\lambda(\lambda t)^{p-1}e^{-\lambda t}}{(1-I_k(\lambda t)\Gamma(p)}$ $I_k(\lambda t)=$ $\int_0^t \lambda^{p-1}e^{-t}dt/\Gamma(p)$	$S(t)=1-I_k(\lambda t)$
Gompertz distribution	$f(t)=\lambda p e^{\lambda t}e^{p}\exp(-pe^{\lambda t})$	$h(t)=\lambda e^{pt}$	$S(t)=$ $\exp((\frac{-p}{\lambda})e^{\lambda t-1})$

The parametric proportional hazard (PH) is a well- known parametric survival model [79]. It is similar to the semi parametric Cox's model in terms of how the hazard function is calculated. They both assume proportional hazard and have the same way of describing the hazard ratio. However, the main difference between the two is that the former assumes a distribution for the baseline hazard, while this is not the case in the Cox model [80].

The accelerated failure time (AFT) [45] is another popular parametric survival technique. It uses variables information to construct a predictive model and determines their effect on survivability. Parameters estimated using the AFT model explain the influence of each variable on the mean survival. It assumes that a change in covariates may accelerate or decelerate the event time of a patient by a constant factor called the acceleration factor. AFT is preferred over the PH model as it could be used with more hazard distributions. Both the parametric PH and AFT techniques are not frequently used in clinical trials [80].They are commonly used in manufacturing.

The AFT explains the relation between survival probabilities and some covariates $\{x_k\}$ using Equation (2.18):

$$S(t) = S_o(t).\exp(\beta_1 x_1 + \beta_2 x_2 + ...\beta_k x_k) = S_o(t).\exp(\mathrm{B}X)) = \frac{S_o(t)}{\eta(X)}, \tag{2.18}$$

where; $S_o(t)$ is the baseline survival function, B is the vector of regression coefficients, and $\exp(\mathrm{B}X)$ is the acceleration factor denoted by $\frac{1}{\eta(X)}$ which is the ratio of survival times comparable to any value of *S(t)*. It is assumed that the covariates influence is constant and multiplicative on a time scale, so a change in any covariate value affects the survival by a constant factor which is the acceleration factor. It may accelerate or decelerate the event time for any patient.

The logarithmic linear form of the AFT model with respect to time is given by Equation (2.19). In other words, the AFT model produces a relation between the logarithmic transformation of survival times and covariates in a linear manner using Equation (2.19):

$$U = \log T_i = \mu + \mathrm{B}X_i + \varepsilon_i \sigma, \tag{2.19}$$

where; *log* T_i is the log linear model for random variable T_i , associated with the survival time with the i^{th} patient in the study, μ is the fixed effects or intercept σ is a scale parameter, X is

the vector of covariate values, ε_i is a random variable which is assumed to have a certain distribution and represents the error distribution. For each distribution of ε_i there is an equivalent distribution for *T*. The AFT is given the name of the distribution of *T* not log T or ε_i [80]. The survival function of *T* is determined using Equation (2.20):

$$\begin{aligned} S_i(t) &= P(T_i \geq t) = P(\log(T_i) \geq \log t), \\ &= P(\mu + \rho' x \ + \sigma\varepsilon_i \geq \log t), \\ &= P\left(\varepsilon_i \geq \frac{\log t - \mu - \rho' x}{\sigma}\right), \\ &= S_{\varepsilon_i}\left(\frac{\log t - \mu - \rho' x}{\sigma}\right). \end{aligned} \tag{2.20}$$

The parametric models are less frequently used in clinical trials than Cox regression models, due to stronger assumptions involved about the hazard distribution. In addition, a test must be made to check if the family of distribution that was chosen is suitable or not. Also, the AFT model has a difficulty in calculating regression parameters. Therefore, clinicians and statisticians prefer to use the Cox semi-parametric model to avoid the effort of model checking, and reduce the risk of misspecification of the distribution assumed [81]. However, the Weibull parametric PH model was adopted by Carroll [82] to model prostate cancer progression in two groups of patients treated with different treatment. Wang, et al [83] used the AFT with different hazard distributions for predicting the benefit of adjuvant chemoradiotherapy in gallbladder cancer. Royston, and Parmar [84] used the parametric PH model to model prognosis of both breast and bladder cancer. Parametric PH model with different hazard distributions was used in [81, 85] to predict survivability of patients with gastric cancer, progression of HIV. While in [61, 86-88] it was used for predicting survival rates of colorectal cancer, breast cancer and kidney transplant candidates.

2.6 Machine Learning Survival Analysis Methods

Recently, machine learning techniques have been used widely in the field of survival analysis in medical applications [5]. They are preferred over statistical survival methods such as; Kaplan-Meier, Cox's proportional hazard model, and accelerated failure time model. This is due to their capability to deal with complex relations between covariates which will improve prediction performances accordingly. Also, they are able to tackle nonlinearities in a dataset which is not the case with statistical approaches [10, 89]. Moreover, they are better than the non-parametric

survival statistical methods (Kaplan-Meier method) in considering the variable information in constructing the predictive model. Additionally, they do not assume a hazard distribution to construct the predictive model like parametric methods (Accelerated failure time model), therefore the difficulty of choosing the best hazard distribution to fit the model is avoided Finally, they do not assume that the hazard ratio is constant over time like the semi-parametric technique (Cox's semi-parametric model) which might not be applicable for some datasets [90]. Several methods have been used in the literature. These methods include Artificial Neural Networks (ANN) and Bayesian Networks (BN). The former is considered to be one of the most popular and widely used machine learning techniques especially in the medical area as it has an ability to yield good prediction even with noisy data [91] like the EVAR datasets used in this thesis.

2.6.1 Artificial Neural Networks for Survival Analysis

The artificial neural networks (ANN) are models stimulated by the biological nervous system of animals, in specific the neural networks of the brain [92]. The learning process in ANN is accomplished by changing the weights of the connected neurons with the help of the training data and a learning algorithm. A testing set is used later in order to test the performance of this ANN model that might be used for predicting the classes of new examples. The multilayer perceptron (MLP) is the most common and widely used ANN in the medical area. Survival analysis through MLP-ANN can be classified into three approaches; the direct prediction of survival time, prediction of survival status or probability, and ANN extension of Cox proportional hazard model [13].

2.6.1.1 Direct Prediction of Survival Time or hazard

An MLP-ANN model is able to predict the survival time of patients. Two applications are in prediction of breast and skin cancer [93, 94]. The authors of these papers trained the ANN with the censoring and mortality time**s** as the target value. They tried to reduce the bias that occurs due to censorship by not updating the network weights in the case of censored patients if only their predicted survival time is greater than actual censoring time. However, the bias still exists which might affect the results. Also, they did not allow the use of the full data, since the network will not be trained by censored cases with predicted survival time greater than their censoring time. The authors in [95] split their data into 7 partitions according to the follow-up time intervals existing in the data. They deleted censored patients; however, they overcame the problem of the small sample size for predictive model construction by repeating patients by a classification rule. Patients in each partition were classified as 1 if breast cancer relapsed and 0 if it did not relapse and they were repeated as zero until breast cancer relapsed. MLP-ANN models were constructed for each of these partitions to predict breast cancer relapse at each time interval. The drawback of this method is ignoring censorship and replicating data examples

which will lead to a highly biased predictive model [96]. The authors in [97] designed an MLP network that detects the areas of feature subspace that are homogenous to a given treatment in order to predict the survival time for skin and lung cancer. It also deals with censoring by comparing censored patients to those having the same survival experience. They presented a learning process that considers patients similar if they have a short difference in survival time and vice versa. Taşdelen, et al [98] used three MLP networks to estimate the prognosis of headache in elderly people with three different observation time durations. These models were constructed while considering patients that do not experience the event of interest during the specified observation time as event free. Ohno-Machado [27] implemented two scenarios for ANN. The first one consisted of several ANN models with a single output neuron and the other with only one ANN with multiple output neurons. Both of them were constructed to model prognosis of coronary heart failure disease. He compared their results with the standard statistical survival models to show their superiority.

The authors in [36, 37] employed MLP-ANNs to predict survival times for different kinds of cancer. They used only uncensored patients to construct the model to avoid bias and left censored and a few uncensored patients to test it. However, this might be only applicable for low level censored data with few censored examples. In [99] an MLP-ANN was used to predict the survival time of lung cancer patients. The dataset contained a low level of censoring so the authors let the censoring indicator be an input variable to the ANN model and ignored censorship. Fornili, et al [100] proposed an MLP-ANN to model the hazard distribution and predict breast cancer survivability. It uses a likelihood approach to predict the hazard function with the presence of censoring. Gohari , et al [101] proposed another method that uses ANN for predicting survivability of colorectal cancer at 1,3, and 5 years; however they ignored the effect of censoring and considered patients who survived till 1,3,5 years as zero class output.

The previously discussed ANN methods have several drawbacks. The first drawback is the partial use of the examples involved in a dataset in approaches like [93, 94]. In other words, the ANN stops training for censored cases. This will limit the ability of the ANN to be well trained especially when the number of censored patients is large like that included in the EVAR datasets. Secondly, is deleting censored patients or using only censored [36, 37] cases to construct a predictive model which lead to biased predictions and the bias increases when the level of censoring increases. Therefore, it cannot be used in our case. Although the method proposed in [95] overcame the problem of the small sample size resulted from deleting of censored cases by repeating patients through a classification rule to construct a predictive model, it cannot be used with the few examples of uncensored cases available in the EVAR datasets. The ANN to have efficient training, there should be enough number of examples. Thirdly, is considering censored patients as event free in approaches like [98, 101] which might

also lead to a highly biased predictive model especially in a high level of censoring such as that located in the EVAR datasets. Finally, is increasing the number of output neurons of the ANN like in [27]. This lead to an increase in the complexity of the ANN and estimation process, therefore it is not recommended to be used in this study.

2.6.1.2 Prediction of Survival status or Probability

The authors in [33-35, 102-105] used ANN for predicting breast cancer recurrence or 5 year survival in hemodialysis, liver, lung and breast cancer. The datasets used contained either medium or low level censored patients, so the authors deleted them before constructing the model or considered them as event free according to a threshold. This leads to biased predictions. Abdelghani and Guven in [106] modified Delen's approach [105] by considering only patients that died before 5 years as not survived if the cause of death was breast cancer. This improved ANN predictions. However, Ravdin and Clark [40] did not ignore censored patients or consider them as patients with zero class value to overcome this problem. Alternatively, they coded the time variable and considered it as an additional input variable to the MLP-ANN to predict breast cancer recurrence. They divided the observation time into *n* time intervals. Then, transformed the data by repeating patients that had recurrent breast cancer for all time intervals with survival status equal to 0 which is then transferred to 1 at the time interval it recurred and remains one for the rest of interval. For the case of censored patients, they were replicated only for the time interval lower than or equal to their censoring time. Although this method dealt with censorship, it increased the size of the dataset and the ANN correspondingly. Therefore, the ANN has a higher computational cost.

Liestbl, et al [41] overcame the huge increase in the dataset size that appeared in Ravdin and Clark [40] method. They used the Cox proportional hazard model to determine the event occurrence times for censored patients. They transformed the data output class value into *n* binary values equivalent to time intervals to construct one MLP with *n* output neurons, instead of constructing *n* MLPs and replicating patients with all of them. Each patient status will be equal to 0 then it changed to 1 at the death time interval. Later, Brown in [28] modified Liestbl, et al [41] technique by considering a censored patient to have survived until the end of any of *n* time intervals, if he/she has survived more than half of this time interval before being censored. The survival status for censored patient is undefined at time intervals after his/her censoring time. The error for undefined statuses will be equal to 0 when training the ANN model which prevents updating the weights for these cases. The main disadvantage of this method is the partial use of the examples included in the dataset available when training the ANN.

Chi, et al [42] and Street, et al [107] employed an MLP-ANN to predict breast cancer recurrence probability. The dataset target output was transformed into *n* intervals as well as the other

Brown and Liestbl methods [28, 41]. However, the recurrence probability for a patient was considered 1 till recurrence time and then 0 afterwards and for the street method -1 afterwards. For censored cases, KM plots were used to estimate the probability of recurrence. Although these three methods solved censoring issue, they still increased the size of the ANN architecture, as they increase the number of output neurons leading to an increase in ANN complexity. Additionally, as mentioned before, the KM estimates are over optimized for highly censored datasets which might consequently affect the results.

Biganzoli, et al [108] proposed an MLP-ANN, named partial logistic regression ANN (PLANN). It is a modification of the Ravdin and Clark method [40]. They used the time variable as an additional input to the network with has one output layer of just one node that indicates the survival status. However, patients that experienced the event of interest are not repeated after the event occurrence like Ravdin. Censored patients are repeated with survival status equal to 0. Moreover, logistic transfer functions were used in both hidden and output layers. PLANN expresses the hazard as a function of time and input variables without proportionality assumptions. Additionally, it does not assume that the effect of variables is constant over time. However, due to the large amount of data repetition, the network is not scalable (unbalanced) and highly biased and might over-fit [13]. Extensions were done on Biganzoli's PLANN to take into account the cause-specific hazards in [43] and employed Bayesian theory when learning the network. This prevents over-fitting and gives a natural sorting for input variables. Other examples of Bayesian extensions for discrete time model referring are PLANN-ARD (Automatic Relevance Determination) [7, 109] and for a continuous time model referred to as Conditional Hazard Estimating Neural Network (CHENN) [110] which perform feature selection by sorting variables according to their importance to the model and remove irrelevant variables with a low relevance score . CHENN was used later by Damato, et al [111] to predict choroid cancer. PLANN and its versions were used later by several authors to predict pancreatic, breast, intraocular, ocular, uveal, and colorectal cancer survivability [15, 110, 112-116]. Spelt, et al in [117] used an ensemble PLANN with weight decay method to overcome over-fitting and predict liver survivability after colorectal cancer. Hamdan and Garibaldi in [118] proposed a fuzzy system using PLANN to predict ovarian cancer survivability.

An MLP-ANN was used in [119] which considered the effect of censoring on its training without transforming data. This was done using the concordance index (CI) as an objective function which is a metric used to measure the quality of predictive survival models [120]. The network was used to predict the survival status of prostate cancer. Kalderstam, et al [121] proposed using ensemble (several) MLP-ANN networks constructed by different parts of the

data and trained directly on the CI using a genetic algorithm to predict breast cancer recurrence. The advantage of using CI is that it accounts for the censorship issue. However, these methods carry out parametric assumptions related to the effect of variables with the survival time. Moreover, maximizing the CI requires a high computational cost, as it is a discrete optimization process [122].

These ANN methods illustrated earlier have several limitations in the way of handling censoring which make them inappropriate to be used with the highly censored EVAR datasets used in this study. The first limitation is deleting censored patients or considering them as event free according to a threshold like in [33-35, 102-105]. This leads to biased predictions and this bias increases with the high level of censoring as mentioned earlier. Secondly, is the splitting of survival time into intervals and repeating patients to these intervals like in [40, 108] and using one or several ANN models like [41] to make prediction. The repeating process leads to unbalance, biased, and complex predictive model. It also increases the complexity of estimation process and increases the noise level for noisy datasets like our EVAR data. Therefore, these approaches will not be applicable in this study. Thirdly, is the increase of the size of ANN model which as mention before increase the complexity of the estimation problem, for this reason, it is not recommended to be used in this thesis. Finally, is using only uncensored cases to train ANN like in [28] which produce insufficient results in datasets with a high level of censoring like the EVAR due to small number of examples used to train the ANN model. Hence, these methods cannot be used in our case.

2.6.1.3 ANN as an Extension to Cox's Model

As mentioned before, Cox proportional hazard models cannot capture complex and nonlinear relations of the survival data. Therefore, an ANN was used as an extension to Cox's model to overcome this problem. The first extension was proposed by Faraggi and Simon [123]. The method generalized the Cox's model to enable it to deal with complex and nonlinear relations between data. The MLP-ANN constructed has one hidden and output layer with logistic and linear activation functions respectively. They replaced the linear function of Cox model ($\beta' x$) by the nonlinear output of the ANN. It was used later in [124] to compare its performance with the standard Cox model for classifying breast cancer patients. Afterwards, Xiang et al [125] proposed a technique that replaced the censored survival times which are the input to Cox model with their expected values estimated by KM method, then used an ANN outputs to replace the linear function of cox's model and compared the results with Faraggi and Simon method. The authors in [126] proposed using an ANN as extension to parametric survival models with log-logistic, lognormal survival distributions to predict breast cancer survivability. While those in [122] introduced another approach on extending an ANN to Cox in which the

linear function of the later is replaced by the output of the neural network optimized by the CI. The advantage of using ANN-Cox extensions is that they conserve all the benefits of using the Cox's statistical model. Moreover, they consider the time varying effects. It can also be used with parametric survival models, [13]. However, they are considered as a sub-optimal solution to baseline variations. Moreover, they make parametric assumptions on the effect of variables with survival time [122]. Choosing the appropriate survival distribution is a challenging task. It is difficult to choose the best distribution that fits the data. Complex tests are needed to test this assumption. For these reasons, these methods were not used in our study.

2.6.2 Bayesian Networks for Survival Analysis

This is a probabilistic network that uses probability theory to calculate joint probabilities between variables. It is also considered as a graphical model, as it looks like a graph representing relations between numbers of nodes (also called vertices) which correspond to random variables. These nodes may be either connected by arcs to each other or not connected. These nodes may be considered as a parent node when an arc comes out of it and enters another node (child node). When two children nodes are not connected, but they have the same parent node, this indicates a conditional independency between these two variables. Figure 2.5 shows a simple structure of a Bayesian Network showing how nodes (variables) are either connected by arcs or not representing conditional dependencies and independencies. For example, node 1 is not connected to node 4, but these two nodes are conditionally independent given node 5. Hence, this type of network presents a group of conditional dependencies and independencies between all variables of a dataset used to create the network itself [127].

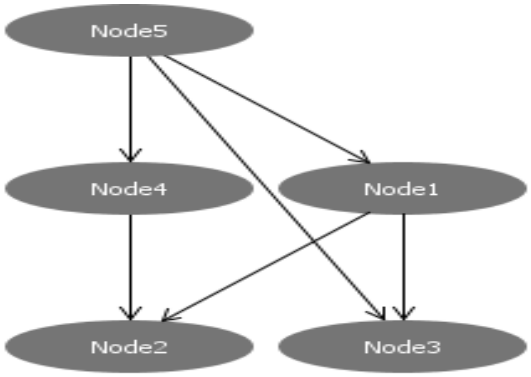


Figure 2.5 A simple example of a Bayesian Network structure showing how nodes (variables) are either connected by arcs or not, representing conditional dependencies and independencies. Each node represents a variable in a dataset.

BN was used for survival analysis in [128] in order to build a model that predicts the risk of death from sickle cell disease within 5 years. A patient that survived more than 5 years was considered as one and vice versa even if they were censored. Two different datasets were used

for constructing and validating their results. Later, a hybrid Bayesian network was constructed in [129] to predict breast cancer prognosis. A patient that survived more than 5 years was also considered as 1 and vice versa even if he/she is censored. The hybrid network was a combination of ANN and BN. The output of the ANN was used to calculate a confidence value which was used as an input node to the BN. In [130] a BN was designed to deal with censoring and predict breast and colon cancers prognosis using expectation maximization method. The authors in [11] proposed approaches that dealt with censoring through weighting and repeating each censored patient once as experiencing the event and once as event free. They constructed a Bayesian classifier (Naïve Bayes) to predict the recurrence of prostate cancer. Štajduhar, and Dalbelo-Bašic [131] used the weighting process proposed by Zupan et al [11] to deal with censoring and constructed a BN to predict breast cancer recurrence. They compared the results with BN learned by considering censored patients as event free as well as the Cox model. Then, they proposed a weighting technique in [6] based on KM analysis to deal with censorship using BN. Authors in [132] used BN for estimating perioperative risk of clostridium difficile infection following colon surgery. Rancoita, et al [133] used a decision tree referred to as a survival tree to solve the censoring issue, and constructed a BN to impute missing covariate data so that the survival tree is able to use all the data and classify diffuse large B-cell lymphoma and marginal zone lymphoma. For each of them, two separate datasets were used. The authors in [5] proposed an approach called an inverse probability of censoring weighting using KM estimator to handle censoring by weighting it with its inverse probability for times greater than its censoring time. A BN was then constructed to predict cardiovascular diseases.

The main drawback of the previously discussed method is the way of handling censorship by considering patients as event free, repeating patients, or weighting censored patients by KM analysis. The disadvantages of repeating censored patients and considering them as event free is the production of unbalanced, complex and high bias prediction results especially with highly censored data. KM analysis produce good results with low to medium censoring, however for high censoring, KM estimates are over optimized which might affect results consequently. For these reasons, these methods were not used in this study.

2.6.3 Multiple Classifier System

Recently, many studies have used multiple classifiers systems (MCS) for classification instead of single classifiers. Wolpert stated in [134] that there is no classifier that is suitable for solving all classification tasks; because each one is ideal in a specific area. Thus, MCS is important as it merges the outputs of multiple classifiers [135, 136]. Classifications are combined using a fuser which usually improves prediction performance and is a widely used method in pattern recognition as stated in [137, 138]. Figure 2.6 shows a MCS structure combining three classifiers.

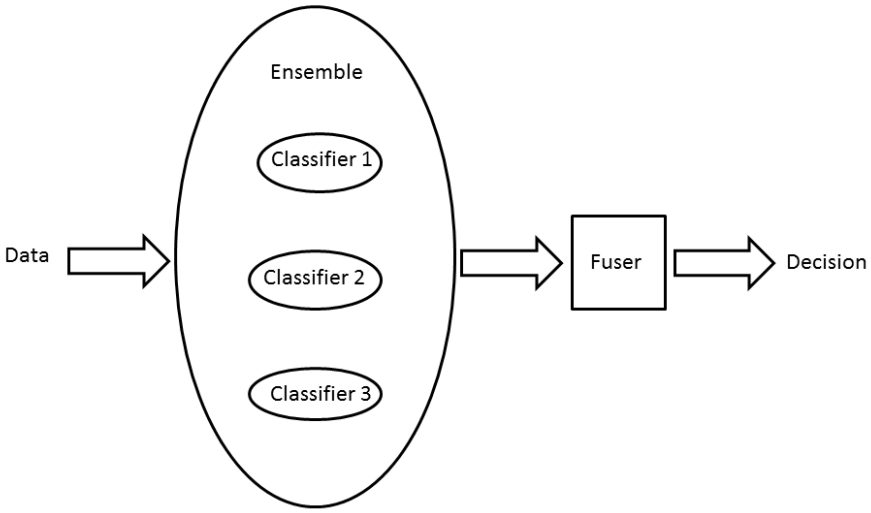


Figure 2.6 A Multiple Classifier System combining three classifiers by a fuser to come up with a final decision.

2.6.3.1 Multiple Classifier System Advantages

Researchers constructed a MCS to solve their classification task due to its various advantages. It can deal with small size datasets in model construction through boosting and bagging techniques. For large datasets, MCS partitions data and uses it to train different classifiers and combine their estimates by a combination or fusion rule [14]. Moreover, it merges the strengths of all algorithms used for training which usually improves prediction performance of the best classifier in the pool. Under some circumstances, this enhancement has been verified analytically in [139]. Additionally, MCS could produce a final optimal model for those classifiers that use learning algorithms which depend on the initial point such as C4 decision tree classifier. This is due the fact that MCS could begin at different starting points in the search space [140]. Finally, it has the ability to select the local optima model from the number of classifiers available [14].

2.6.3.2 Multiple Classifier System Topologies

In order to construct a MCS, the system topology should be decided initially. MCS has two topologies; parallel and serial. The former is the most common way used to connect classifiers [141]. The same input enters all classifiers, and the final prediction is determined based on outputs of all classifiers. On the other hand, in the serial topology, classifiers are connected in series following some sorting over them. If the first classifier predictions are not accurate enough, input variables will enter the second classifier. Classifiers are added iteratively according to their order until predictions are enhanced [142, 143]. It is more preferable when the

classification cost is essential. The first classifier is the weakest with the lowest computational cost and the second is stronger and has a higher cost than the previous one [144]. The AdaBoost ensemble classifier is an example of serial topology [145]. Figure 2.7 shows the difference between the two topologies.

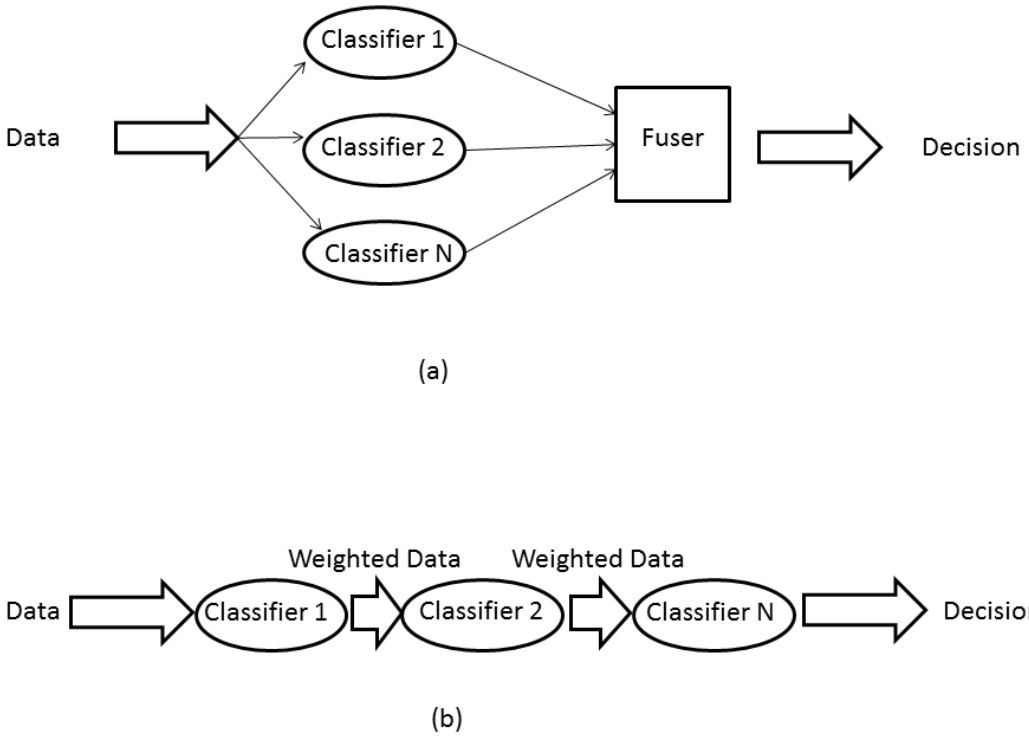


Figure 2.7 The difference between topologies of a multiple classifier system; (a) parallel (b) serial.

2.6.3.3 Multiple Classifier System Fusion Design

Fusion is about the way predictions of individual classifiers are combined to produce final decisions [14]. Popular MCS methods include bagging, boosting, stacking, and voting [146]. Note that classifiers of the MCS could be trained by either same training set, different subdivisions of the training set, or different partition of the feature space [147].

- **Bagging**

It is also known as bootstrap aggregation [146]. Bootstrap resampling method is used to produce a number of data subsets *N* that was randomly generated from the original data. Each subset (consisting of training subset and testing subset *n*) is employed to construct an individual classifier such as an ANN. As a result, several ANN classifiers will be created and combined in the pool using a suitable combining method.

Bagging usually enhances the prediction of an individual classifier when the model is unstable. This is due to the nature of boostrapping when generating subsets of data. It tries to ensure that these subsets are completely different. The classification will be greatly enhanced as a consequence. However, when the predictive model is stable, bagging will either make no or slight improvement on prediction performances [148].

- **Boosting**

This is similar to bagging in which the training of each classifier is made with a different training set. However, it is different in its topology. Boosting classifiers are serially connected as shown in Figure 2.7 part (b) [147]. They try to improve the performance of a weak classifier to become more powerful. Generally, it starts by building an ensemble of serial class weighting resampling technique. Samples that were misclassified using the first weak classifier are given higher weights. These samples are the input of the second classifier in the ensemble. Again, samples that were not correctly predicted are given higher weights than those that were correctly classified. This process is repeated until all classifiers of the ensemble are processed. Note that, the primary classifiers are weak, while the later ones are stronger. AdaBoost (Adaptive boosting) proposed by Freund and Sapphire [149] is the most popular and successful boosting method**.**

- **Stacking**

This is another MCS method that was first introduced by Wolpert [150]. Stacking consists of two levels of classifiers. Level 0, is called the base learner or classifier. They are formed from a number of classifiers like bagging and boosting, however they are trained with different learning algorithms (which is not the case in boosting and bagging). Figure 2.8 shows the structure of the stacking method. Level 1 is called the meta-classifier. Usually, training is done using leave one out or k-fold cross validation methods. The same training subset of data is used to train the base classifiers to produce a number of predictions equivalent to the number of classifiers in the level 0. These predictions along with the original class label are considered as a new set of data for another learning problem and they are the input to the meta-level classifier [151]. It actually acts as a combiner to generate a final decision on the test subset which was not used in the learning process; however, the combination is done using a learning algorithm. It is more complex and harder to investigate theoretically. As the meta-classifier is constructed with predictions of level one classifiers and it does not model the original problem variables, it is less preferred than voting, bagging and boosting [152].

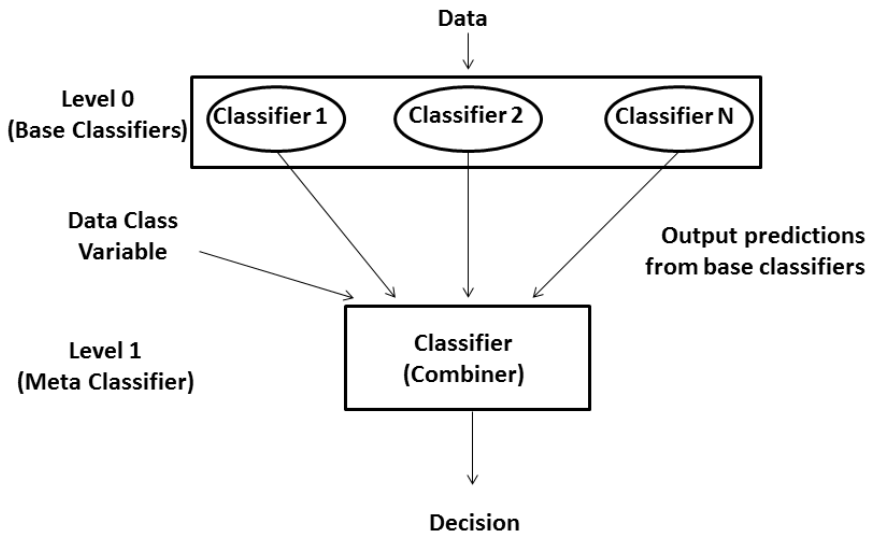


Figure 2.8 A simple structure of a stacking method which includes base classifiers passing their prediction along with the original data class variable to a meta classifeir to produce a final decision.

- **Voting**

This is the common way to combine classifiers which were either learned by the same learning algorithm such as bagging and boosting, or by different learning algorithms. Generally each classifier in the pool gives a vote for prediction. Several voting methods are used to make a final decision such as; average voting, majority voting, weighted majority voting, maximum or minimum probability voting, or product rule voting. Most MCS methods usually produce slightly different results [147, 148]; though, many researches favor the voting ensemble for its simplicity, ease of implementation and good performance [146, 153].

- **Average voting** averages the probability distributions of the predictions; the class with the higher probability is the class of the final decision.
- **Majority voting** gives a final decision of 1 if most of the classifiers in the pool indicate that the prediction is 1 and vice versa.
- **Weighted majority voting** similar to majority voting, however a weight is given to the prediction or vote of each classifier in the pool according to its performance.
- **Maximum or minimum probabilities** choose the maximum or minimum class probability of all the classifiers in the pool to decide the final decision.
- **Product rule** combines a posteriori probabilities distributions produced by each classifier in the pool by the means of a product rule which can be found in [154].

2.6.3.4 MCS for Survival Analysis

The MCS approaches were used with the statistical survival analysis approaches such as Cox's regression and accelerated failure time models. The authors in [155] have proposed a method to

produce an ensemble of Cox's model for variable selection. They used majority voting for combining prediction to classify primary biliary cirrhosis of the liver. Those in [156] proposed a boosting ensemble that uses a Cox's model based on the concordance index to perform variable selection in breast cancer patients. A boosting approach was introduced in [157, 158] for censored survival data that uses AFT for predicting colon and lung cancers recurrence.

The multiple classifier system was also used with machine learning classifiers. The Bagging or boosting ensemble method was used in [159-163] based on survival decision trees to perform variable selection and predict neck, head, colon breast, and ovarian cancer recurrence and cardiovascular disease survivability. However, Kalderstam et al [121] used a bagging method with ensemble ANNs that were trained by a genetic algorithm which used the concordance as the fitness function to predict breast cancer recurrence. The ANNs were trained directly on the concordance index to deal with survival data and censoring problem. Predictions were combined with average voting. Boosting ensemble was used in [164] based on survival trees which used the concordance index to train the trees and predict breast cancer. The stacking approach was employed [165] to predict pancreatic cancer survival. It depends on the fact that the overall prediction may be improved, if the best machine learning method performance is selected to predict each patient. The selected method can be used afterwards to predict current patient label. In order to deal with censoring, a threshold was used to label patients according to survival time. Authors used ANN, BN, Naive Bayes, decision tree, and SVM to construct the model. They also combined it with feature selection to enhance performance. Hsieh, et al [166] used BN, ANN and SVM classifiers to construct a stacking ensemble approach and predict the postoperative morbidity and explain the causal relations between preoperative variables and outcomes in patients that experienced an endovascular aortic repair surgery. The authors considered patients that did experience postoperative complication as event free; therefore, they combined the MCS model construction with standard FS techniques

The voting fusion method was used extensively in survival analysis due to reasons previously discussed. Unweighted majority voting ensemble approaches were proposed in [167-169] for prostate, esophageal cancer prognosis, and breast cancer. These methods were based on survival trees or SVM. Survival trees can deal with censoring, however for SVM censored patients were deleted and not used in constructing the models. Rathore and Agarwal [170] combined Naïve Bayes, decision tree, and multiple association rule classifiers using the majority voting ensemble approach to predict breast cancer survivability. To deal with censoring, patients were labeled according to survival time using a threshold. In [171] five classifiers were combined using majority voting to predict colon cancer survival times. The authors also labeled censored patients according to survival time using a threshold. Then, they used a resampling technique to deal with unbalance problem. Afterwards, they used standard FS methods to select the most

relevant attributes. Agrawal et al. [172] used five different classifiers based on decision tree to construct an ensemble system using average probability voting to predict lung cancer survivability, while in [173] seven classifiers were fused using maximum probability voting ensemble for breast cancer prognosis. Both methods used a threshold for labeling censored patients before entering the model.

The methods argued earlier have a main drawback in dealing with censored patients which make them unsuitable to be used with the high level of censoring included in the EVAR dataset. All of these methods handle censoring by considering them as event free using a threshold or ignoring censorship which may be applicable to low to medium censoring , but not high censoring. others methods used survival model like survival decision trees which deals directly with censoring , however they have several drawbacks such high chances of over-fitting and instability of the generated model. Therefore, they were not used in this study.

2.7 Evaluation metrics and Tests for Survival analysis

2.7.1 Log Rank Test

Usually in clinical trials, doctors need to examine the effect of a new therapy on patients and compare it with an old one. They start drawing survival curves (usually Kaplan curves to determine the probability of survival using the two therapies). Then, they employ statistical tests in order to compare between two or more survival curves. The most popular known test is the log rank test.

The log rank test is commonly used to compare two or more survival functions. It tests the null hypothesis that there is no difference in survival curves of the two groups of patients finds. This is done by determining at each time t_j (j =1, 2, 3 ….m), the number of observed dead patients (d_j) (where the event has occurred) and the total numbers of patients still at risk (n_j) preceding t_j . Afterwards, the test calculates the number of expected deaths (E), the difference between the observed and expected deaths (U_L), and the variance (V_L) between the observed and the expected deaths in the two groups of patients [174]. The steps for calculating the log-rank test is described below and can be found in [45, 175] with an illustrative example.

First, calculate the expected number of deaths in the first group of patients (E_{1j}) using Equation (2.21):

$$E_{1j} = \frac{n_{1j} d_j}{n_j}, \tag{2.21}$$

where; n_{1j} is the number of patients of the first group who are at risk of death at time interval t_j , n_j is the total number of patients at risk of death in the two groups of patients at time interval t_j , and d_j is the total number of dead patients in the two groups of patients at time interval t_j .

Second, calculate the difference between the observed and expected deaths in the first group of patients U_L using Equation (2.22):

$$U_L = \sum_{j=1}^{r}(d_{1j} - E_{1j}), \tag{2.22}$$

where; d_{1j} is the number patients in the first group that were dead until the time interval t_j .

Then, calculate the variance of death for the first group at time interval t_j (v_{1j}) using Equation (2.23):

$$v_{1j} = \frac{n_{1j}n_{2j}d_j(n_j - d_j)}{n_j^2(n_j - 1)}, \tag{2.23}$$

where; n_{1j} is the number of patients of the second group who are at risk of death at time interval t_j .

Next, the total variance in the first group (V_L) is calculated using Equation (2.24):

$$V_L = \sum_{j=1}^{r} v_{1j}. \tag{2.24}$$

Finally, a chi squared statistic (χ^2) is performed to determine if there a significant difference between the two groups using Equation (2.25):

$$\chi^2 \approx \frac{U_L^2}{V_L}. \tag{2.25}$$

A statistical coefficient known as a p-value associated with the chi-squared test is determined from the distribution function of the chi-squared random variable. In other words, it is determined from the probability that a chi-squared variable on one degree of freedom is greater than or equal the χ^2 value. The p-value indicates the probability under the assumption of the

null hypothesis. Usually, a p-value less than 0.05 indicate a significant difference between the two groups of patients.

2.7.2 Concordance Index

Concordance Index (CI) is generally used as a criterion to test discrimination in the predictions of predictive model dealing with censored survival data. CI is the probability that, given two randomly selected patients, at least one of them must have experienced the event of interest at a shorter follow up time (comparable patients); this patient must have a higher probability of the event occurrence than the other (concordant patients) [120]. In other words, the CI counts the number of concordant and comparable patients and divided them by each other. Patients are comparable if at least one of them must experience the event of interest and his/her follow up observation time is less than the other. Patients are considered concordant if they are comparable and the predicted probability of survival for the patient that experienced the event at time *t* is greater than the other. The higher CI means that the predictions of the predictive model are more concordant and discriminative. The maximum value of CI is one, while the minimum is zero. It can be calculated using the equations (2.26), (2.27), (2.28):

$$comparable = \sum_i \sum_j \pi_{comparable(i,j)}, \tag{2.26}$$

$$concordant = \sum_i \sum_j \pi_{concordant(i,j)}, \tag{2.27}$$

$$CI = \frac{concordant}{comparable} . \tag{2.28}$$

Where; $\pi_{comparable}$ is a binary variable which indicates only comparable cases, $\pi_{concordant}$ is a binary variable which indicates only concordant cases. *i* and *j* are two different patients.

2.7.3 Standard Performance Measures

Several measures are used to evaluate the performance of predictive models. In this section the most important ones are discussed.

- **Classification Accuracy** is the number of correctly classified (predicted) instances over all the number of instances used in the testing set.
- **Sensitivity (true positive rate)** is a ratio between correctly classified positive instances and the actual positive class of the instances.
- **Specificity (true negative rate)** is a ratio between correctly classified negatives instances and the actual negative class of the instances.

- **Receiver Operating Characteristics Curve (ROC)** is a curve that plots the sensitivity as a function of (1- specificity). Usually, the area under the ROC is used to evaluate the performance of the model. The greater the area indicates better performance. The maximum area that could be reached is 1. It is used for standard data type in which all class values are known.

2.8 Feature and Model Selection for Survival Data

2.8.1 Feature Selection

Feature Selection (FS) is a primary step applied to large dimension datasets. Several techniques were used to search the feature space in order to choose a lower number of attributes that can be used to build a predictive model. Their purposes are to reduce data dimension by removing redundant and irrelevant variables. FS could help in lowering the complexity of the predictive model and make it more understandable and faster. Adding more variables to the predictive model may increase its fit. However, if too many variables are added, this could lead to over-fitting. Hence, FS improves the model prediction accuracy and generalization by reducing over-fitting when choosing the optimal features [176]. Feature selection is very popular in the field of data mining when dealing with large medical datasets like microarray or gene expression data.

FS consists of four steps. which are feature generation subset, feature evaluation, stopping criteria, and feature validation [177]. The former generates subset of features according to a specific searching strategy. They are then evaluated using an evaluation metric. Feature searching usually stops when a certain criterion occurs. The variables chosen are then validated to determine their significance on prediction.

2.8.1.1 Feature Searching Strategy

Feature searching strategies search for reduced subsets of features in the feature space that influence prediction. The search strategy concerns with the way the feature subsets are formed. Feature searching strategies are classified into three categories; complete (exhaustive), heuristic, and meta-heuristic methods. Complete methods search for a best possible subset that optimizes a specific performance metric (i.e. classification accuracy). They look for optimality by exploring every feature combination in a dataset. However, this investigation may lead to an increase in the complexity and computational cost [8].

Heuristic methods such as; greedy hill climbing and best first are iterative approaches that are different to complete search as they look for the suboptimal feature subsets taking into consideration the speed of search [8]. Greedy hill climbing adds or deletes one feature iteratively to generate feature subsets. It examines these local changes to select the most

important variables. It is divided into three strategies. The first strategy is forward selection in which the search starts with an empty subset and then adds a new feature iteratively forming new subsets. The second strategy is the backward elimination search which starts with all the features available in a given dataset and deletes irrelevant features one by one. The best subset that improves a certain criteria is the one chosen. The third one is the floating stepwise strategy which alternates between the forward and backward methods to remove unimportant features. It gives another chance to eliminated variables to re-enter feature selection process once again. For this reason it was used in this thesis. The best first strategy is another iterative heuristic approach similar to hill climbing in searching feature space by making local changes to feature subset to remove unnecessary and irrelevant features. However, it allows backtracking to a more favourable previous feature subset and continues searching from it. This happens when the search path being examined offers a less promising solution.

Meta-heuristic methods such as; genetic algorithms are nature dependent. They offer good solution to feature selection in a reasonable time without investigating all feature space like exhaustive methods. The genetic algorithm is the most common meta-heuristic FS search approach. It simulates the analogy with biology theory of natural selection and genetics [8]. The group of chromosomes is known as a population and the genes as individuals. The two genetic actions that could take place are crossover and mutation. Crossover means that breeding is done by swapping over genes of each parent. However, mutation is made by randomly replacing one bit of the parent genes. Next, a fitness function is used to evaluate the development of genes after reproduction throughout the evolutionary process [178]. The advantages of using the GA are; it is a robust and parallelized technique. However, GA may have slow convergence especially with large datasets and its optimum solution may be inconsistent, as the probabilistic nature of GA could produce different optimal values so it might be ensnared in a local optimum [179]. For this reason they were not used in this study as they might increase the computational cost of feature selection process.

2.8.1.2 Feature Selection Methods

In the context of classification, feature selection techniques can be organized into four categories, depending on how they combine the FS search with the construction of the classification model: filter or ranking methods, wrapper methods, embedded, and hybrid methods. The literature has shown that most variable selection methods were done on standard supervised data. However fewer papers applied feature selection for censored survival data. Most of them focused on using Cox proportional hazard model with greedy searching strategy, penalized or shrinkage variable selection methods [9] which will be discussed later. The former methods however suffer from instability [180] which will be illustrated later in this chapter.

2.8.1.2.1 Filter or Ranking Method

Feature ranking is the easiest way to select features. It starts with calculating a score for each input variable which usually explains its relation with the output target class. Then, this score is used to sort features in a descending order. Finally, the variables with higher scores are used for constructing a predictive model. Usually, a separation threshold is known previously to select these features and the low scoring variables are removed [181]. Several techniques are used to determine ranking scores. They are categorized into two categories; correlation and informative based methods. These methods depend on either the correlation between variables and output class or the amount of information between each variable and the output.

The most simple and straight forward filter method used in survival analysis is called univariate Cox regression which calculates a Cox score for each feature used to fit the Cox's model separately. A Cox score is simply a value calculated for every variable not in the model to determine whether it should be added to the Cox proportional hazard model or not. More details of the Cox's score could be found in [182]. This score measures how well each variable predicts survival. It was used in [183-186] to predict survival of lung cancer, prognosis of gastric and cervical cancers, and tumor relapse. Tan, et al [187] used it to filter variables that are not associated with survival time of ovarian cancer, which were used later for survival prediction using SVM. In order to deal with censoring, the authors selected patients with short survival even if it uncensored and label them as unflavored. They selected other patients with long survival (censored and uncensored) and labeled them as favored. These patients were used to train an SVM model to predict the labels of the rest of patients. Afterwards, a final model is constructed with the whole data using leave one out cross validation. The Cox score was modified in [188] to enhance prediction results. Others used a Wald test or likelihood test statistic instead of a Cox score to quantify variable association with survival prediction such as in [189] to predict survival of breast cancer.

Some authors used other metrics for filtering the attributes such as; Blanco, et al in [190] who used mutual information to sort variables and select some of them according to a threshold before entering Bayesian network and Naïve Bayes classifiers. The dataset contained a low level of censoring; therefore the model was appropriate for the prognosis of survival of cirrhotic patients treated with a transjugular intrahepatic portosystemic shunt. They considered patients who survived more than 6 months as 1 and vice versa. Others [191, 192] used correlation analysis to filter variables entering Cox's model to predict ovarian and breast cancer recurrence. Neuvirth et al in [193] also used the correlation information to remove highly correlated variables. They employed the standard deviation values of the features to filter features with low standard deviation scores, and then used the Cox regression to predict patient's risk of diabetes.

The authors also did binary classification with logistic regression and KNN classifier with only uncensored patients. They also performed forward greedy FS wrapped around both models as well as univariate Cox analysis as a filter FS method. In [194] a chi-squared test value was applied to determine the association between variables and survival times of lung cancer. Then, the ten highest ranked features were selected to train an ANN in order to predict the survival times of gene expression data from multiple laboratories. In order to deal with censoring the authors considered patients that survived longer than the median survival time as low risk and vice versa. In paper [95], the authors split data into 7 partitions according to the follow-up time intervals existing in the data. They deleted censored patients; however, they overcame the problem of the small sample size for predictive model construction by repeating patients through a classification rule. They then used decision trees as a filter FS method to select prognostic variables that can predict breast cancer relapse for each of the 7 partitions. These selected features were used afterwards to construct 7 neural network models, one for each time interval. The problem of this method is both deleting censored patients and repeating patients which may lead to a biased predictive model.

The problem of most filter methods is that they are univariate which means that the scoring or evaluating function is examined for only one feature at a time. Moreover, when the selected features enter the predictive model, they could not ensure good survival prediction, as they might be highly correlated and/or collinear to each other which affect results consequently [195]. Another common disadvantage is that they ignore the interaction with the classifier (the search in the feature subset space is separated from the search in the hypothesis space). Also, that might select redundant variables as most filter methods do not consider relations between variables. Another drawback of the previously discussed methods is the way they handled censoring by deleting, repeating patients, using only uncensored cases, or considering censored patients as event free according to a specific threshold. These methods of dealing with censoring lead to highly biased predictions in cases of highly censored data like our EVAR datasets. However, the advantages of filter feature selection methods are that they can be easily scaled to very high-dimensional datasets; they are computationally simple and fast. Additionally, feature selection needs to be performed only once, and then different classifiers can be evaluated. The aim in this study is to determine which features have great influence in predicting the risk of EVAR re-intervention. As mentioned before, filter methods do not involve classifiers when selecting features. Therefore, they were not used alone in this study. Factor analysis feature reduction method which may be considered as a filter method was used in this study as an initial step to remove unimportant variables. This method does not depend on the class value to reduce variables so it may be applicable to be used with censored data. In order to deal with censoring a new method was proposed in Chapter 3 which was then embedded in a wrapper feature

selection procedure which involves the classification process into the feature selection procedure in Chapters 4 and 5. Also, a filter ranking method with used with wrapper feature selection process to further reduce features selected in this thesis in Chapter 5.

2.8.1.2.2 Wrapper Method

This method wraps the model hypothesis search within the feature subset search. In this setup, the whole feature space is searched and various subsets of features are generated and evaluated. The evaluation of a specific subset of features is obtained by training and testing a specific classification model. The working procedure of wrappers is the same as that of the filters except that the measurement stage is replaced by a learning algorithm. This is the main reason that the wrappers always perform slowly. However, wrapper method could achieve better feature selection results in most cases. FS using wrapper method stops when the results start to get worse or the number of features reaches a predefined threshold.

Many authors use a greedy search strategy which is wrapped around Cox' regression model for survival analysis. Authors in [196-198] used stepwise variable selection with Cox's model to predict prostate and lung cancer survival and relapse, and overall survival of metastatic renal cell carcinoma and use the p-value of the likelihood function for selection. In [199] both forward selection and backward elimination were used with a Cox's model to remove unimportant variables and predict sudden cardiac death in patients with heart failure. A Backward elimination strategy was merged with a multivariate Cox's model in [200] and the Wald test statistic was used as a selection criterion. However, in [201] the p-value of the likelihood is used as a criterion for selecting the risk factor for late renal allograft failure. Authors in [202] employed forward FS with Cox's model to predict the risk of recurrent venous thromboembolism. While authors in [12] employed a genetic algorithm for selecting features based on Naïve Bayes classifier prediction as its fitness function to predict colorectal cancer survivability. To deal with the censoring issue, the authors divided patients into two groups. The first group contains patients who died with survival time less than 30 months and labeled as 1. The second group contains patients that have survived with survival time greater than 70 months and labeled as 0. Khosla, et al [203] proposed a new approach called conservative mean which is wrapped around SVM classifier to predict the risk of stroke occurrence. The authors considered patients that did not experience a stroke for 5 years as event free labeled as 0. They considered that the number patients that experienced the stroke is small and non-homogenous, therefore prediction results will be affected by how the training and testing sets are sampled. To overcome this, they used a conservative mean criteria for feature selection which is the difference between the standard deviation and the mean of the AUC of each feature when enter the SVM predictive model. Partial logistic artificial neural network (PLANN) was used with

backward feature elimination [117] to select the most important features and predict liver survivability after colorectal cancer. PLANN handle censoring by dividing observation time into *n* intervals and repeating patients to these intervals. The repeating process leads to imbalanced and biased predictive models especially with censoring of high level [13]. It also increase the complexity and training time of these models and may increase the noise level existing in the datasets. Authors in [187] applied Cox's model to perform FS on censored data before entering a SVM classifier to perform prediction. Other authors used wrapper FS methods[12, 193] with Bayes classifiers and K-nearest classifier. While in [194], a chi-square test was used to measure degree of association between variables and observation time to choose the most relevant variables. These variables are then used to construct an ANN.

Advantages of wrapper approaches include the interaction between feature subset search and model selection, and the ability to take into account feature dependencies. They evaluate a subset of features which enables figuring out interactions between variables. Therefore, they were used in this thesis to select the features that influence the prediction of the risk of EVAR re-intervention. The limitation of methods that has been shown in the literature is that most of them are based on Cox's model. As mention before, Cox's model is incapable of identifying complex and non-linear relations between data which are included in most medical datasets. Therefore, it is not recommended to be used in our EVAR data. Another limitation is the way they handled censoring which was by deleting, ignoring, using only uncensored patients, or considering censored patients as event free. These methods of handling censoring are not appropriate to deal with the high level of censoring which exists in our EVAR datasets. Another drawback is that they have higher computational cost than filter techniques, especially if building the classifier is computationally expensive. Hence, in this study, they were merged with factor analysis and feature ranking procedures which may be considered as filter methods to select a reduced number of features the have better prediction on aortic complications.

2.8.1.2.3 Embedded Method

The search for an optimal subset of features is built into the classifier construction, and can be seen as a search in the combined space of feature subsets and hypotheses. Just like wrapper approaches, embedded approaches are specific to a given learning algorithm. Embedded methods have the advantage that they include the interaction with the classification model, while at the same time being far less computationally intensive compared to wrapper methods. Authors in [204] embedded variable selection in Bayesian network construction to perform feature selection with several types of censored data. They extended a technique called Max-Min Parents and children algorithm to survival analysis. Pang, et al [205] used feature elimination search strategy to embed FS in a random survival forest classifier to reduce the high dimension of three censored survival gene expression data. Random survival forest classifier is

an extension to standard random forest classifier in which the output is survival time with a censor indicator. However, the main drawback of tree based methods are instability, variable selection bias and over-fitting [167]. Van Belle, et al [206] embeds FS in least square-SVM construction to predict breast cancer survivability by changing the loss function so that the maximal variation for each variable is reduced, leading to models have with dependence on the features. This will lead to non-relevant features having a very small maximal variation, while relevant variables will still have large maximal variations. Note that, the least absolute shrinkage and selector operator can be considered as an embedded method, as the variable selection process is involved in the regression model construction.

2.8.1.2.4 Hybrid Method

Sometimes, two classes of the previously mentioned methods are combined together to obtain the advantages of both of them. This is called hybrid FS. It usually outperforms the other three FS methods. Therefore, they were adapted in this study. The Hybrid FS technique for survival data was used in [207] to filter features not related to survival according to the concordance index metric. Afterwards, feature elimination FS is embedded in support vector regression for censored data (SVRc) to predict overall survival in patients with lung cancer and select the model with the highest concordance index. This is done by eliminating features which have low impact in the model measured by multiplying the weight of each feature by the square root of the standard deviation from variable values for all patients. SVRc is a modification on standard SVR in which the loss function is modified to be able to deal with censoring. Authors in [208] used variance information to filter variables which were then followed by univariate Cox analysis to filter out feature related to survival. Afterwards, a Cox's model was used with greedy backward elimination to select the best variables that predict breast cancer survival time according to concordance index. Stebbins, et al [209] used stepwise, forward, and backward searching strategies wrapped around Cox's model to select the best features that predict mortality in Acute ST-Segment Elevation Myocardial Infarction. They then applied the Wald test to rank the selected features according to their significance. Choi, et al [197] used cross validation to produce several subsets of features using Cox's model, then sorted them according to their frequency distribution to predict prostate cancer relapse.

Most of these methods are based on Cox's model and as mentioned before machine learning techniques are preferred due to their ability to detect complex and non-linear relations between data that improve prediction. Therefore, methods based on Cox's model were not recommended to be used with the EVAR data. However hybrid methods are used in this study as they usually have higher performance than other methods as mentioned earlier. For this reason, a hybrid feature selection approach was proposed in this study. The proposed method uses factor analysis and feature ranking methods with wrapper feature selection approach based on machine learning

classifiers to select variables that have impact on the prediction of the risk of EVAR re-intervention.

2.8.2 Model Selection

Model selection is popular in the field of statistical analysis. It aims to find a model or models with either the best parameter combinations or the best parameter and variable combinations from a number of candidate models. They may be considered as a variable selection technique when the purpose is to find the model which has the optimal explanatory variables. In this subsection, our aim is to discuss model selection methods used for feature selection and refer to it as model-feature selection methods.

2.8.2.1 Model Stability

Model Stability is an important issue in model selection process. Stability refers to how different models produced from the model selection algorithm are close to each other when training misclassification cost occurs [210]. In other words, the models are stable if each one has a slight or no change in prediction error when some data examples used for training are replaced by other cases. Usually, unstable algorithms can lead to high variance. Furthermore, in model-variable selection, eliminating important variables from the model selected lead to under-fitting the data. It may also bias the prediction results. Adding additional unnecessary features leads to over-fitting and increases variance.

2.8.2.2 Bias-Variance Dilemma

In machine learning, the classification error (C_e) produced from the predictive model is decomposed into three components, namely bias and variance and irreducible error (e) as shown in Equation (2.29) which shows the expected square classification error. The irreducible error is the noise error that cannot be reduced by any model. Hence, in order to lower the classification, both bias and variance must be reduced. Error due to bias can be considered as the difference between the average predicted value and the correct one, i.e. if the model is trained with different training sets, how close the prediction is to the original true one at each training case. Error due to bias is determined using Equation (2.30). Low bias means that the prediction is very close to the original on average and vice versa. However, variance error is the one that indicates how far each predicted example is from the same example when predicted from a model trained with other training set. High variance means that the prediction is widely spread and changing the training set highly affects the predictive model. Error due to variance is calculated using Equation (2.31).

$$C_e = E(Y - \hat{f}(x))^2 = Bias^2 + Variance + e, \tag{2.29}$$

where; E refers to as expected, $\hat{f}(x)$ is the estimated classification function that approximates the true function $Y=f(x)$, Y is the original output class of variable x, and e is called the irreducible error which is the noise error that cannot be reduced by any model.

$$Bias = E(\hat{f}(x)) - f(x). \tag{2.30}$$

$$Variance = E[(\hat{f}(x) - E[\hat{f}(x)])^2]. \tag{2.31}$$

Usually, models with low bias are complex and have high variance. Their prediction varies widely when changing the training examples of the model. However, simple models have low variance, but the bias is large. In order to produce a stable algorithm a tradeoff between variance and bias must be made [211]. Additionally, as mentioned before model selection can be used for choosing the optimal explanatory feature subset. Adding more features leads to a decrease in bias, but it increases variance. So, the tradeoff between the two is important. One should choose the right model that balance both variance and bias [176].

2.8.2.3 Model-Feature Selection Methods

Several techniques have been proposed in order to choose the optimal model. They can be categorized into; methods that maximize the likelihood subject to some penalty, shrinkage methods such as LASSO, cross validation methods, Bayesian model selection, and Bayesian model averaging methods. Note that, all model selection techniques can be used with cross validation.

2.8.2.3.1 Penalized Methods

As mentioned before, increasing the number of parameters or variables improves fitting. However, increasing it too much may lead to over-fitting. Penalized methods overcome this problem by introducing a penalty for the number of parameters or variables in a given model. Akaike (AIC), Bayesian (BIC) information criteria are the most well-known techniques [211]. The former was introduced by Akaike in 1977. It measures the quality of each candidate model. It is based on minimizing the Kullback Leibler distance which is the distance between the true and candidate models. AIC takes into consideration the number of free parameters in the candidate model and the goodness of their fit. The chosen model is the one that minimizes the AIC which is equivalent to the lowest distance to the true model [212]. It can be calculated using (2.25)

$$AIC = -2 \cdot \ln(L) + 2 \cdot K \ , \tag{2.25}$$

where; L is the maximum likelihood of the model given the data and K is the number of parameters in a given model.

The Bayesian information criterion (BIC) is another well-known metric for model selection. It was introduced by Schwarz [213] in 1978. BIC also measures the quality of each candidate model. The penalty term in BIC is greater than AIC. BIC takes into account the size of data for each model which is not the case for AIC. Therefore, some researchers prefer it when they deal with models with small or different sample sizes. Again, the model which minimizes BIC is the one chosen. It is calculated using (2.26)

$$BIC = -2 \cdot \ln(L) + 2 \cdot K \cdot \ln(n), \tag{2.26}$$

where; L is the maximum likelihood of the model given the data and K is the number of parameters in a given model, and n is the number of observations.

Most authors that use AIC and BIC for variable selection in survival data have combined it with Cox's model, such as; [214] used them for variable selection to predict liver disease, then modified the classical BIC approach. Authors in [215] employed them both as stopping criteria for selecting the optimal variables that detects ozone effects on school children's lung growth. Xu, et al [216] used AIC for variable selection for lung cancer prediction. Authors in [217] modified the AIC to be used and applied it on two cancer datasets. While, those in [218] extended the BIC in the Cox model to deal with very high dimensional data and applied it to select the variables of an microarray data. Hofner, et al. [219] used them to model the Großhadern severe sepsis data. These variable selections methods are based on Cox's model as mentioned before, Cox's model is the most common survival method used in medical applications. However, the aim in this thesis is to use machine learning techniques to construct a predictive model as they preferred for the various reasons discussed earlier. Therefore, the proposed feature selection methods in this thesis were compared with Cox's model based on AIC and BIC as a selection criterion.

Both AIC and BIC have been used as evaluation metrics for the wrapper variable selection methods. These criteria may not be practical with complex model situations where the number of variables is extremely large as it may overestimate the true model. Therefore, a new penalized method based on the shrinkage of the absolute regression coefficients was proposed as an alternative [220].

2.8.2.3.2 Least Absolute Shrinkage and Selection Operator (LASSO) Method

LASSO is a very popular variable selection method. It was introduced by Tibshirani R. in 1996 [221] to be used with regression models and was then extended for Cox's survival model by Tibshirani R. in 1997 [222]. The advantage of LASSO is that it produces stable models with high predictive accuracy. This is done by the regularization parameter that shrinks the coefficients of unimportant variables (regressors) towards zero by putting a penalty on their absolute values. Therefore, the produced models will focus only on the explanatory variables that improve their total predictive accuracy and preventing over-fitting that may occur due to collinearity of the variables. Additionally, LASSO has low computational cost. It is calculated using (2.27)

$$Lasso = \arg_{\beta} \min\left[\left\|y - X\hat{\beta}\right\|^2 + \lambda\left\|\hat{\beta}\right\|_1\right], \tag{2.27}$$

where; y is the response variable (output target), X is a matrix with predictive variables(predictors) and observations, β is a vector of their coefficients, and λ is the amount of shrinkage or prenatally on β. As it increases the value of shrinkage increases and vice versa. The term $\left\|\hat{\beta}\right\|_1$ when $\lambda \rangle 0$ is the L_1 norm.

Many authors have used it with Cox regression model. Lee et al, Kammers et al, and Liu and Jiang [223-225] used it to predict survival of breast and lung cancer and other types of cancer respectively. Liu, et al [226] used it to select a smaller set of genes, then proposed a correlation metric to further reduce the number of genes to predict survival of lymphoma patients. It was employed in [227] to select the prognostic variables that predict penile cancer and heart valve failure. It was used in [228] to select the risk factors for pancreatic cancer survival. However, authors of [229] applied it to Cox's and ART models and those of [230, 231] used it only with AFT for breast cancer prognosis. Several extensions were made by authors such as; [232] modified the LASSO technique by adapting weights of L_1 penalty on regression coefficients and referred to it as adaptive LASSO which was used later by [220, 233] to predict and congestive heart failure and lung cancer survival with variables containing missing values. Unlike LASSO, it does not place the same penalty for all coefficients; however, it adapts the weights of these penalties. Large weights are assigned to small coefficients and vice versa. Authors in [234, 235] proposed using LASSO with Cox's model learned through a gradient LASSO algorithm for faster convergence to predict survival of breast cancer. The authors proposed gradient LASSO to solve the problem of collinearity involved with Cox's model when used with high dimensional data. It also avoids the computational cost existed in previous shrinkage methods

used with high dimensional data. LASSO was also extended in [236] by Bergersen who named it weighted LASSO. It used external information on the covariates to select the appropriate penalty weight for regression coefficients and guide variable selection for cervix and head and neck cancers. In this thesis, the Cox-LASSO method was used for comparison with the feature selection approaches based in machine learning classifiers that are proposed in this thesis. As mentioned before, Cox's model is extensively used in medical survival data, therefore it was chosen for the comparison purpose.

2.8.2.3.3 Bayesian Model Selection

In Bayesian model selection, a prior is given for each candidate model. Then the posterior probability of each model given the data is calculated. The model with highest posterior probability is the one chosen. It is calculated using Bayesian theory as shown in Equation (2.32):

$$P(M_i \setminus D) = \frac{P(D \setminus M_i).P(M_i)}{\sum_j P(D \setminus M_j).P(M_j)}, \tag{2.32}$$

where; $P(M_i \setminus D)$ the posterior probability of the candidate model M_i given the data D, and $P(M_i)$ and $P(M_j)$ are the prior probabilities of model M_i and M_j respectively. $P(D \setminus M_i)$ and $P(D \setminus M_j)$ are the probabilities of the data given models M_i and M_j including their free parameters.

If the priors $P(M_i)$ and $P(M_j)$ are unknown, they can be given equal probabilities, so Equation (2.32) can be simplified into Equation (2.33) and again the model with the highest posterior probability is the one chosen.

$$P(M_i \setminus D) \approx \frac{P(D \setminus M_i)}{\sum_j P(D \setminus M_j)}. \tag{2.33}$$

The main advantage of using Bayesian model selection is that it takes into consideration the fitting improvement of the model and not reaching over-fitting. Hence, adding more variables or parameters will not improve the fit. However, it will lower the posterior probability of that model given the data [176]. This method was used for survival analysis in [237-239] combined with a Cox's model to predict lung cancer survivability. Authors in [240, 241] used it with an AFT model for predicting lymphoma and breast cancer patients survival times. These methods are based on either Cox's or AFT survival models which cannot figure out complex relations in a data. Moreover, a prior should be known in order to perform model selection. Hence, these methods were not applied to EVAR data.

2.8.2.3.4 Model Averaging

Model averaging is another model selection technique. As mentioned before, the purpose of model selection is to choose the best model; however, sometimes averaging results of several models may be better than choosing only one single model. This is due to the fact that the averaging method considers the uncertainty of estimation or variables included in the candidate models [242]. Bayesian model the averaging sums all the weighted average of some or all models used to find the predictions. The weighted term refers to posterior probabilities of each model given the data $P(M_i \backslash D)$. Suppose a data D with output target variable O and the candidate models of interest M_i. The posterior predictive probability distribution of O is determined using Equation (2.34):

$$P(O \setminus D) = \sum_{i \in N} P(O \setminus D, M_i).P(M_i \setminus D), \tag{2.34}$$

where; $P(O \setminus D, M_i)$ is the marginal posterior probability distribution of the O given the data and model M_i with all parameters integrated out.

Again, the main advantage of model averaging is that when more than one model has high probability distributions, model averaging take into consideration the uncertainty of estimation generated from these model. This is not case in the Bayesian model selection which chooses only one single model. It was extended by [243, 244] to be used for survival analysis using an iterative approach with a Cox model to predict breast cancer, leukemia and lymphoma survival times. Later, a Bayesian model average approach was proposed in [245] for variable selection of the most prognostic factors related to breast cancer survival time. Authors of [246] employed a model averaging approach combined with Cox model to predict 5 year survival after major surgery.

Since model averaging creates a final single model by average multiple models with high posterior probabilities, it could not be easily interpreted [176]. Also, survival model averaging methods are based on Cox's model which has drawbacks that were previously discussed. For these reasons, they were not used in this study.

2.9 Conclusion

In this chapter, survival analysis techniques used in clinical trials were discussed. They included the traditional survival statistical methods and machine learning approaches. Traditional survival statistical methods construct a predictive model that can deal directly with censoring which is the unique characteristic of survival data. Machine learning techniques such as ANN and BN are preferred over traditional statistical methods due to various advantages. However, they cannot be used directly with survival data due the presence of censoring. For this reason,

the censoring problem should be solved. This chapter discussed different methods to handle censoring. These methods have several drawbacks and they cannot be used with high levels of censored data. Therefore, they cannot be used directly to the EVAR datasets. For this reason, a modified approach will be introduced in Chapter 3 to deal with the highly censored data and construct a predictive model using machine learning techniques.

This chapter also explains the importance of FS in the medical field and discusses different approaches used for medical survival data. The literature showed that most of these methods are based on Cox's proportional hazard model or AFT model. However, as mentioned before, machine learning techniques are favoured. Therefore, a FS approach will be proposed in Chapter 4 that uses individual ANN classifiers to construct a reduced predictive model. The approach embeds the proposed uncensoring method with FS process to deal with high censoring to construct a predictive model capable of predicting the risk of re-intervention after EVAR surgery.

Additionally, Chapter 2 gave a brief introduction to the concept of multiple classifier systems and how several classifiers can be combined using several approaches. It also discussed how they can improve prediction results of individual classifiers and why they are preferred. Thus, in Chapter 5 a hybrid FS method is proposed which uses three popular classifiers to construct a predictive model capable of predicting the risk of re-intervention after EVAR surgery.

Chapter 3: Handling Censoring to Predict the Risk of Endovascular Aortic Aneurysm Repair Re-intervention

3.1 Introduction

In this chapter, a new modified approach for dealing with the censoring issue will be proposed and compared with the technique that was previously proposed by Štajduhar and Dalbelo-Bašić [6]. Two datasets collected from two vascular Centers are used for the model construction and validation. The dataset from Centre 1 is divided into low risk, high risk, and censored groups. Two Bayesian networks are constructed with the low and high risk groups. The likelihood information is then used to determine to which Bayesian network's intrinsic distribution each censored patient belongs to. Afterwards, a three layer MLP-ANN will be used to classify patients into high and low risk groups and its performance will be tested and validated using Centre 2 dataset. Evaluation metrics like the sensitivity, specificity, area under ROC curve, concordance index, and the p-value of the log rank test were employed to test the significance of the results.

3.2 Dealing with censoring

Censoring is the unique feature that differentiates standard data from survival data. It is also the main cause why standard machine learning technique cannot be used directly with survival data types. This is due to the fact that censored patients cannot be all considered as real zero targets (which means that the event of interest definitely has not happened) which could lead the predictive model to be biased toward zero specially when the level of censoring is high. This phenomenon will be illustrated and proven in section 3.6. As mentioned in the previous chapter, the literature had shown there are different scenarios to handle censoring. The first scenario is deleting the censored examples from the dataset or considering them as event free (the event will certainly not occur so the target output will be zero). Though, this will result in a highly biased predictive model which appears clearly when the level of censoring is high (the ratio of the patients that experienced the event to the total number of patients) [26]. Another solution is to construct the predictive model using only uncensored patients; however this could not work as well with a highly censored data as the number of examples used to build the model will be small, leading to inaccurate prediction.

Another situation is to weight censored patients according to KM survival probabilities. Using KM weighting will work well with low to medium censoring, while for high censoring it will give overoptimistic estimated survival probabilities [6]. Another suggestion to deal with

censored data is to split the observation time into time intervals and repeating patients to all time intervals [7]. However, this replication results in an unscaled, complex, and biased prediction [13]. It may also increase the noise in the case of noisy datasets. Hence, a new approach to deal with the high censoring of the EVAR datasets was proposed in this thesis which is a modification on the method in [6]. The new approach is referred to as O & MA, while the other one as S & DB. The new method (O & MA) does not depend on deletion, repetition, and weighting of examples. It also employs all patients available to construct the predictive model after uncensoring censored patients without considering them as event free patients (with zero target). By doing so, it will bias the predictive model toward the zero targets; this phenomenon will also be shown in section 3.6 which shows that predictions will be biased towards the zero class. Therefore, it overcomes all drawbacks of other methods.

3.3 Challenges of the Proposed Uncensoring Approach

The literature has shown that most of the work done in survival analysis was dealing with low to medium censoring level. The problem is more complicated when the level of censoring increases. Moreover, most papers used only one dataset for constructing and validating the predictive model using k-fold cross validation, leave one out, and bootstrapping methods. However, a few of them used multicenter datasets for building and testing their models. The benefits of using multicenter datasets relies on a larger number of research groups and centers, separate areas, locations or maybe countries engaged, the variability of population available in the datasets, and the ability to test the performance on cross center based models which will support and validate the results of the predictive model [247]. Using machine learning techniques instead of statistical survival analysis methods is preferred. As, they deal with complex relations between covariates which will improve predictions accordingly. Also, they are able to tackle nonlinearities in a dataset which is not the case with survival statistical approaches such as; Kaplan-Meier, Cox's model, and accelerated failure time model [89]. Therefore, a new technique was proposed to deal with high levels of censoring using two separate datasets collected from two centers located distinctly in the United Kingdom. These multicenter datasets were used to construct and validate survival models built using machine learning techniques such as ANN and BN.

3.4 Uncensoring Approach

3.4.1 S & DB Uncensoring Approach

In this uncensoring technique, the authors split data into patients that experienced the EVAR re-intervention (REINT) referred to as high risk and those who did not. Patients that did not do such an operation were used to construct a Bayesian network called a censored Bayesian

network. It includes patients with short censoring time and those with longer censoring time exceeding 5 years, who have less probability to undergo a REINT. High risk patients were used to construct another network called a high risk (re-intervened) Bayesian network. Due to the high censoring level located in the data, the constructed censored Bayesian network is a mixture of both the internal distribution of the low risk (lower chance of undergoing a REINT) and high risk patients instead of the low risk patients alone. The likelihood information which is weighted using KM survival probability estimator method was then used to uncensor data. A three layer MLP network was then constructed using the uncensored data. The S & DB technique slightly enhanced the prediction accuracy of both datasets. But, when it was used for cross-center prediction, the neural network constructed with Center 1 data was incapable of predicting the REINT on Centre 2 data which will be clear in the section 3.6. For this reason, to overcome this, the new modified approach (O & Ma) was proposed to deal with the high censoring level located in the datasets. It was able to predict the risk of surgical REINT on patients after 5 year from experiencing an EVAR operation and categorize them as high and low risk patients.

3.4.2 O & Ma Proposed Uncensoring Approach

The main difference between S & DB and O & Ma methods are that the former used both patients with short and long censoring time (less and greater than 5 years) to construct the so called censored Bayesian networks which was employed to uncensor the data using a likelihood information weighted by KM survival probability. However, the latter used only patients with long censoring time exceeding the 5 years (low risk group) to build a Bayesian network called a low risk Bayesian network. This network is used later for the uncensoring process to differentiate between the short censoring times patients into low- and high-risk groups using the likelihood information not weighted using KM. As mentioned before using the KM weighting process may lead to over optimized prediction in case of highly censored data which indicates that most of the patients will be predicted to have lower chances to experience the event which is the REINT in our case.

First of all, in the O & Ma approach, the KM curves of both centers will be plotted separately as shown in Figure 3.1. Afterwards, the time variable is employed to divide patients into three groups; high risk, low risk, and censored. Note that the total number of patients after removing patients with missing values is 457 and 286 for Centers 1 and 2 respectively. The high risk group is the one which has patients that underwent an EVAR surgery within 5 years of observation (re-intervened or high risk patients). The number of patients in this group is 42 and 22 patients for Centers 1 and 2. The low risk group belongs to patients that did not require the operation for a time greater than or equal 5 years (low chance to do the re-intervention or low risk patients. The number of patients in this group is 62 and 20 patients for Center 1 and Center 2 . While, the third group is the rest of patients including all patients who were either deceased

or left the study within 5 years. The number of patients in this group is 353 and 244 for Centers 1 and 2 respectively.

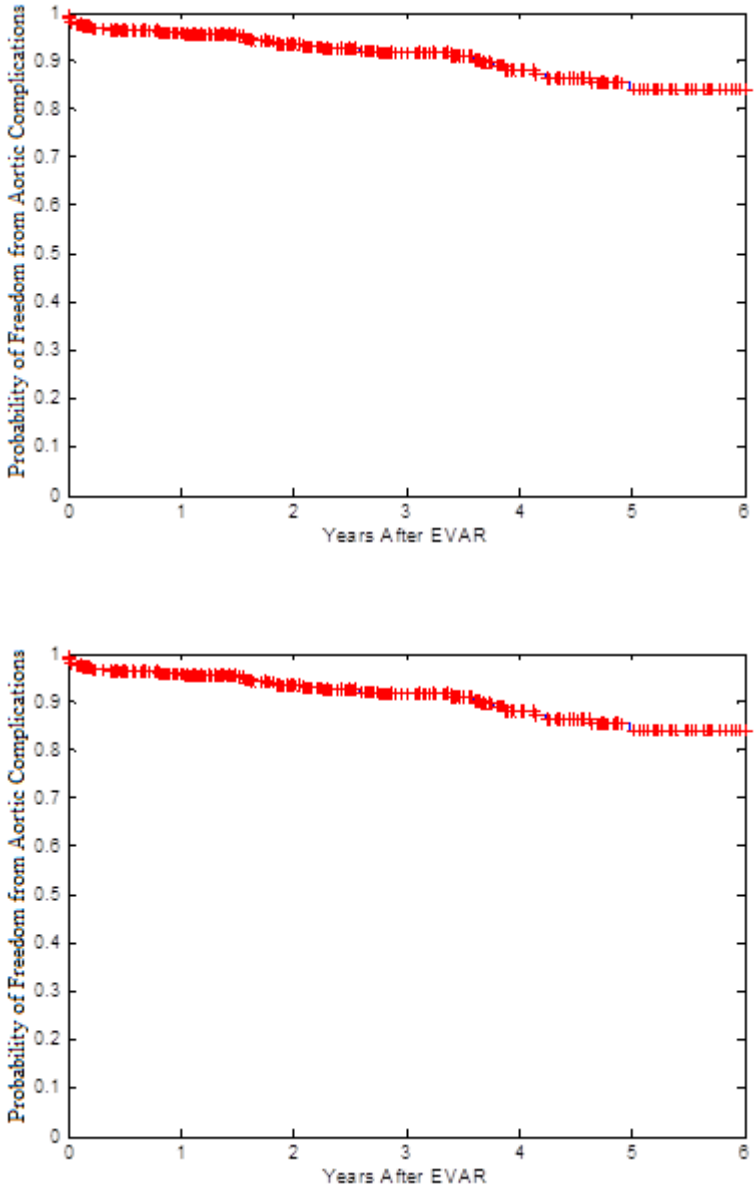


Figure 3.1 The Kaplan Meier curves for (upper) Center 1 data , (lower) Center 2 data.

In the O & Ma approach, the whole data (including all 45 features) was discretized. Then, low and high risks groups after removing the time variable were used to construct two Bayesian networks called B^{low} and B^{high} respectively as shown in Figures 3.2 and 3.3. Each variable V_i represents a node in this network that may be connected to a higher parent node (π) and lower child node. They are directed acyclic graph (DAG) networks given a symbol ξ meaning that nodes are connected in only one direction from parent node to children nodes. Details about Bayesian networks can be found in the Appendix A.

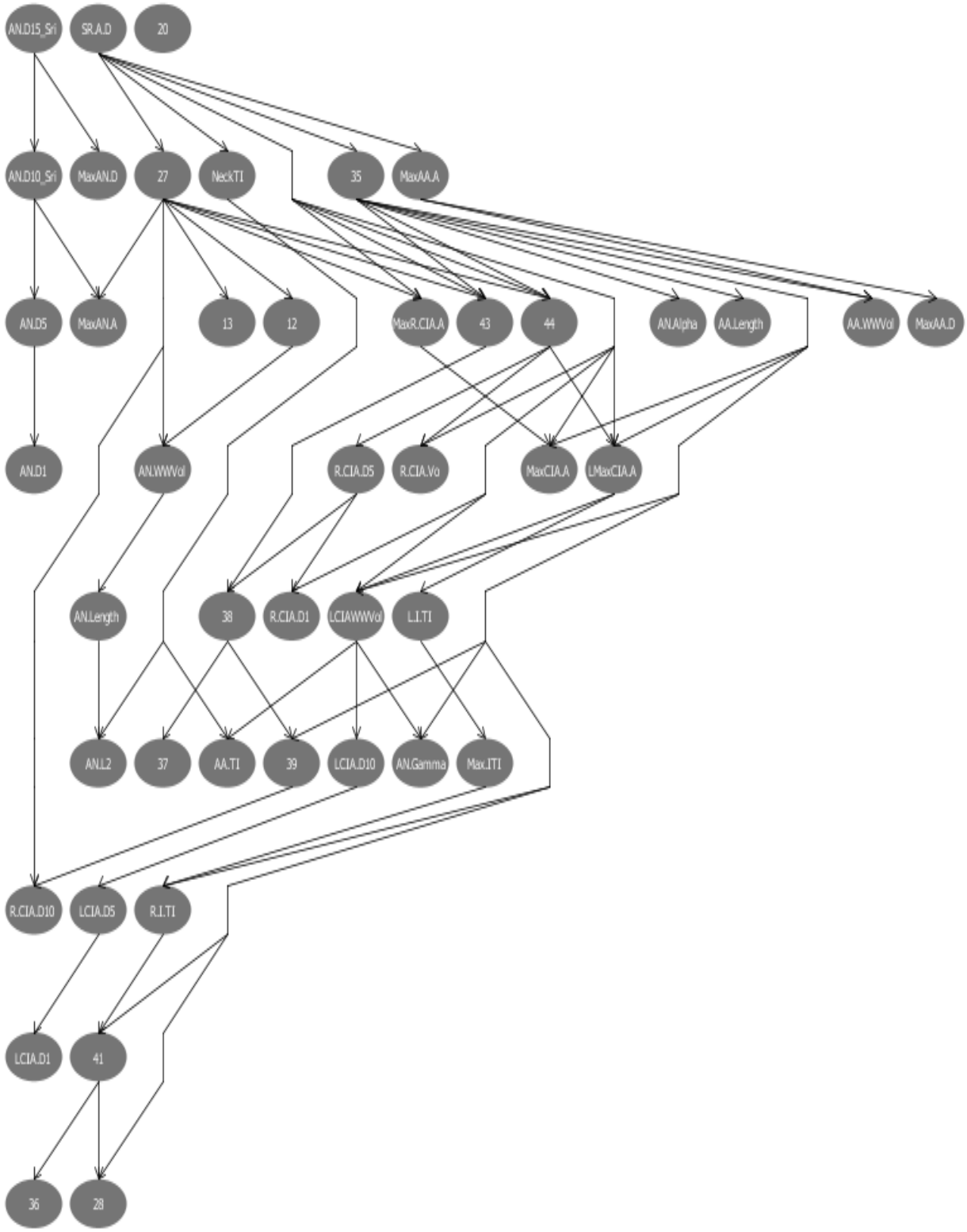


Figure 3.2 The structure of the low risk Bayesian networks constructed using patients of the low risk group. Note that some nodes have numbers which refer to the variable order in the dataset.

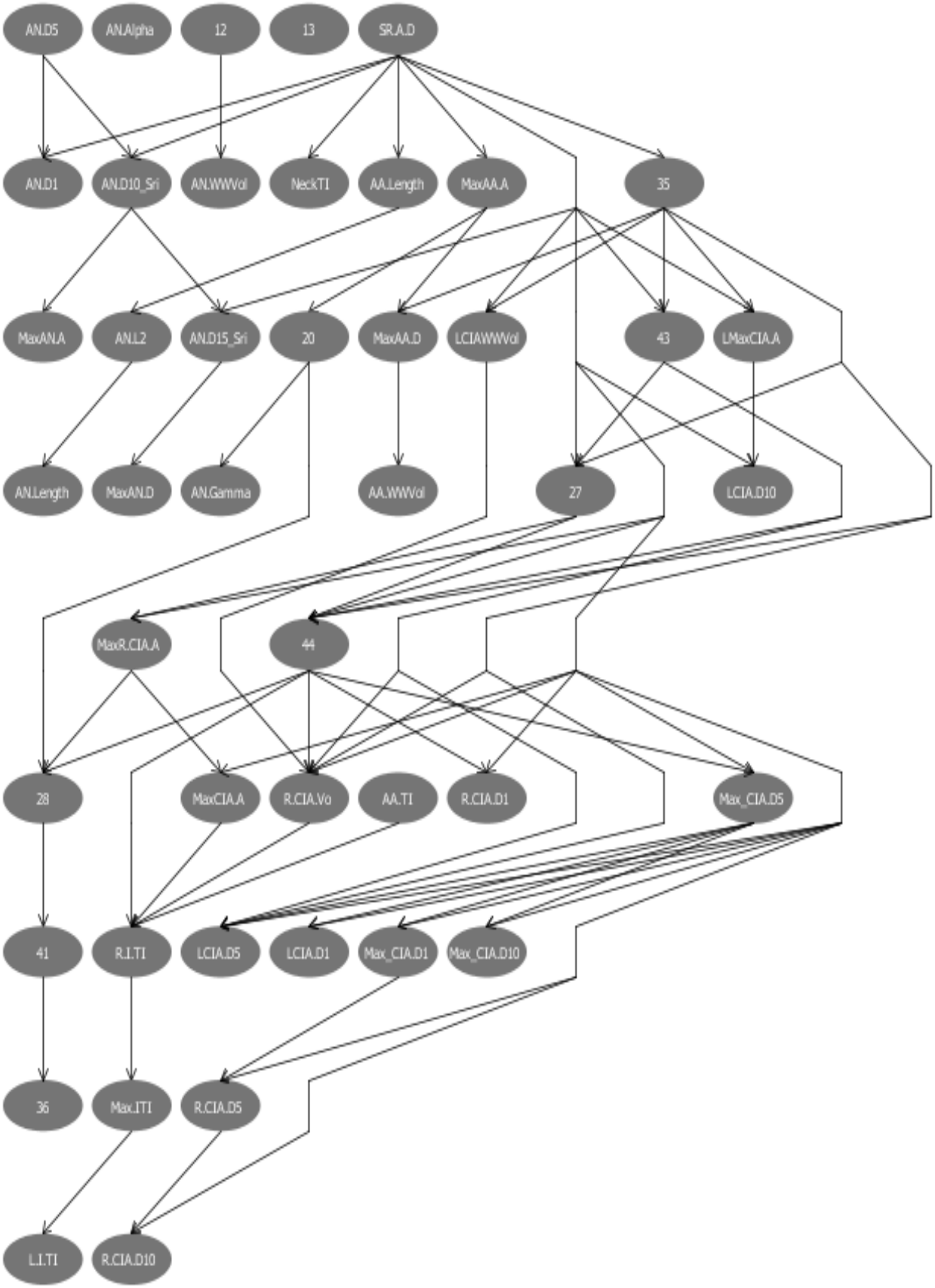


Figure 3.3 The structure of the high risk Bayesian networks constructed using patients of high risk group. Note that some nodes have numbers which refer to the variable order in the dataset.

The uncensoring process is employed to uncensor the censored patients group, where every censored one is compared with the internal structure p of each network p^{high} and p^{low} correspondingly. This structure shows relation between variables. The likelihood $\ell(x_c / p)$ that each censored patient (x_c) belongs to each network is calculated using Equations (3.1) and (3.2) to decide to which group they belong.

$$\hat{\ell}(x_c / p^{high}) = \ell(x_c / B^{high}) = P(x_c / \xi^{high}, p^{high}) = \prod_{i=1}^{n} P^{high}(V_i / \pi(V_i)). \tag{3.1}$$

$$\hat{\ell}(x_c / p^{low}) = \ell(x_c / B^{low}) = P(x_c / \xi^{low}, p^{low}) = \prod_{i=1}^{n} P^{low}(V_i / \pi(V_i)), \tag{3.2}$$

where; $\pi(V_i)$ is the parent node to variable $V_{i,}$, $P^{high}(V_i / \pi(V_i)$, and $P^{low}(V_i / \pi(V_i)$ are the posterior probability of a variable $V_{i,}$ given its parents nodes for high and low Bayesians networks respectively.

Afterwards, the posterior probability that outcome predictions that patients belong to each network given that they are censored (x_c) $P(O / x_c)$ in Equation (3.5) is calculated using Equations (3.3) and (3.4).

$$P(O^{high} / x_c) = \hat{P}(O^{high}) * \frac{\hat{\ell}\ (x_c / p^{high})}{P(x_c)}, \tag{3.3}$$

$$P(O^{low} / x_c) = \hat{P}(O^{low}) * \frac{\hat{\ell}\ (x_c / p^{low})}{P(x_c)}, \tag{3.4}$$

$$P(O / x_c) = P(O^{high} / x_c) + P(O^{high} / x_c) = \frac{\hat{P}(O^{high}) * \hat{\ell}(x_c / p^{high}) + \hat{P}(O^{low}) * \hat{\ell}(x_c / p^{low})}{P(x_c)}, \tag{3.5}$$

Equation (3.5) is then normalized to ignore the effect of probability of a censored instance $P(x_c)$ by dividing Equation (3.5) by $P(O / x_c) * P(x_c)$ to get equation (3.6).

$$P(O^{high} / x_c) + P(O^{low} / x_c) = 1. \tag{3.6}$$

Lastly, a threshold is used to decide which risk group each censored patient belongs to. It is called a censoring correction threshold (P_{Th}). If $P(O^{high} / x_c)$ is greater than P_{Th}, then the patients is consider as high risk to experience a re-intervention and vice versa.

3.5 Artificial Neural Network

The multilayer perceptron (MLP) is the most commonly used ANN in medical applications. The MLP-ANN consists of number of neurons referred as nodes connected together by weights. These nodes are gathered together to form *N* layers consisting of; one input, one output, and one or more hidden layers. More details about MLP-ANN could be found in [248]. In this chapter, ANN models were constructed to predict the risk of EVAR re-intervention after 5 years of the initial EVAR surgery. Before presenting the constructed models, a brief introduction of ANN will be introduced.

3.5.1 Training Multilayer Perceptron Neural Network

Training of an ANN means that the weights connecting the network neurons are updated to minimize the error between its prediction and the true target value. Several algorithms are available for learning ANN. Gradient descend back propagation (BP) optimization method is the most commonly used method for training the ANN. It is simple and usually provides efficient performance in many applications [248]. It depends on calculating the first derivative (gradient) of the loss function relative to the network weights, and then altering these weights in a descending steepest path (the most negative of the gradient) which is the direction in which the error function decreases speedily. In other words, this method generally express significant enhancement in the speed at which the minimum of the error function is found [249] . Although, this direction allows the performance function to decrease rapidly, it does not guarantee the fastest convergence. As, the step size used in this method is usually constant and small to be sufficient to steep gradients to avoid rattling out the minima [250, 251]. Therefore, training will be slow in places where gradient is gentle. For that reason, extension were made on steepest descend to accelerate training of ANN. Examples of these methods are the, conjugate gradient and gradient descent with momentum.

Gradient descent with momentum is a modification done to gradient descent method to speed up the learning process. Momentum takes into consideration fraction of weights that were updated in the previous learning step to modify the current weight based on the present gradient. This will facilitate smoothing-out the descent route of complex error surface by avoiding great alteration in the gradient. In other words, it prevents the slow convergence of steepest descent method when the error surface is not smooth [252]. Conjugate gradient is another extension to the simple gradient descent algorithm. It searches in conjugate directions which produces faster convergence than that of the steepest descend route. The step size for weight adjustment is altered for all iterations of this approach. It is determined by performing a search in conjugate gradient path which minimize the loss function along that line [253]. The conjugate gradient method is preferred over the gradient descent method as it provides greater performance.

Though, it needs larger memory capacity for intermediate results than gradient descent methods. Moreover, it has a lower performance for cases when the error surface is quite uniform or smooth, and when the error function is not quadratic. It is also harder to implement [252].

3.5.2 Network Generalization

A critical issue when training neural network is how well it will generalize on new examples or patients. A model with low complexity may not be able to predict new examples leading to under-fitting. However, too complex models may fit the noise not the data, which may lead to over-fitting leading to predictions that are far away from range of the training. In other words, it will perfectly classify training data, but this is not the case for test data; therefore it should be avoided during training of the network. Several ways are used to prevent over-fitting among them are weight decay, noise injection, early stopping, model selection, and ensemble committee. The weight decay methods works by adding a penalty on the error function which causes large weights to decay to smaller absolute values or zero when no further update is scheduled. The reason for that is; large weights may increase that network complexity and lead to incorrect prediction consequently. In order to reach good generalization using weight decay method, the weights in the network needs at least three decay constants (input-hidden, hidden-hidden, and hidden-output). Updating all of them usually has high computational cost. In other words, it will increase convergence time as the weight decay parameter should be optimized. It also does not put into consideration the function the network is generating, it only penalize the weights. Noise injection is another method which depends on adding an artificial noise to training data, this consequently make the training process unsteady, making the learning process harder to find a solution that fits the training data which prevents over-fitting accordingly [254]. However, it does not work well with datasets containing high level of noise.

Early stopping is the most common approach for avoiding over-fitting. It splits data into two sets training, and validation, it uses the former to learn the network and calculate the error on the validation set. When the error starts to increase for several successive times, the training process is stopped and the performance of the network is measured on the test set which was not used in the building the model. Its main drawback is that the validation error may increase, and then start to decrease once more. It may have several decreasing regions, choosing an inappropriate one will affect predictions. However, this method is considered to be simple, fast , and usually perform similar or better performances than other regularization and noise injection methods [255, 256]. There is no specific method that is optimum; therefore, early stopping was used in this thesis with a threshold of 20 as the number of repeating times the validation error increases. Moreover, cross validation model selection procedure was used to select the number of hidden neurons that reduces chances of over-fitting. In addition, it was employed and combined with feature selection and iterated nested cross validation to reduce chances of over-

fitting, avoid bias and generate stable predictive model in chapter 4. Also, in chapter 5, a multiple classifier system was constructed which combines classifiers predictions to generate more powerful system and avoid over-fitting.

3.5.3 ANN Construction

An ANN consists of one input and one output layers; however it may have several hidden layers. Previous studies showed that only one hidden layer is sufficient to solve complex relations with appropriate number of neurons [257-259], therefore a three layers MLP-ANN was constructed with the uncensored Centers 1 data to predict the risk of re-intervention after 5 years from EVAR surgery and classify Centers 2 patients into high-risk and low- risk groups. Moreover, two ANNs were built using each EVAR datasets of Centers 1 and 2 separately in order to evaluate the performance of each of the uncensoring approaches. Note that this was done using 10 fold cross validation test. A gradient descent with momentum back-propagation (BP) optimization algorithm was used to train the networks. A log sigmoid transfer function was used as activation functions for the hidden and output layers. The number of hidden neurons for each ANN was determined through an iterative approach starting from one hidden neuron and then adding one iteratively. The criteria for choosing the number of hidden neuron is maximizing the area under ROC curve or minimizing the p-value of the log rank test in the testing dataset. The chosen number of hidden neurons was 18 and 14 for building the separate ANN models for Centers 1 and 2 distinctly and 17 and 15 for the ANN constructed with Center 1 and tested with Center 2 using O & Ma and S & DB uncensoring approaches respectively. Early stopping method was used to terminate ANN training process in order to avoid over-fitting. The method in [260] was used to determine the percentage split for training and validation. This method indicates that if the number of training examples is lower than the number of free parameters or weights by a factor of 30, then over-fitting will certainly occur and early stopping will improves generalization performance.

3.6 Results

3.6.1 Separate ANN for Center 1 and 2 respectively

The number of hidden neurons was 18 and 14 for Centers 1 and 2 respectively. As mention before, 10 fold cross validation was used to test the performance of the uncensoring approaches on each Center separately. The results are shown in Tables 3.1 and 3.2. A comparison was made between the prediction results of the two ANN constructed using the two datasets before and after the two uncensoring approaches and discussed in Tables 3.1 and 3.2. Note that the correction threshold used to uncensor datasets is equivalent to ratio between patients used to

train a high risk Bayesian network to that used to learn both high and low risk networks (0.4 for Center1 and 0.5 for Center 2).

As mentioned before, the number of patients that did the re-intervention of the EVAR operation (high risk patients) is 42 and 26 for Center 1 and 2 respectively. After using the two uncensoring approaches, the number of high risk patients has reached 69 (S &DB approach) and 157 (O & Ma approach) for Center 1, while for Center 2 is 70 (S & DB approach) and 75 (O & Ma approach). Tables 3.1 and 3.2 indicate that the two methods have successfully uncensored the datasets as the area under the ROC curve (AUROC) for prediction results has been increased using both approaches. However, the O & Ma technique outperforms the S & DB method as the AUROC in the former are 0.86 and 0.775 while that in the latter are 0.753 and 0.661 for Center 1 and 2 respectively which were 0.593 and 0.497 before using the uncensoring approaches. Moreover, the two predictive models before using the uncensoring approaches were biased towards the low risk (zero targets) with true negative and true positive values of 0.942 and 0.089 for Center 1, and 0.942 and 0.077 for Center 2. This indicates that only 8.9% and 7.7% of the patients who experienced the REINT in Center 1 and 2 have been correctly classified by the censored model, which proves that censoring is the main reason why standard machine learning techniques could not be used directly on censored survival data as it will fail to work properly in constructing survival models.

Table 3.1 Comparing prediction results of Center 1 before uncensoring (a) and after uncensoring using the S&DB (b) and O&Ma (c) techniques.

(a) Before uncensoring (Center 1)			
Class	**True**	**False**	**Area under Roc**
High risk (positive)	0.089	0.911	0.593
Low risk (negative)	0.942	0.058	
(b) After uncensoring with S&DB approach(Center 1)			
High risk (positive)	0.478	0.522	0.753
Low risk(negative)	0.92	0.08	
(c)After uncensoring with O&Ma approach(Center 1)			
High risk (positive)	0.713	0.287	0.86
Low risk (negative)	0.853	0.147	

Table 3.2 Comparing prediction metrics of Center 2 before uncensoring (a) and after uncensoring using the S&DB (b) and O&Ma (c) techniques.

(a) Before uncensoring (Center 2)			
Class	**True**	**False**	**AUROC**
High risk (positive)	0.077	0.923	0.497
Low risk (negative)	0.942	0.058	
(b) After uncensoring with S&DB approach (Center 2)			
High risk (positive)	0.4	0.6	0.661
Low risk (negative)	0.817	0.183	
(c)After uncensoring with O&Ma approach (Center 2)			
High risk (positive)	0.613	0.387	0.775
Low risk (negative)	0.811	0.189	

3.6.2 Center 1 Neural Network Model for Prediction of Center 2 patients

A three layer ANN was constructed using Center 1 to predict the risk of EVAR re-intervention on Center 2. Note that, the discrimination threshold used to separate between high and low risk prediction was chosen according to the ROC analysis which indicates values of 0.5041and 0.4641 for S & DB and O & Ma technique as shown in Figure 3.2. KM curves were plotted for each predicted risk group and the log-rank test was used to determine the significance between the risk groups and also indicate whether the model has correctly classified patients. Usually a p-value less than 0.05 indicate that the two risk groups are significantly separated, distinguished and different. Also, the CI metric was used to test the discrimination of prediction results.

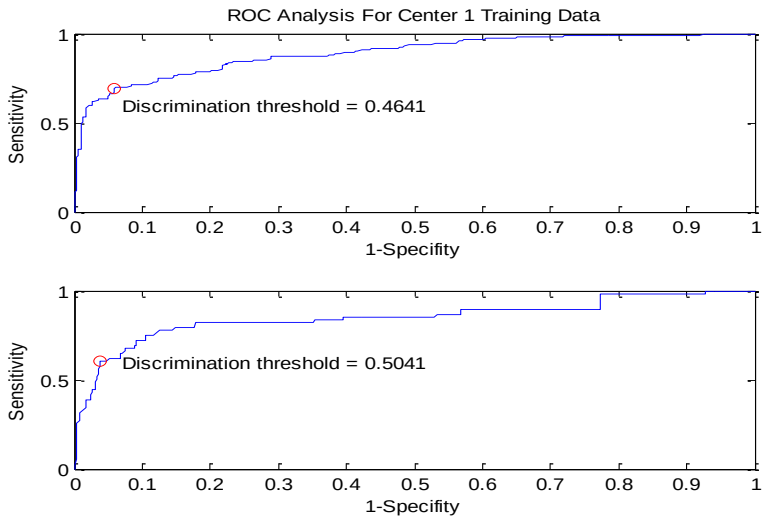
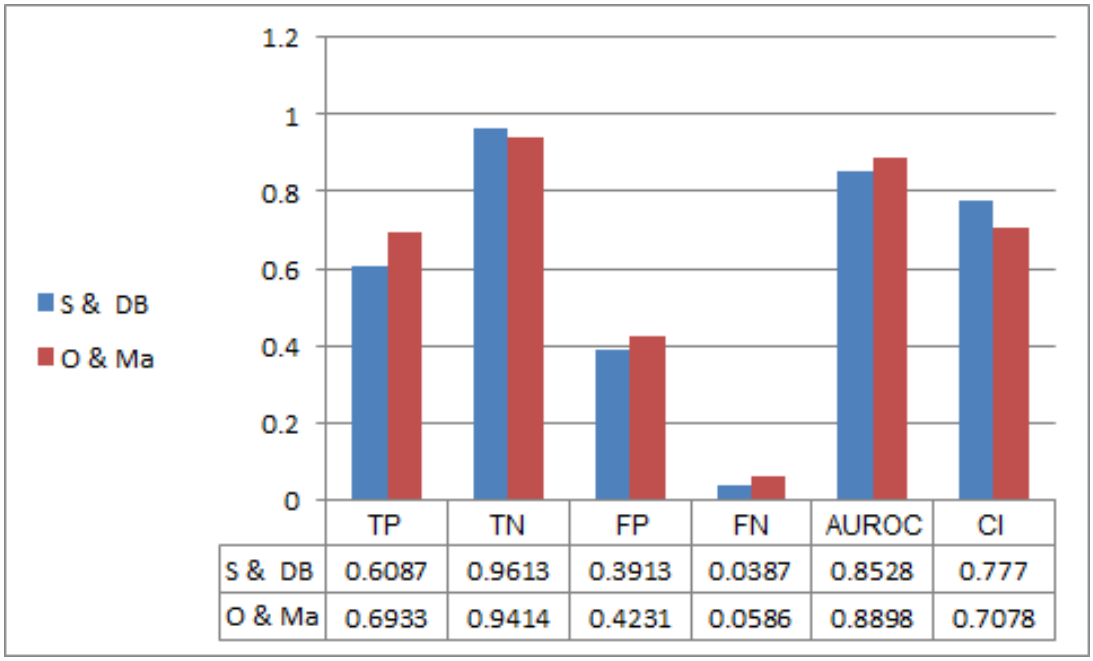


Figure 3.4 The ROC analysis for Center 1 training data using (Upper) O & Ma uncencensoring approach, (Lower) S & DB uncensoring approach to determine the discriminating threshold for each construct ANN.

3.6.2.1 Comparison between Center 1 Results of S & DB and O & Ma Methods

The results in Figure 3.5 show that the AUROC of the trained model with uncensored Center 1 using O & Ma method was 0.8898 which is better than that of S &DB (0.8528). Moreover, the true positive and negative rates of O & Ma method (0.6933 and 0.4231) are greater than the other method (0.6087 and 3913). However, the concordance index (CI) of S & DB technique (0.777) is higher than O & Ma (0.7078).



	TP	TN	FP	FN	AUROC	CI
S & DB	0.6087	0.9613	0.3913	0.0387	0.8528	0.777
O & Ma	0.6933	0.9414	0.4231	0.0586	0.8898	0.7078

Figure 3.5 Comparison of training classification results of the neural network model trained with uncensored Center 1 data using S & DB and O& Ma methods.

The Kaplan Meier curves were plotted for the predictions of Center 1 using both methods are shown in Figure 3.6. The resulted p-value of the log rank was lower than 0.0001 for Center 1 using both methods indicating a good discrimination between risk groups of aortic complications. This discrimination is obvious from Figure 3.5 which shows that the two risk groups of EVAR re-intervention are well separated.

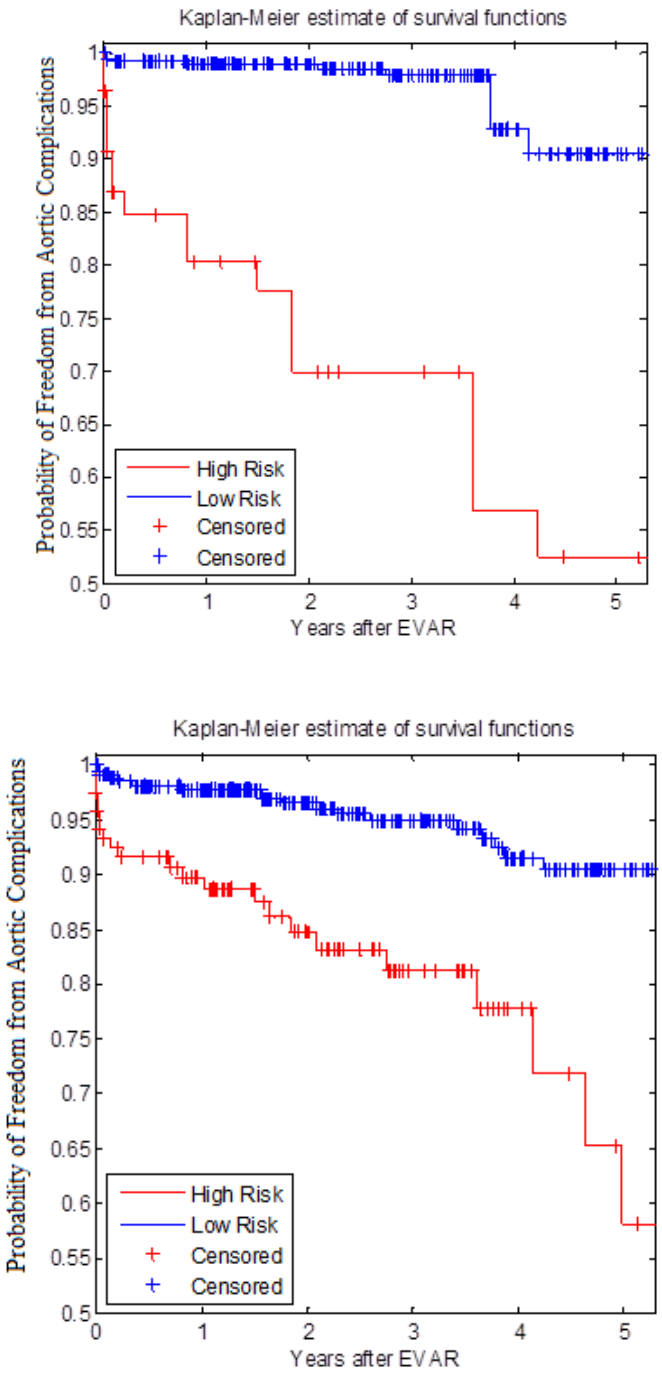
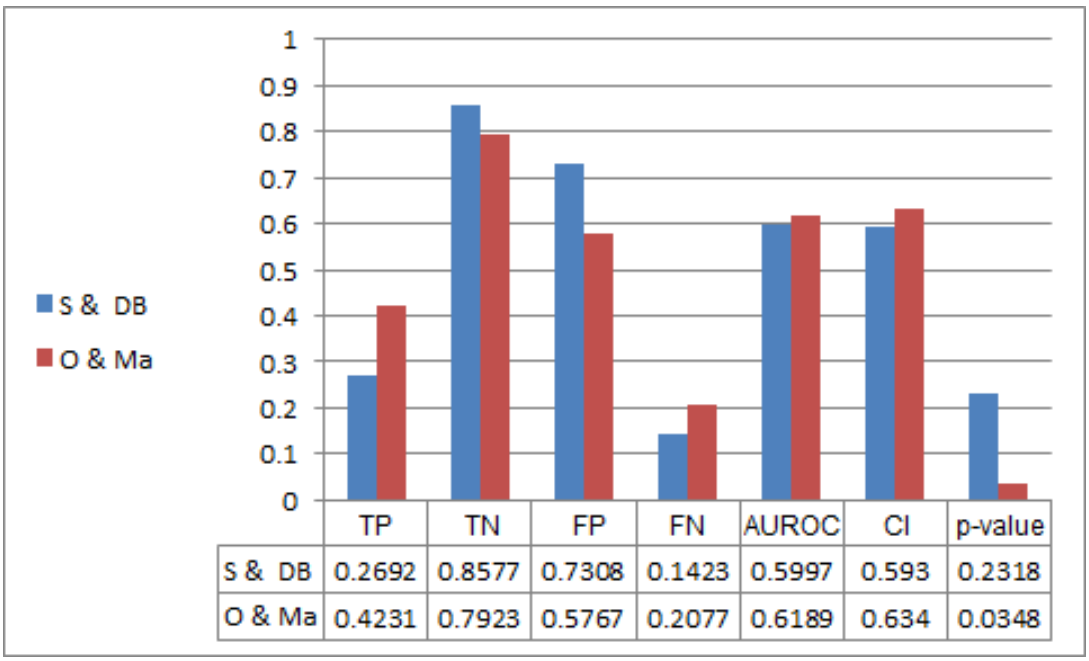


Figure 3.6 Kaplan Meier Curves prediction of Center 1using the neural network model built from the uncensored Center 1 dataset with S &DB (Upper) and O & Ma (Lower).

3.6.2.2 Comparison between Center 2 Results of S & DB and O & Ma Methods

As mentioned before, it is preferred to use external validation method to produce a clinical model that can be used for medical application. This model should establish corresponding performance on other patients from another Center who have different population from the one used to construct the model. The opportunity to cross check the model prediction upon different centers strengthens the results and study outcomes[15, 16]. Therefore, Center 2 prediction is

important in validating the model performance. As it is shown from Figure 3.7, the AUROC is better using O & Ma (0.6189) method comparing to the value of 0.5997 achieved using the S & DB technique.. The CI is better used for censored survival data type as discrimination criteria, therefore it was used for Center 2 as well as a metric to show the ability to separate risk groups and compare performance between O & Ma and S & DB methods. It turns out to be 0.6340 for Center 2 which is better than that of S & DB method (0.593). Moreover O & MA technique has higher ability than the other method to correctly classify high risk patients, as its true positive rate is equal 0.432 versus 0.2692 of S & DB approach.



	TP	TN	FP	FN	AUROC	CI	p-value
S & DB	0.2692	0.8577	0.7308	0.1423	0.5997	0.593	0.2318
O & Ma	0.4231	0.7923	0.5767	0.2077	0.6189	0.634	0.0348

Figure 3.7 Comparison of testing classification results of the neural network model trained with censored Center 2 data using S & DB and O& Ma methods.

The Kaplan Meier curves were plotted for the predictions of Center 2 and were shown in Figure 3.8. The resulted p-value of the log-rank was equal to 0.0348 for Center 2 using O & Ma method, meaning that the model has succeeded to differentiate between the low and high risk groups. However, the p-value of the log-rank was equal to 0.2318 using S & DB method, indicating that this technique was incapable of distinguishing between risk groups of aortic complications.

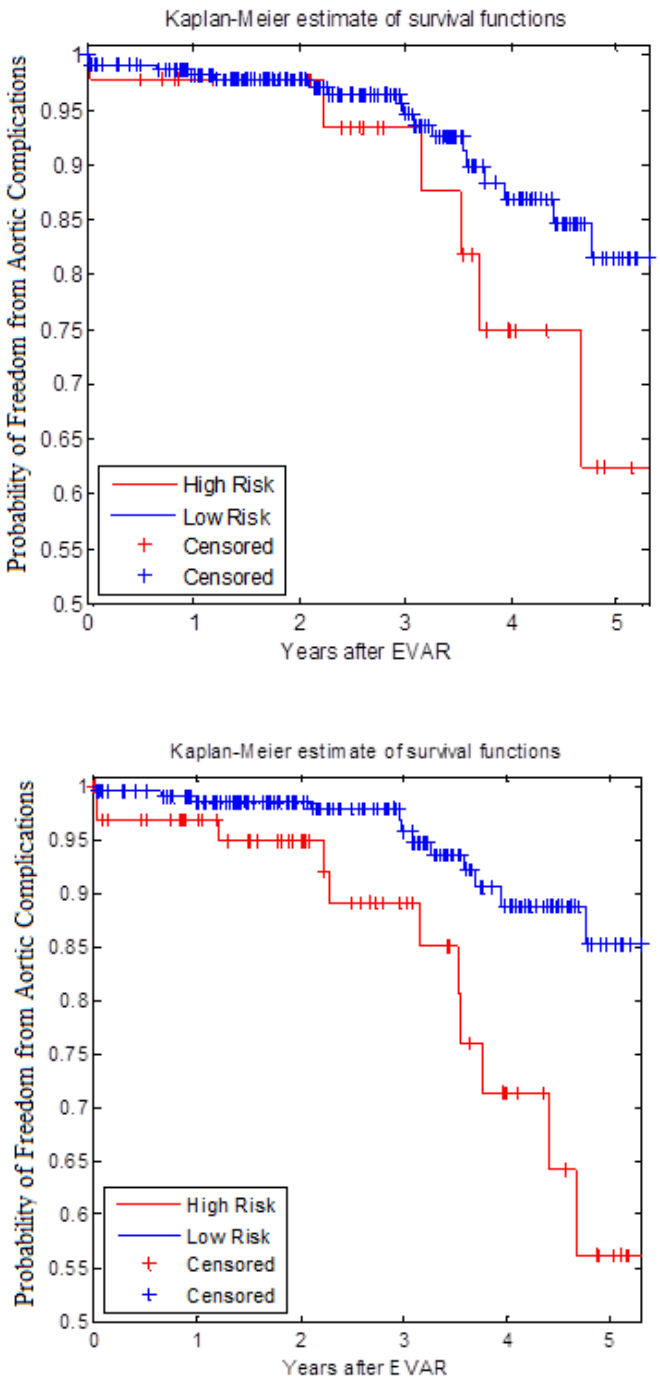


Figure 3.8 Kaplan Meier Curves prediction of Center 2 using the neural network model built from uncensored Center 1 dataset produced with S&DB (Upper) and O& Ma (Lower) methods.

3.7 Discussion

A Bayesian network and gradient based back-propagation neural network is capable of building a predictive model of REINT for EVAR patients. The Bayesian networks has successfully uncensored the clinical datasets. As shown in the Table 3.1 and 3.2, the two BP neural network models that tested the efficiency of the uncensoring algorithms had increased the AUROC from 0.593 to 0.86 and 0.497 to 0.775 for Center 1 and Center 2 patients respectively. Additionally, the proposed O & Ma algorithm outperforms S & DB when using the uncensored Center1 data to build neural networks models to predict risk of re-intervention after EVAR for censored Center 2 data. The results in Tables 3.3 and 3.4 showed that O & Ma prediction is more concordant (0.634) than S & DB (0.593). Moreover, the model constructed using S & DB was unable to distinguish between the two risk groups of censored center 2 data (p-value of the log rank test=0.2318), however O & Ma succeeded in differentiating between them (p-value of the log rank test=0.0348). Therefore, low risk patients can be monitored at less regular intervals to reduce their exposure to radiation and costs involved with frequent monitoring.

Advantages of this Bayesian-Neural Network approach are; first, explaining relations between variables in a visual form, as medical researchers are not aware of data mining and machine learning techniques. Since, the Bayesian network is a graphical probabilistic network; it shows joint probabilities between variables in the form of a graph consisting of nodes representing these variables and arcs showing relations between them which explain causality. Hence; it is considered the perfect tool that can be used in medicine to show causal influence between variables and their probabilistic relations [261]. Moreover, physicians could calculate the condition and marginal probabilities of the network which demonstrate the uncertainty of the studied medical domain [190]. Second, this approach was capable of uncensoring a highly censored dataset. This uncensored data can be used later with any standard supervised machine learning technique as the time variable was embedded in the uncensoring technique. Third, the neural network is a powerful classifier as it successfully predicted aortic complications and distinguished between the high-risk and low-risk groups of patients.

Matlab and Weka Softwares [262] were used for the implementation of the proposed techniques. Weka is a free, simple and user friendly data mining software that can be embedded easily in a more complex system. It provides a large number of classifications, clustering, and variable selection methods implemented in java and do not requires coding.

3.8 Clinical Findings

This study demonstrated that prior to surgery, it is possible to identify that the majority of patients will have a negligible 5-year risk of developing aortic complications after EVAR. This finding has great clinical relevance because the need for surveillance after EVAR is determined wholly by the incidence of aortic complications; a validated ANN model might therefore allow surveillance to be safely abandoned in patients undergoing EVAR. Previous scoring systems have not been able to isolate patients with a sufficiently low risk of aortic complication to allow surveillance to be safely abandoned [263, 264]. However, the proposed O & Ma ANN predictive model showed a good discrimination between low and high risk groups which allow more patients to undergo an appropriate surveillance plan.

The impact of aortic morphology on long-term outcome of EVAR is likely to be complex and well-suited to ANN analysis, with considerable potential for interaction between aortic volume, shape, diameter, angulation, endograft configuration, and patients' cardiovascular health. The majority of patients are low-risk as classified by the proposed O & Ma - ANN model which was in-line with the clinical practice [21]. This proves that O & Ma predictions make medical sense. The ANN provided a more powerful prediction of the low-risk and high -risk groups than has previously been possible by existing methods taking aim at providing patients with individualized risk profiles [263, 264] , which resulted in superior performance in simulated surveillance studies. This finding is not unexpected due to the construct of ANNs; that enables them to outperform logistic or survival models by including multiple interacting and complex covariate effects for event prediction.

3.9 Conclusions

Machine learning techniques have been used widely for survival analysis in clinical trials. Censoring is a common problem that appears when dealing with survival data. Censored patients cannot be ignored or discarded, as this may bias and this reflects on the results, especially when the data is highly censored. In this chapter, an approach has been proposed to deal with the high censoring issue that appeared in the EVAR datasets of the two different vascular centers.

Two Bayesian networks were produced called high-risk and low-risk networks and they were used to relocate censored patients to either the low or high risk groups. This was done by calculating the output probability that any patient belongs to low or high risk network given that he/she is censored. If the output probability of the high risk network given a censored instance is greater than a specific threshold, this patient is considered a high risk patient and vice versa.

After applying the proposed algorithm on the two EVAR datasets, two MLP-ANN models were used to test the efficiency of the algorithm on each dataset separately. Next, a third one was constructed using Center 1 dataset and employed to classify patients of Center 2 into high or low risk categories. The proposed O & Ma approach has successfully increases the AUROC for both Centers, Moreover; it was able to predict the risk of REINT on censored Center 2 data after 5 years with the BP neural network model built with uncensored Center 1 dataset. The results showed that the model prediction was more concordant (0.634) than S & DB approach (0.593). It also succeeded to differentiate between the low and high risk groups with a p-value of the log rank test equals to 0.0348 which is not the case for S& DB approach (p-value of the log-rank test equals 0.2318). The high risk means that the patients will more likely need re-intervention and therefore frequent monitoring of conditions, whereas less frequent monitoring is needed for low risk patients. A future observation plan could then be planned to new EVAR patients with the neural network model built by the new proposed approach.

Chapter 4: Feature Selection and Model Selection through Validation and Un-censoring for Predicting the Risk of Endovascular Repair Re-intervention

4.1 Introduction

Feature selection (FS) and model selection (MS) are important topics in data mining; especially when dealing with real medical datasets of large size. FS searches for a reduced number of variables that have the ability to improve prediction using a selection criterion. MS chooses one optimal (or more) model from a number of candidate models formed from either several classifiers or the same one but with different parameters. MS can be considered as FS; when selecting between models with reduced number of variables which is the main aim of this chapter. FS techniques tend to lower the classifier's complexity and speed up the classification task. In addition, they enhance generalization and prevent over-fitting [8]. Clinicians need them extremely in order to decrease the effort and time need to measure the unnecessary variables to build a reduced predictive model.

FS is widely used for standard medical data, though this task becomes more complex for survival data type due the presence of censoring [8]. Most of the work done for FS was dealing with normal standard data, fewer work was done to deal with censored survival data. The literature has shown that most of the papers employing variable selection methods for survival data are based on Cox's proportional hazard model [9]. Cox's model is the most common statistical survival analysis model; however, machine learning techniques are preferred over Cox's model for their advantages [10]. In this chapter a feature model selection technique is proposed to deal with the highly censored EVAR survival data in order to produce a final stable model using neural network machine learning classifier. The proposed approach is able to predict the risk of EVAR re-intervention after 5 years to patients from two different Centers located in the UK, which allows it to be potentially applied for cross-center predictions. The proposed model is compared with the two states of art survival variable selection techniques: Akaike and Bayesian information criteria (AIC, BIC) that are used with Cox's model. The results show that the proposed approach outperforms the other two FS approaches constructed using Cox's model.

4.2 Feature and Model Selection for Survival data

Model selection methods are considered as feature selection techniques when the purpose is to choose between several subsets of variables generated during model selection. Feature selection

is very useful especially in medical areas, as it reduces the time needed and the effort made by physicians to measure irrelevant and redundant features. It could avoid over-fitting that might occur during the learning process of the predictive model. It may also lower its complexity and speed up prediction process [8]. Feature selection techniques are used widely for standard data, though this task is more complex for survival data t due the presence of censoring. The literature has shown that a large number of papers employ variable selection methods that use Cox's proportional hazard model [9] such as; penalized L1 Least absolute shrinkage and selection operator (LASSO) and partial least squares embedded methods. Other papers wrapped feature selection around Cox's model and used several criteria such as the Akaike or Bayesian information criterion, hazard ratio, and concordance index (CI) calculated from its prediction to select features. Some used it as a filter approach and perform univariate analysis to calculate Cox's score metric for selection, while others used Wald or likelihood test criteria instead to quantify variables association with survival prediction. However, a feature model selection is proposed in this chapter which uses neural network machine learning classifier instead of standard statistical survival models due to its various advantages which were previously discussed in Chapter 2 section 2.6. The proposed feature selection method used factor analysis feature reduction technique as an initial step to group observed variables related to latent factors and remove variables not related to any latent factor. Moreover, it embeds the uncensoring approach that was previously proposed and discussed in Chapter 3 with a feature selection process to handle censoring in order to construct and validate a predictive model capable of predicting aortic complications 5 years after the EVAR surgery. The results of the proposed approach were compared with two popular survival feature selection methods that employ Cox's proportional hazard model.

4.3 Feature Reduction and Transformation

Multivariate data with high dimensions cannot be easily visualized and understood. The purpose of feature transformation and reduction techniques (also known as multivariate analysis) is to transform this data into a new transform domain and reduce its dimension so that most of the classification related information is compressed in a smaller number of features, leading to a reduction of the necessary feature space dimension which will help in visualising and understanding relations between variables of a dataset. This can remove the redundancy which usually exists in some variables due to measurement errors obtained by devices. However, the main limitation of these methods is that they end up by producing a large number of variables chosen in the reduced dimension space [265] .

Principal component analysis (PCA) and Factor analysis (FA) are the most popular feature transformation and reduction techniques [265]. PCA explains variations among variables of a

datasets. It analyzes the data by projecting data points into a new uncorrelated and orthogonal variables called principal components using Eign decomposition [266]. Principal components are linear combinations of variables which maximize variations. FA analyses variables of a data by identifying new smaller unobserved variables called common factors (similar to principle components in PCA); which capture correlation among observed variables. FA also tries to group variables that are related. This helps in understanding the underlying structure of the data which cannot be directly observed, and also discovering redundant variables [265]. Variables that does not refer to any group and does not have a large loading to one or two factor may be considered as unimportant variable. FA is preferable than PCA and therefore was used in this study, as the latter assumes that the only source of variance in a data is the common variance shared by variables. Unique variance indicates the unreliable variance of measurement errors and the part of variance not shared by common factors. However, FA considers the existence of both common and unique variances [265]. One aim of this study is to select a reduced number of discriminate features that distinguish between risk groups of aortic complication. Another reason to use FA rather than PCA is that the second loses its local features of the data. In other words, there is no direct relation between the feature space of PCA and the original variables of the data [12]. While in FA, variables not related to any of the latent variables, or do not have large loading to one latent factor can be removed as they are considered unimportant. This leads to a reduced number of features that can be used later in the feature selection process that will be proposed in this thesis.

Some terms have to be defined before explaining how factor analysis works:

- **Observed variable** is the original variable which is measured directly.
- **Factor also known as Latent Factor** is like in the principal component analysis, it is indirectly measured by seeing its correlation with observed variables. There are two types of factors: unique and common.
- **Factor loadings** are the values or weights (called loadings) multiplied by each factor then added together to form a linear combination of factors that produce observed variables.
- **Common Variance** is the type of variance that is common or shared between all variables of a dataset.
- **Unique or Specific Variance** is the type of variance that is unique to only one variable which usually occurs due to error during measurement.

- **Common Factor** is the unobserved variable that shows the common variance between two or more observed variables.
- **Unique Factor** is the unobserved variable that shows the unique variance of only one variable.
- **Communality** is the portion of the variance produced due to common factors. It is calculated by adding the square of factor loadings.
- **Specificity** is the portion of the variance produced due to specific factors.
- **Score** is the value of a data point in a new transformed axis (factors or principle components). In case of PCA they are the actual, but in FA they are estimates of the latent factors.
- **Oblique angle** is any angle other than the 90 degrees.

As mentioned before, factor analysis assumes that there are a number of unobserved factors or variables smaller than the dimension of a dataset that can better express relationships between variables. In factor analysis, the observed variables are the linear combinations of a smaller number of factors which is not the case in PCA where principal components are the linear combination of the observed variables. In other words factors create variables, while variables create principal components [265].

To understand how factor analysis is performed [265], assume a set of variables $x_1, x_2, \ldots x_p$, and a set of common factors $f_1, f_2, \ldots, f_m$. where $m < p$. Each variable x is a linear combination of common factors f and unique factor e which is obtained due to error in device measurements [267], where; l are constants called factor loadings.

Equations (4.1)-(4.3) calculates the variables:

$$x_1 = l_{11} f_1 + l_{12} f_2 + \ldots\ldots\ldots l_{1m} f_m + e_1. \tag{4.1}$$

$$x_2 = l_{21} f_1 + l_{22} f_2 + \ldots\ldots\ldots l_{2m} f_m + e_2. \tag{4.2}$$

$$x_p = l_{p1} f_1 + l_{p2} f_2 + \ldots\ldots\ldots l_{pm} f_m + e_p. \tag{4.3}$$

The main reason why factor analysis is applied to a data is to reduce its dimensions into a smaller number of factors. They can easily visualise groupings between variables, and help in understanding the underlying structure of the data which cannot be directly observed. Sometimes after applying FA, the initial factors clearly visualise these groups. Hence, factor rotation is needed. Rotation here means that the original factor axis is rotated into a newer

rotated axis called “rotated factors” [266]. It searches for new axes that create large loading on some factors while very small loadings on others. The two well-known rotation techniques are varimax which produces orthogonal uncorrelated rotated factors, and promax; which is an oblique rotation produce non-orthogonal correlated rotated factors. Promax rotation is sometimes preferred over the varimax as it is a faster and simpler method [268]. However, in this study varimax rotation was used to generate uncorrelated rotated latent factor in order that the variable related to different latent factors will be uncorrelated.

4.4 Challenges of the proposed method

Generally, feature and model selection techniques lead to a model with lower complexity and variance than that of the full model. However, it may lead to an unstable model. Additionally, the collinearity between ignored and selected features may lead to selection bias. Omitting important variables from the final model may lead to under-fitting and omission bias. Adding unnecessary variables may lead to over-fitting and increase the variance of the final model. Usually, the leave one out cross validation and boostrapping resampling methods are used to train a model with different subsets of data which may reduce the high bias. However, cross validation method is preferred due to its lower computational cost [269]. The first challenge in this work is to reduce bias and instability, and improve consistency of the feature selection using an iterated nested cross validation method. The second one is to apply feature selection techniques on highly censored data such as the EVAR survival data. Third, is to use machine learning techniques to solve censoring issues and build a model to predict the risk of re-intervention after EVAR. Since, cross-center testing is essential when the target is to validate the model for wider applications. The last challenge is to construct a predictive model using Center 1 EVAR data and testing it using other data collected from Center 2 (multicenter study).

4.5 The Proposed Feature-Model Selection Technique

This work was proposed to utilize the most up-to-date and appropriate feature selection, survival analysis and machine learning techniques to produce an accurate, efficient and stable EVAR re-intervention prediction model. The proposed technique uses Center 1 for feature selection and model selection and Center 2 for its validation and assessment which enable this model to be used for cross- center prediction. It is divided into seven main steps. Figure 4.1 illustrates the main three areas of contributions of the proposed approach which are feature selection, uncensoring, and classification with their interactions which are highlighted in blue colour.

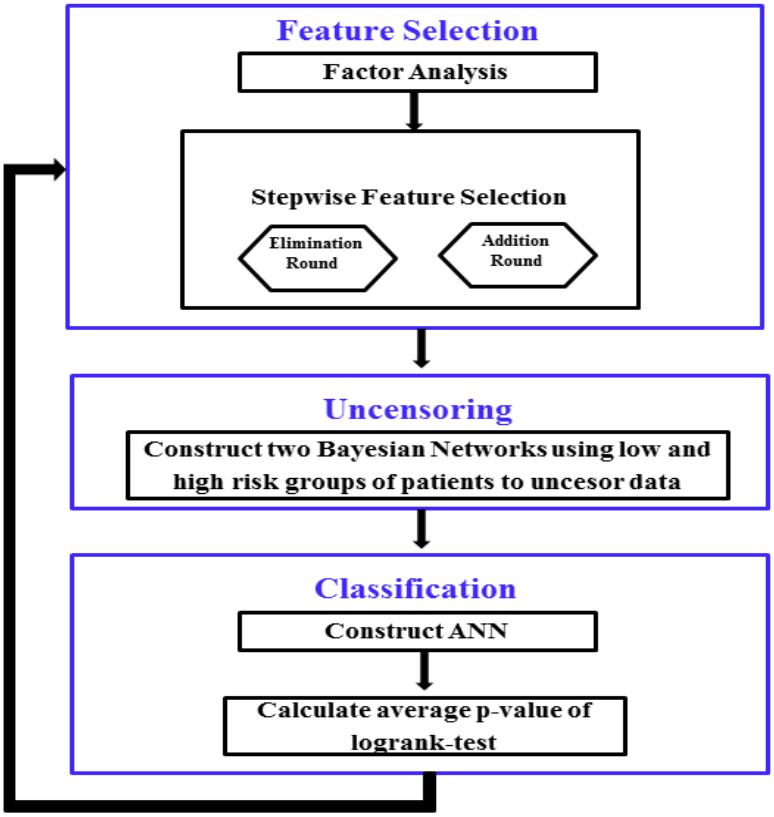


Figure 4.1 The main three areas of contributions of the proposed approach (feature selection, uncensoring, and classification) and their interactions which are highlighted in blue colour.

The details of the proposed algorithm are illustrated in the flowchart shown in Figure 4.2. The first step of the proposed approach is feature reduction using factor analysis technique. This was done after using Kaiser-Meyer-Olkin and Bartlett's tests to examine if factor analysis is needed for the data [270, 271]. The second one is nested cross validation splitting. The third is the uncensoring step which uses a Bayesian network to uncensor data. The fourth is an iterated nested cross validation step followed by an ANN model construction step to predict the risk of re-intervention. The sixth step is stepwise feature selection (SWFS) which uses the p-value of the log- rank test as a criterion. Steps two to six are repeated for each fold produced from the cross validation splitting step. They are repeated until a model with a minimum number of features is produced which also minimizes the p-value of the log-rank test. The last step is to choose among the different models constructed using each training fold generated in the cross validation step. The final model is the one which has the smallest p-value of the log rank test on the remaining censored test fold that was not used in training. This p-value is used to determine if the ANN model built with the selected features was capable of differentiating between the two risks groups of the censored Center 2 EVAR data. Usually, when it is lower than a significance level of 0.05, the two risk groups are considerably different, discriminative and separable.

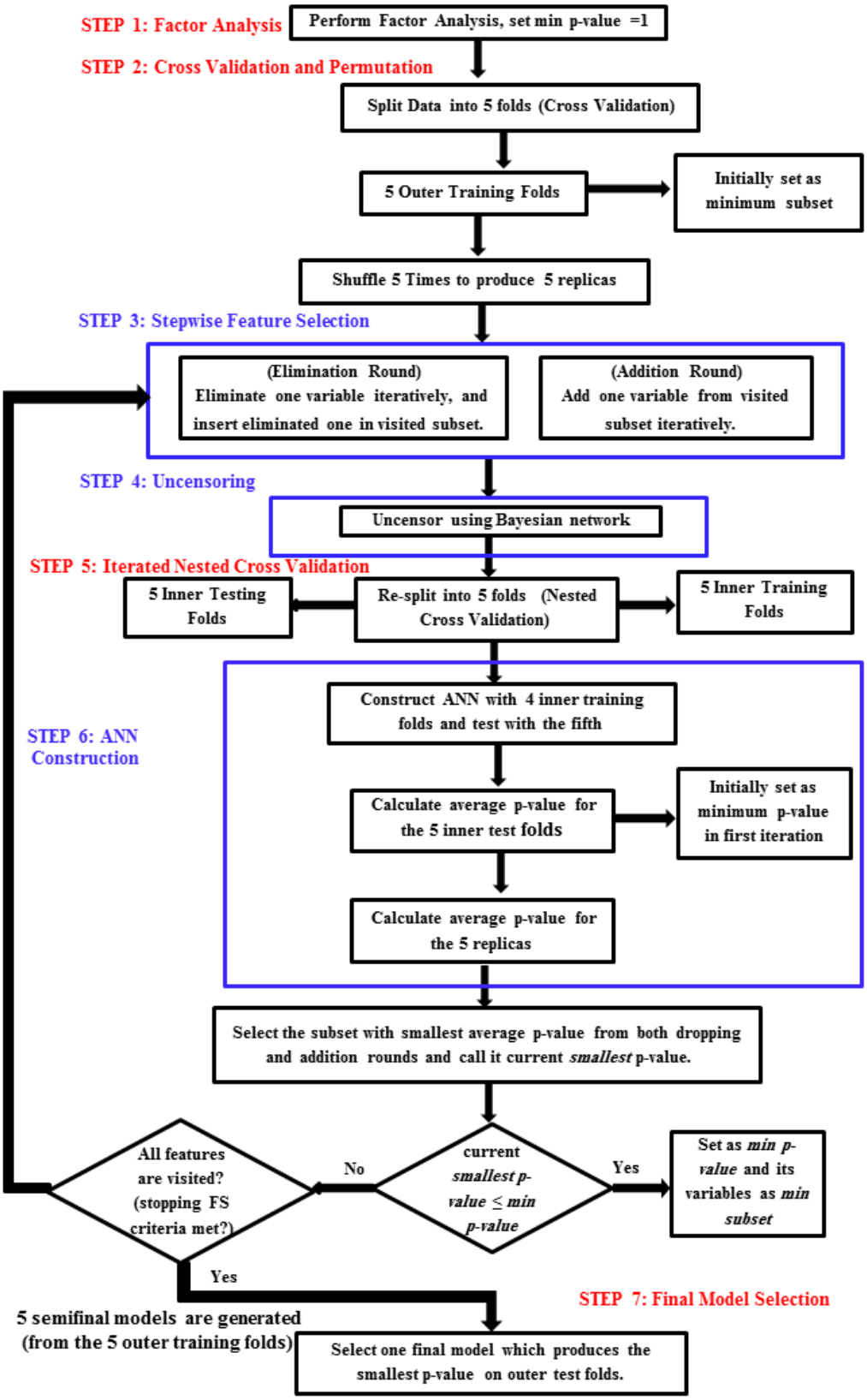


Figure 4.2 A flow chart showing the details of the steps of the proposed feature model selection technique.

4.5.1 Feature Reduction using Factor Analysis Step

The main purpose of using factor analysis in this study is to drop variables not related to any latent factor, not transform data into newer domain and use these compressed variables. Factor analysis was employed in this work as a first step to reduce the dimension of the datasets available. This will reduce the computational cost required later in the proposed feature selection algorithm. The number of latent factors has to be determined first in order to start FA. In order to determine the number of latent factors used in factor analysis, a scree plot is produced which shows the eigenvalues accompanied with principle components or latent factor listed in descending order versus the number of components or factors. As shown in the scree plot of the PCA of center1 data in Fig. 4.3, only six components showed almost 80% of the variance in the data. Therefore, six factors were used for Center 1 data.

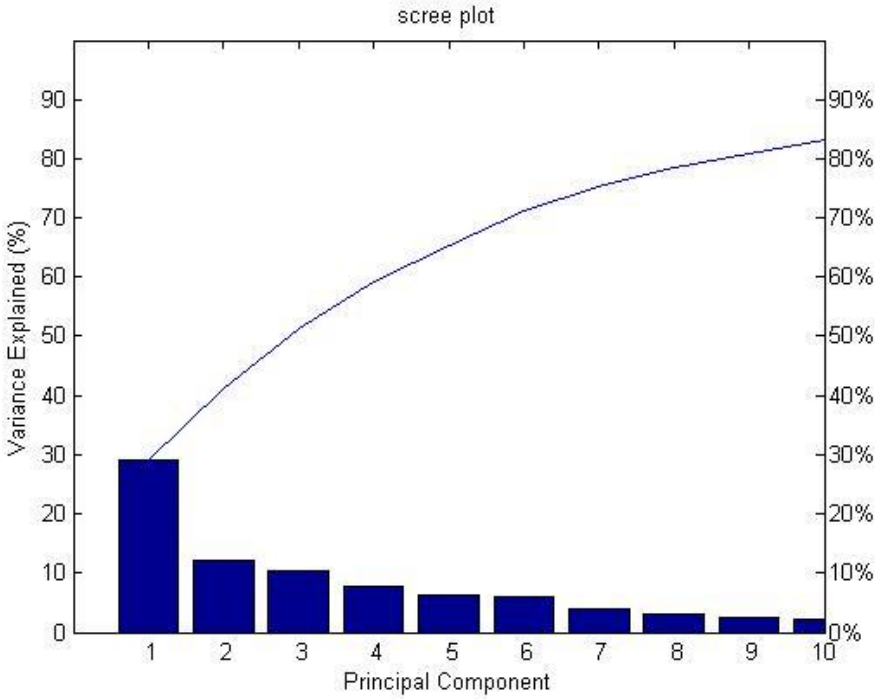


Figure 4.3 A scree plot using Center 1 EVAR data to select the number of latent factors used in the factor analysis process.

Factor analysis is performed on the data using the six latent factors chosen by the scree plot. Some variables can be removed from the dataset as they may be considered as unnecessary. The most common criterion used to drop these variables is communality which is the part of the variance generated from common factors. In this work, the threshold of communality (1-Specificity) according to which features are dropped is determined from the histogram of specific variances (Specificity), produced by specific factors. As shown in Figure 4.4, it is clear that there is a sudden drop in the specificity at 0.25. This threshold was chosen as a result for dropping variables.

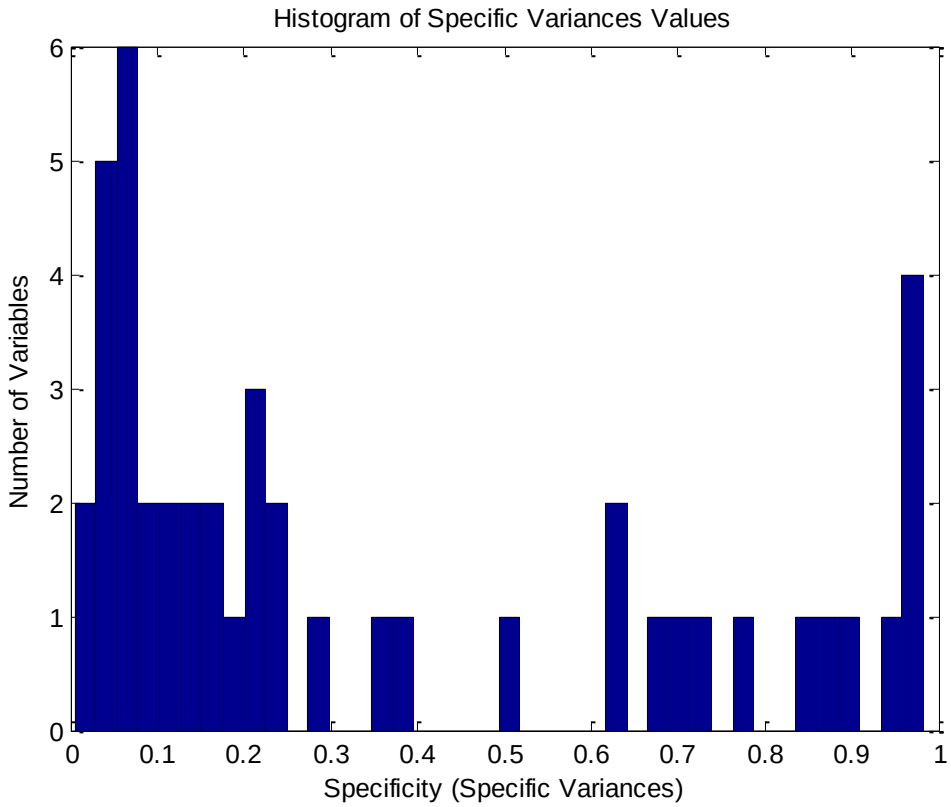


Figure 4.4 A histogram of the specificity values (specific variances) for Center 1 EVAR variables.

4.5.2 Cross Validation and Permutation Step

Center 1 EVAR data is divided into 5 folds using 5 folds cross validation method. Each separate 4 folds is called the outer training fold and the fifth is called the outer testing fold. Each outer training fold is used in nested cross validation, stepwise feature selection, uncensoring, and ANN construction steps later. During the stepwise feature selection step, it is permuted 5 times to produce its replicas after being uncensored (in the uncensoring step).

4.5.3 Stepwise Feature Selection Step

In the stepwise feature selection (SWFS) step, a canonical greedy stepwise search was used to select features using the p-value of the log-rank test as a performance metric. It is divided into two stages; feature elimination and addition. The main advantage of this strategy is that the eliminated features are given another chance to re-enter the feature selection process. It starts with the elimination stage in which all the features of each outer training fold is eliminated iteratively and the rest are used in the uncensoring and ANN construction steps after being permuted and re-split into inner nested folds to calculate the average of the p-value of the log-rank test. This prevents over-fitting and overoptimistic p-value predictions and enables the production of a stable model. The subset with the smallest averaged p-value is the one chosen

and called "minimum subset" and its average p-value is called the "minimum averaged p-value". Features eliminated will be inserted in a subset called the "visited subset" and will be given another chance to enter the FS process again in the addition stage. This is done by adding features of the visited subset iteratively, and then repeating the uncensoring and ANN construction steps. If the subset with the smallest average p-value has a p-value lower than the "minimum averaged p-value", then this subset is set as the "minimum subset" and its p-value as the new "minimum averaged p-value". This procedure is repeated until all features are visited.

4.5.4 Uncensoring Step

After, splitting Center 1 censored EVAR dara into the outer training folds and performing the stepwise feature selection process. The censoring time variable for each outer training loop was used to divide patients into three groups. The first group belongs to patients that experience the re-intervention at a time lower than or equal to 5 years (re-intervened or high risk patients). The second group refers to patients that did not need the re-intervention until a time greater than 5 years (low chance to do the re-intervention or low risk patients). Finally, the rest of the patients are considered as the censored group which is the third group (those who died or leave the follow up observations before 5 years). The same proposed uncensoring technique (O & Ma) illustrated in the previous chapter was used to deal with the censoring issue.

4.5.5 Iterated Nested Cross Validation Step

After uncensoring all the replicas of each outer training partition, each replica is divided into 5 nested folds. Every different 4 folds are called inner training folds and are used to build an ANN (in the ANN construction step), while the fifth one is called inner test fold and is used for measuring the performance of the ANN using the p-value of the log-rank test of this inner test set. This is done for the different inner training and testing folds. Afterwards, the average of the p-value is then calculated. This process prevents overoptimistic results. Note that; each outer training fold was used to construct the ANN with different number of neurons. The process is repeated for every replica of the outer training fold, and then the average of all of them is calculated to produce the p-value of the nested cross validation which is used in the stepwise feature selection step as a criterion to select the attributes. It produces a stable model that tunes the tradeoff between variance and bias. Nested cross validation is also used as a method of avoiding over-fitting used due to the small sample size available in the data.

4.5.6 ANN Construction Step

A three layer MLP-ANN trained with gradient descent with momentum back propagation algorithm was used to build a model for every outer training fold with different number of neurons. Sigmoid was used as an activation function. Other parameters are kept with their

default values. Each model was used in the feature selection process to determine which one has the optimal features that minimize the p-value. The model has the ability to distinguish between the two risks groups of Center 1 data used in the ANN construction, and can be validated with Center 2 data. In Chapter 3 early stopping was used to control over-fitting. In this Chapter, cross validation model selection procedure was used to select the number of hidden neurons that reduces chances of over-fitting. In addition, it was employed and combined with feature selection and iterated nested cross validation to reduce chances of over-fitting

4.5.7 Final Model Selection Step

Finally, five models will be produced from the five outer training folds that were used to build an ANN with different numbers of neurons. Each one is tested with the outer test fold which is not used in the ANN construction and the feature selection steps. The model that produces the minimum p-value in the outer test fold is the one chosen as a final model. The number of features and neurons produced in this model is used to train an ANN using Center 1 data. Center 2 data is used to validate the selected model. Note that the low and high risk Bayesian networks constructed using patients of the low and high risk groups of the final seven reduced features selected in the final model are shown in Figures 4.5 and 4.6.

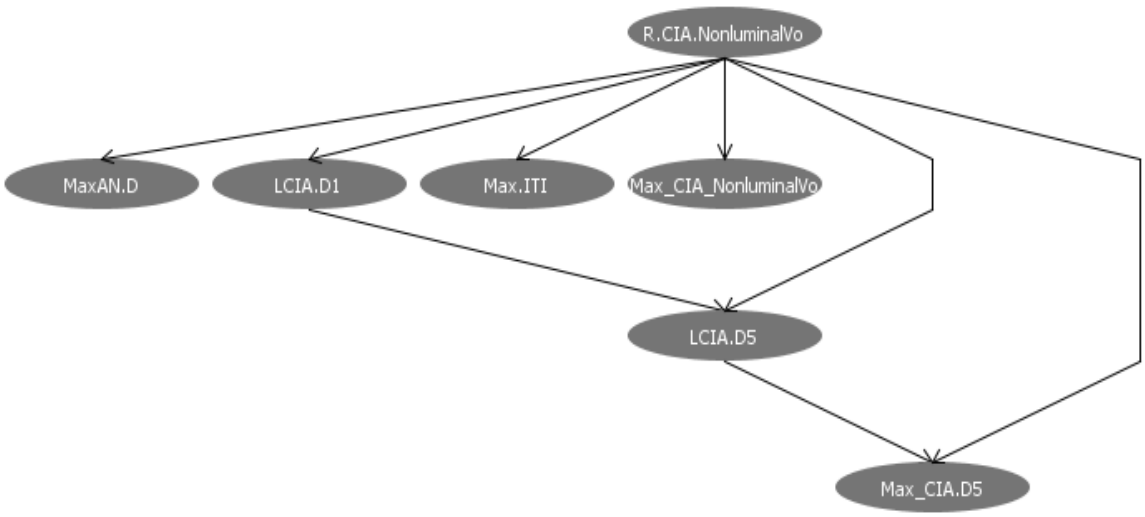


Figure 4.5 The low risk Bayesian network constructed using patients of the low risk group of the final seven reduced features selected in the proposed feature selection approach.

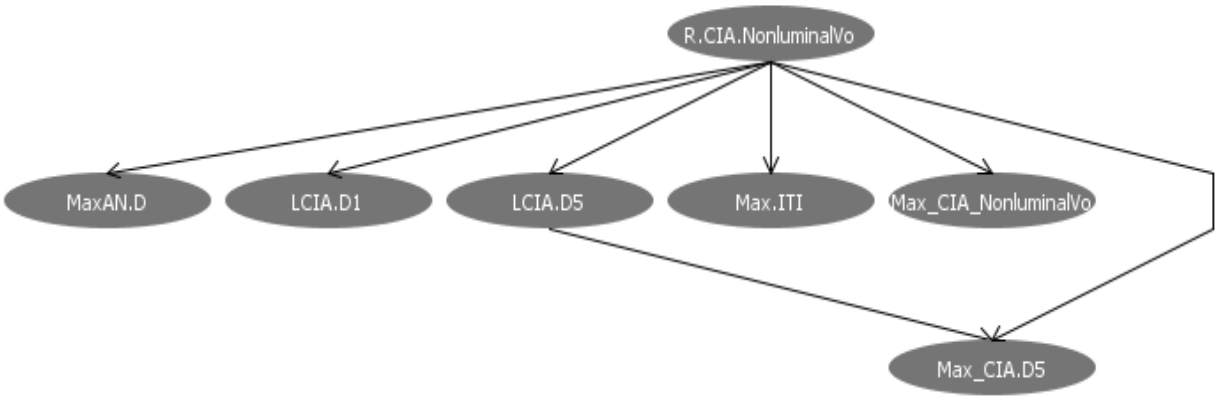


Figure 4.6 The high risk Bayesian network constructed using patients of the high risk group of the final seven reduced features selected in the proposed feature selection approach.

4.6 Results of the Proposed Feature Selection Algorithm

4.6.1 Results of the Proposed Feature Selection and Model Selection Method

One common approach for feature selection is to utilize the whole data set for the selection procedure. Re-sampling techniques such as cross validation and bootstrapping can split data into separate parts. Parts will be used for feature selection and evaluation. Other part not used in model construction will be used for validation and assessing the performance of the final feature selected model. In the proposed algorithm, Center 1 data was used for the feature and model selection (number of hidden neurons of ANN), while Center 2 was employed for the validation and assessment. Cross and nested cross validation resampling techniques were used to split Center 1 data. The inner nested loop was used for producing an unbiased, stable algorithm and overcoming over-fitting and the overoptimistic predictions that might be produced. The outer loop was used to choose the model that produces the smallest p-value of the log-rank test on the outer test set. Note that, the ANN of final chosen model has the number of neurons equal to seven. Table 4.1 shows the p-values of the log-rank test of the five outer test models with their equivalent number of hidden neurons used to train the five outer training ANN models.

Table 4.1 The p-values of the log-rank test of the five outer test folds and the equivalent number of neurons of the ANN models trained with the five reduced outer training folds.

Number of hidden neurons	P-value of the log-rank test
5	0.046
6	0.053
7	0.022
8	0.030
9	0.040

In theory, the more K (number folds) you have in k-fold cross validation is better as it will usually lower prediction bias. Though, large number of K results in high computational cost Also, large K indicates that a small number of sample combinations is possible, consequently limiting the number of iterations that are different [272]. Larger K is usually used with big datasets and smaller K is used with smaller dataset. Large K may lead to higher variance estimators. On the other hand, low values of K leads to higher bias, lower variance predictors [273]. In this thesis, repeated nested cross validation is a solution that was used to tradeoff between bias and variance; however this increases the computational cost. Therefore, 5-fold cross validation was chosen for splitting data. Another, reason for choosing this number is that the EVAR datasets used in this thesis are not big enough to use K equals to 10. Moreover, as mentioned before, increasing K reduces bias, thus K equals to 3 is very small which may lead to biased prediction.

Table 4.2 shows the results of using all the features available in Center 1 data and the results after factor analysis and the stepwise feature selection steps. We see from this table that the number of features has been reduced from 45 to 27 after the FA step. Both sensitivity and concordance index (CI) for Center 2 prediction has been increased from 0.46 to 0.57 and 0.6 to 0.612 respectively. The p-value for Center 2 has improved from 0.036 to 0.034. It is also obvious from Table 4.2 that final features selected in the final model after the SWFS step is lowered to seven. The p-value, CI, and sensitivity for Center 2 predictions have improved to 0.022, 0.6305, and 0.73 respectively.

Table 4.2 Results of the proposed feature selection algorithm using all features, after factor analysis step, and after stepwise feature selection step. We see that the number of features, p-value of the log-rank test, concordance index, and sensitivity were used for comparison.

The proposed algorithm	Number of features	p-value	Concordance Index (CI)	Sensitivity
All Features	45	0.036	0.6 (0.0677)	0.46
FA Features	27	0.034	0.6120 (0.0715)	0.57
Stepwise selection Features	7	0.022	0.630 (0.0739)	0.73

4.6.2 Results of the Final Model of the Proposed Method Compared to Cox's Models based on AIC and BIC

In this section, the results of the final model produced by the proposed algorithm are compared with the most well-known semi-parametric method used in medical applications: Cox's proportional hazard model. AIC and BIC are popular survival feature selection techniques that are used with Cox's model to produce a stable algorithm with a reduced number of features. The parameter estimates from the final models were multiplied by each variable to generate a risk score. The threshold that separated the two risk groups for Center 1 was 3.1 using AIC which is equivalent to the mean value of the risk score. The threshold for BIC is 2.4. The same threshold is applied for Center 2 data. Tables 4.3 and 4.4 show the results of the proposed algorithm compared with the results of AIC and BIC for Center 1 and 2 respectively.

Table 4.3 shows the results of the proposed feature selection algorithm compared to the AIC and BIC Cox's algorithms. We see that this table indicates that these methods have successfully distinguished the two risk groups of Center 1 as they all have p-values lower than 0.00001 which is beyond the significance level 0.05. The CI of AIC and BIC Cox's models are 0.7898 and 0.7624 which is greater than the proposed algorithm 0.7407. However, the proposed algorithm outperforms them in the number of features selected in the final model which is seven, while in the AIC and BIC methods it is 14. Moreover, the sensitivity of the proposed algorithm is 0.77 which is better than 0.69 and 0.38 of the AIC and BIC Cox's models. This means that the final model produced from the proposed feature selection approach can predict more patients that experienced the EVAR re-intervention which is the event of interest in our study. This superiority in the sensitivity rate shows the advantage of using the proposed algorithm over the other two methods in predicting the risk of re-intervention. This also appears clearly in the predictions of Center 2 as shown in Table 4.4, the proposed algorithm outperforms the AIC and BIC Cox's models in the sensitivity which is 0.73 versus 0.35 and 0.23 respectively. Additionally, the results of the final model of the proposed feature selection approach have better p-value of 0.022 than the other two algorithms 0.034 and 0.029. Also, it has the same CI as the BIC model (0.630) and outperforms the AIC with CI equals to 0.610.

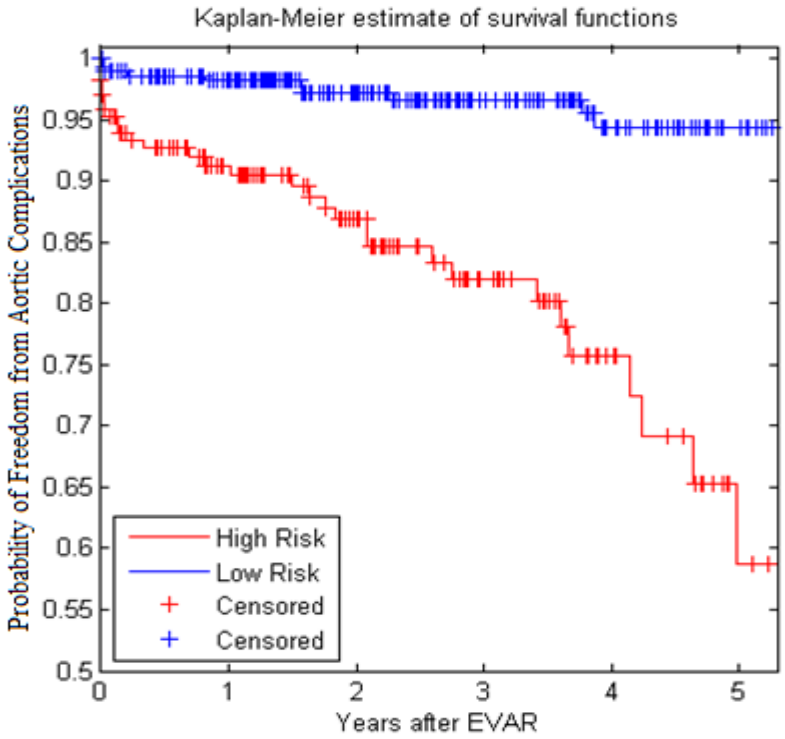
Table 4.3 Results of Center1 data using the proposed feature selection algorithm compared with Cox's model using AIC and BIC. We see that the number of features, p-value of the log-rank test, concordance index, and sensitivity were used for comparison.

Algorithm used	Number of features	p-value	CI (SD)	Sensitivity
Proposed Algorithm	7	‹0.00001	0.7407(0.0439)	0.76
AIC Algorithm	14	‹0.00001	0.7898(0.0408)	0.69
BIC Algorithm	14	‹0.00001	0.7624 (0.0465)	0.38

Table 4.4 Results of Center 2 data using the proposed algorithm compared with Cox's model using AIC and BIC. We see that the number of features, p-value of the log-rank test, concordance index, and sensitivity were used for comparison.

Algorithm used	Number of features	p-value	CI (SD)	Sensitivity
Proposed Algorithm	7	0.022	0.630 (0.0739)	0.73
AIC Algorithm	14	0.034	0.610 (0.0725)	0.35
BIC Algorithm	14	0.029	0.630 (0.0685)	0.23

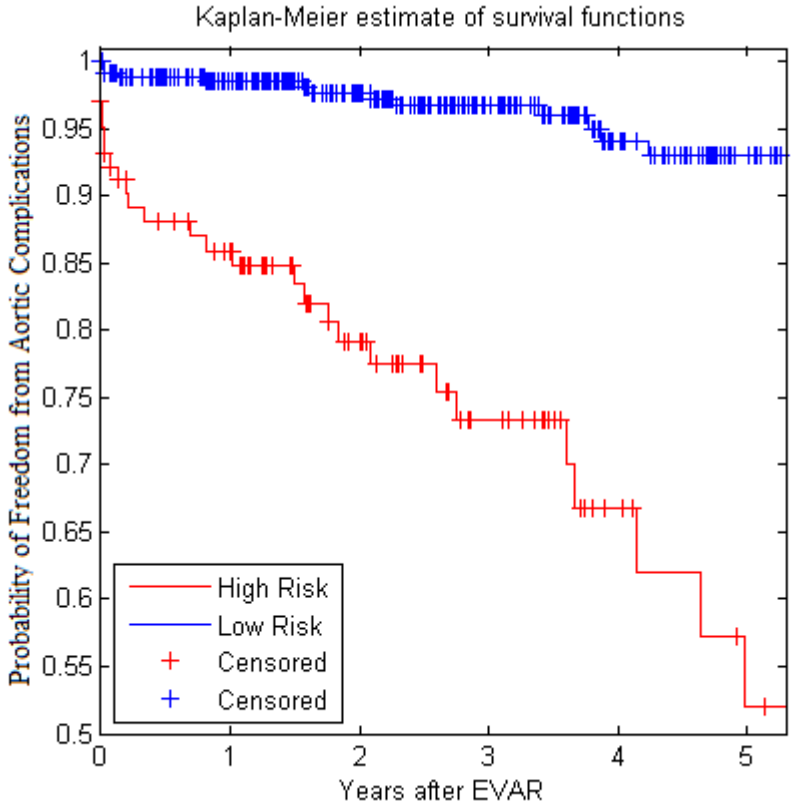
Figure 4.7 shows the Kaplan-Meier curves of the two risk groups using the final selected model of the proposed algorithm for Center 1 compared with the Kaplan-Meier curves of the AIC (Figure 4.8) and BIC (Figure 4.9) methods respectively. Figure 4.10 shows the Kaplan-Meier curves of the two risk groups using the final selected model of the proposed algorithm for Center 2 compared with the Kaplan-Meier curves of the AIC method (Figure 4.11), and BIC method (Figure 4.12) correspondingly. Moreover, these figures include the probability of freedom from aortic complications within the 5 years after EVAR along with the number of patients at risk of each group (low and high risks).



Year		0	1	2	3	4	5
Freedom from Aortic Complications	**Low-risk**	-	98%	97%	96.6%	94.4%	94.4%
	High-Risk	-	90.5%	84.7%	82%	75.7%	58.7%
Number at Risk	**Low-risk**	288	239	177	124	81	45
	High-Risk	169	122	85	52	25	8

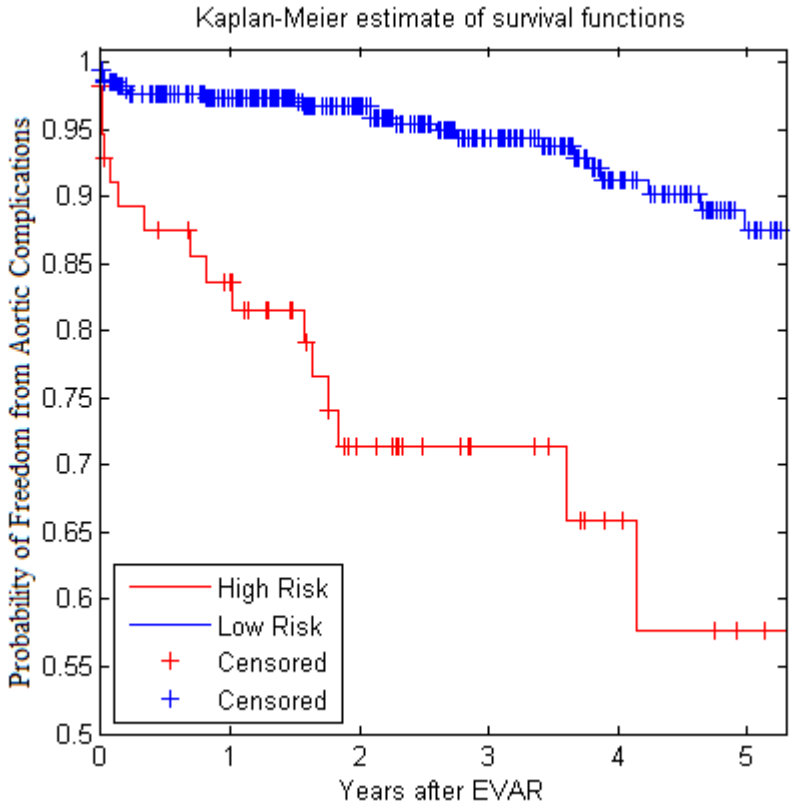
Figure 4.7 Kaplan-Meier curves of each risk group of Center 1 predictions using the final selected model of the proposed feature selection algorithm. This explains that the two risk groups of patients are separable, as the p-value of the log rank test is lower than 0.00001 which is below the significance level of 0.05.

Figure 4.7 which is the Kaplan Meier curves of each risk group of Center 1 predictions using the final selected model of the proposed feature selection algorithm. This shows that the proposed algorithm classified 169 of the center 1 patients as high risk which is equivalent to 37% of the patients. Freedom from aortic complications of the high risk patients reached 58.7% vs. 94.4% in low risk patients at year five as shown in figure 5 ($p<0.00001$ log-rank test). However, the AIC and BIC Cox's model predicted 104 and 58 patients as high risk as shown in figure 4.5 and 4.6 which is equivalent to 23% and 13% of the patients. Freedom from aortic complications of the high risk patients reached 52% vs. 93% and 57.6% vs.87.4% in low risk patients at year five ($p<0.00001$ log-rank test).



Year		0	1	2	3	4	5
Freedom from Aortic Complications	**Low-risk**	-	98.5%	97.7%	96.7%	94%	93%
	High-Risk	-	86%	79%	73.4%	66.73%	52%
Number at Risk	**Low-risk**	353	284	211	144	90	42
	High-Risk	104	77	51	32	16	9

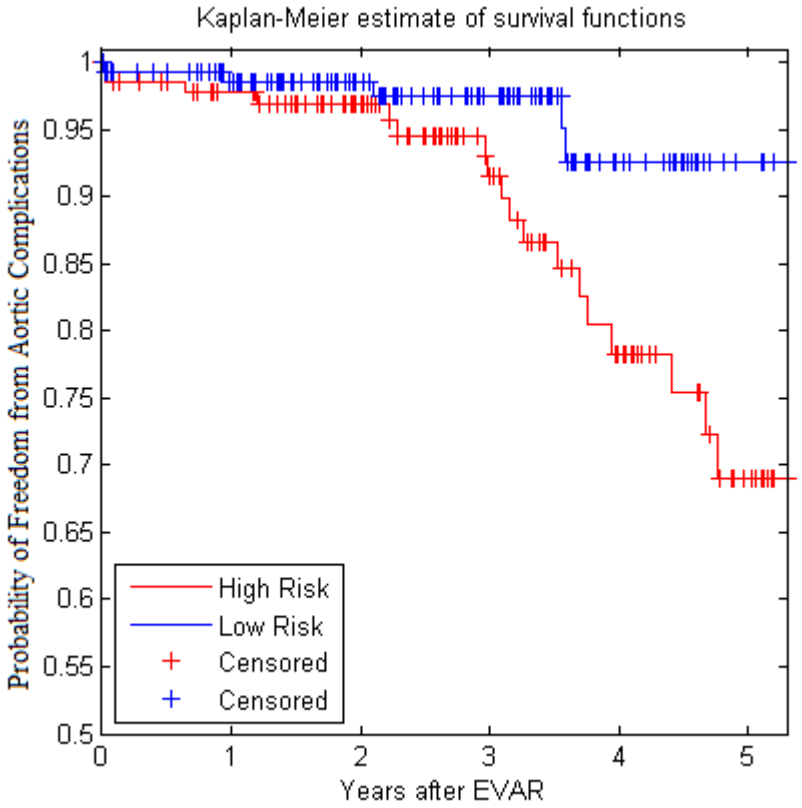
Figure 4.8 Kaplan- Meier curves of each risk group of Center 1 predictions using Cox's model with AIC. This explains that the two risk groups of patients are separable; as the p-value of the log rank test is lower than 0.00001 which is below the significance level of 0.05.



Year		0	1	2	3	4	5
Freedom from Aortic Complications	**Low-risk**	-	97.4%	96.7%	94.3%	91.2%	87.4%
	High-Risk	-	83.6%	71.4%	71.4%	66%	57.6%
Number at Risk	**Low-risk**	399	318	238	161	97	48
	High-Risk	58	43	24	16	9	5

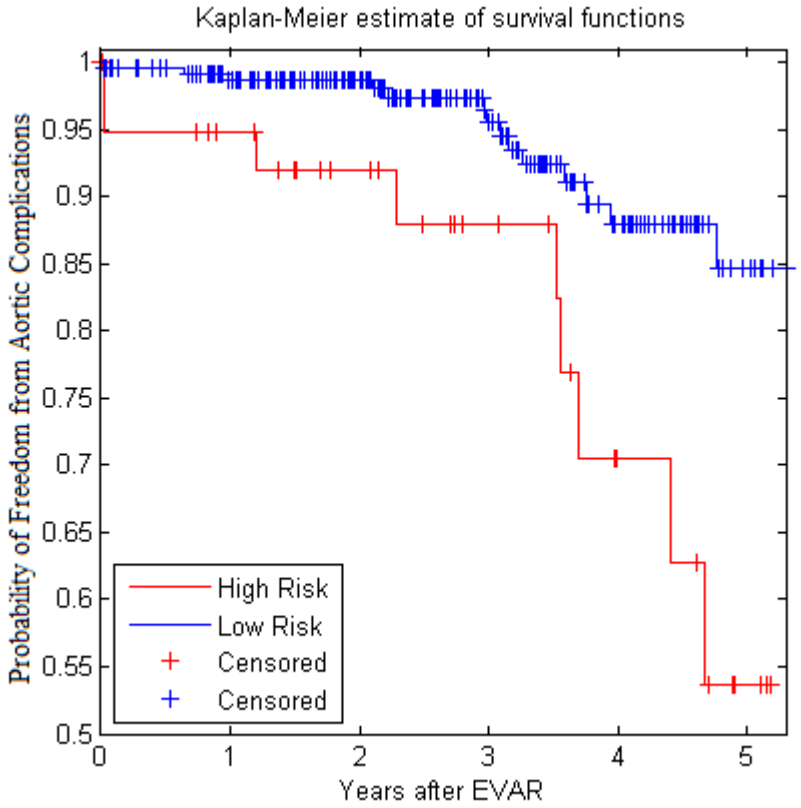
Figure 4.9 Kaplan -Meier curves of each risk group of Center 1 predictions using Cox's model with BIC. This explains that the two risk groups of patients are separable; as the p-value of the log rank test is lower than 0.00001 which is below the significance level of 0.05.

Figure 4.10 also shows that the proposed algorithm classified 136 of the Center 2 patients as high risk which is equivalent to 47.5% of the patients. Freedom from aortic complications of the high risk patients reached 69% vs. 92.6% in low risk patients at year five (p=0.022 log-rank test). However, the AIC and BIC Cox's models predicted 25 and 41 patients as high risk as shown in Figures 4.11 and 4.12 which is equivalent to 8% and 14.3% of the patients. Freedom from aortic complications of the high risk patients reached 39.5% vs. 83% and 53.7% vs.84.6% in low risk patients at year five (p=0.034 and 0.029 log-rank test).



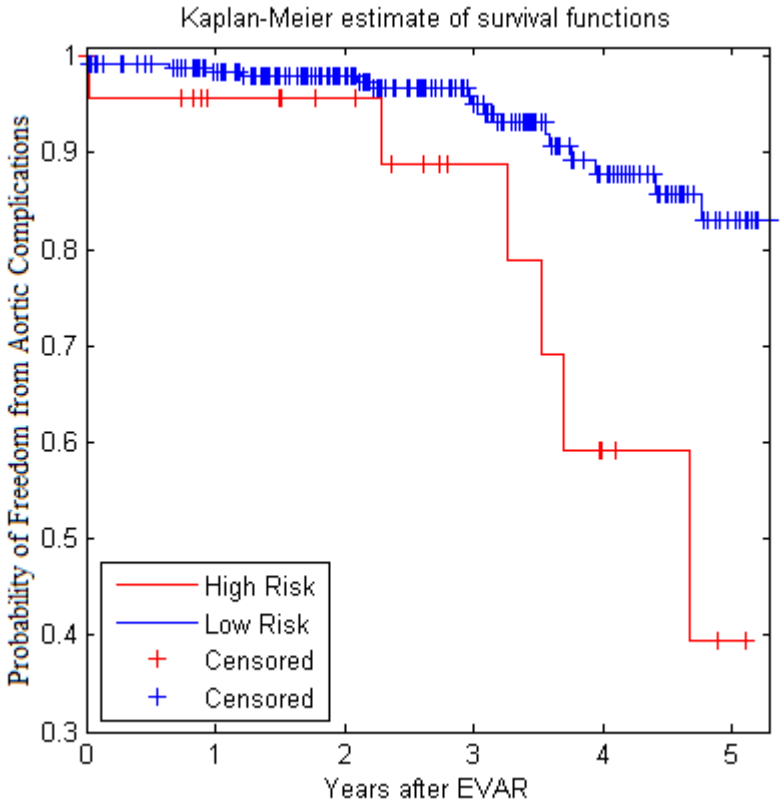
Year		0	1	2	3	4	5
Freedom from Aortic Complications	**Low-risk**	-	98.5%	98.5%	97.5%	92.6%	92.6%
	High-Risk	-	97.7%	97%	91.5%	78.3%	69%
Number at Risk	**Low-risk**	150	128	93	60	27	5
	High-Risk	136	120	91	60	35	10

Figure 4.10 Kaplan- Meier curves of each risk group of Center 2 predictions using the final selected model of the proposed feature selection algorithm. This explains that the two risk groups of patients are separable; as the p-value of the log rank test is lower than 0.022 which is below the significance level of 0.05.



Year		0	1	2	3	4	5
Freedom from Aortic Complications	**Low-risk**	-	98.7%	98.7%	95.5%	88%	84.6%
	High-Risk	-	94.8%	92%	88%	70.5%	53.7%
Number at Risk	**Low-risk**	245	214	159	102	53	15
	High-Risk	41	34	25	18	10	1

Figure 4.11 Kaplan Meier curves of each risk group of Center 2 predictions using Cox's model with AIC. This explains that the two risk groups of patients are separable, as the p-value of the log rank test is 0.034 which is below the significance level of 0.05.



Year		0	1	2	3	4	5
Freedom from Aortic Complications	**Low-risk**	-	98.4%	98%	95%	87.8%	83%
	High-Risk	-	95.6%	95.6%	88.5%	89.2%	39.5%
Number at Risk	**Low-risk**	261	230	169	111	58	15
	High-Risk	25	19	15	9	4	1

Figure 4.12 Kaplan Meier curves of each risk group of center 2 predictions using Cox's model with BIC. This explains that the two risk groups of patients are separable; as the p-value of the log rank test is lower than 0.029 which is below the significance level of 0.05.

4.7 Clinical Findings

The influence of aortic morphology on long term prediction of EVAR is complicated and suitable to be analyzed with ANN, with significant possible interface between aortic volume, aortic shape, aortic diameter, and aortic angulation. Current approaches have shown that aneurysm diameter predicts re-intervention after EVAR [274, 275], however evidence also recommended that other features of aneurysm morphology have influence to long-term clinical success [276-278]. Further complex concerns such as endograft configuration and deployment,

or intermediate markers of patients' cardiovascular risk phenotype, could possibly be used to train ANN in prospective studies which increase the clinical significance of prediction and make it more reliable. Also, adding of more operative factors such as graft size, and endoleak at completion or post-operative variables such as endoleak at early surveillance scans could enhance significantly the discriminatory power of ANNs [25]. The proposed feature selection approach based on individual ANN has selected the maximum aneurysm neck diameter, diameter of the left common iliac artery 1 and 5mm below internal iliac ostium, maximum common iliac artery diameter 5mm proximal to internal iliac origin, maximum iliac tortuosity index, maximum common iliac thrombus volume, and right common iliac artery non luminal volume. These features were examined by clinicians who approved that they have outstanding validation terms for the prediction morphology for current endografts available. Results were compared with current clinical method such as; SGVI (St George Vascular Institute), which showed that the proposed method has superior performance. The seven features ANN of the proposed method has greater predictive ability in classifying low and high risk patients than the SGVI score [263]. Concordance index and p-value of the predictive model using the selected features shows good clinical sense. In addition, the sensitivity rates are very promising compared to SGVI score, as it potentially indicates an increase in the event detection (EVAR re-intervention in this study) without affecting the cost of collecting unnecessary additional variables and surveillance's cost.

4.8 Conclusion

The proposed feature model selection technique was capable of building and validating a predictive reduced model using the two datasets collected during 2004 to 2010 from two vascular centers. The final reduced model was able to predict the long-term risk of aortic complications after EVAR. This will help clinics to put a future follow up surveillance plan for different risk groups EVAR patients.

This work has four challenges; the first one is to reduce instability and improve consistency during the feature selection and prediction processes, and this was resolved using the iterated nested cross validation method. Second is to apply feature selection techniques to highly censored data such as the EVAR survival data. Third, is to use machine learning techniques such as Bayesian networks and Artificial Neural networks to solve censoring issue and build a predictive model. The last challenge is to construct a predictive model using Center 1 EVAR data and test it using another data collected from center 2 to justify the model in future cross center applications.

The proposed algorithm was capable of overcoming all these challenges. It used a Bayesian network to uncensor the highly censored EVAR datasets and an ANN for prediction the risks. It

employed factor analysis reduction method and stepwise feature selection to reduce the number of features used for building the predictive model. It used cross validation for model selection (choosing the number of neurons of ANN) and iterated nested cross validation to generate a stable feature selection algorithm and produce the final reduced model of seven features only instead of the full model of 45 attributes. The final reduced model was validated using Center 2 data and the results showed it was capable of predicting the risk of EVAR re-intervention. It was also able to successfully distinguish between the two risks groups of each center as the p-value of the log rank test was lower than 0.00001 for Center1 and 0.022 for Center 2. This proves that the model can be used in cross-center predictions.

Two popular feature model selection techniques AIC and BIC were compared to the proposed algorithm. Both methods are employed with the most well-known predictive model used in survival analysis known as Cox's proportional hazard model. The same stepwise searching strategy used in the proposed algorithm is applied to the two methods. The final

The results showed that the proposed algorithm, AIC and BIC Cox's model distinguished between risk groups of Center 1 patients as the p-value was lower than 0.00001 which is beyond the significant level 0.05. The concordance index of AIC and BIC Cox's models are 0.7897 and 0.7622 which is greater than the proposed algorithm 0.7407. However, the proposed model has only seven features while the AIC and BIC cox's models have 14 features. Moreover, the sensitivity of the proposed algorithm is 0.77 which is better than 0.69 and 0.38 of the AIC and BIC Cox's models. This means that it could predict more patients that experienced the EVAR operation which is the event of interest in our study and this shows the advantage of using the proposed algorithm over the other two methods in truly predicting the risk of re-intervention. This also appears clearly in the predictions of Center 2. It expressed better sensitivity of 0.73 than the AIC and BIC Cox's models which are 0.35 and 0.23 respectively. Additionally, it has lower p-value of 0.022 than the other two algorithms 0.034 and 0.029. It has CI of 0.630 which is the same as the BIC model, but it outperforms the AIC with CI equals to 0.610. This makes the proposed algorithm very attractive for cross-center applications.

The number of patients of Center 1 that were classified as high risk using the proposed method, AIC and BIC Cox's models are 169, 104, and 58 patients versus 136, 25 and 41 for Center 2. This means the proposed algorithm identifies the risk of EVAR re-intervention better. Therefore, it will enable clinicians to put a more regular monitoring schedule in the future follow up and surveillance plan for those who have high risk of needing re-intervention, and the lower risk patients can be monitored less regularly. This would help in balancing and developing a cost-effectiveness surveillance system.

Chapter 5: Multiple Classifiers System for Predicting the Risk of Endovascular Aortic Repair Re-intervention through Hybrid Feature Selection

5.1 Introduction

Lately, the idea of using multiple classifiers system (MCS) is of interest to many researchers in machine learning field. Wolpert stated in [134] that there is no classifier is suitable for solving all classification tasks; because each one is ideal in specific area [14]. Therefore; combined classifiers is essential. It merges the outputs of multiple classifiers using a fuser in order to improve predictions. They can be used with feature selection to produce a reduced predictive model. Hybrid feature selection method combines advantages of feature selection classes. Merging hybrid feature selection with multiple classifiers system (MCS) usually improves prediction. Though, care must be made to prevent generating of unstable models in which predictions are extremely affected by any change in training data used to build it [135]**.** In this chapter, a hybrid feature selection technique was used to reduce the dimension of two highly censored EVAR data. Feature selection was done in two phases; first stage feature selection (FSFS) and iterated feature ranking (IFR). A Bayesian network was employed to the solve censorship issue. Moreover, MCS were constructed using (SVM, MLP, and KNN) to produce a stable reduced predictive model able to predict the risk of aortic re-intervention. A cost-effectiveness surveillance system is established consequently. It will allow doctors to select the appropriate follow up observation plan. MCS predictions were compared with single classifier performances. They were also compared with Cox's model combined with AIC, BIC, and LASSO.

5.2 Hybrid Feature Selection

Feature selection methods are divided into filter, wrapper, and embedded. However, recently, many researches focused in merging two or more classes of feature selection techniques to form a new class of feature selection known as hybrid feature selection. The main reason for doing that is that the hybrid method is capable of joining the benefits of these feature selection approaches. It also enables the production of a reduced predictive model. Therefore, filter and wrapper FS classes are combined together in this chapter to produce a hybrid feature selection method capable of featuring the important risk factors related to re-intervention risk. The results section will show that combining both of them will improve the prediction performance as well as reduce the final number of features selected.

5.3 Multiple Classifier System

MCS is a hybrid technique that combines the predictions of several classifiers mixed together in a particular way. Wolpert's theorem [134] mentioned that there is no particular classifier that can be used for all pattern recognition problems. MCS gathers powers of each learning algorithm in order to outperform the performance of each single classifier. In medical field, it is equivalent to taking the opinion of several doctors to reach a more confident final decision. Nevertheless, sometimes MCS results do not compete with the performance of the best individual classifier in the pool. However, it prevents the chance of poor decisions that might be taken with a particular inappropriately chosen model. For that reason, an MCS was proposed in this chapter which combines the strengths of three popular machine learning classifiers. Hybrid feature selection was combined with it in order to produce a reduced predictive model capable of predicting the risk of EVAR re-intervention. The prediction performances was compared with individuals ones. The results that will be discussed later show that the MCS performance beats that of individual ones. Prediction results of the proposed method outperformed the prediction using the state of art survival variable selection methods used with the common statistical survival method (Cox's regression model).

The MCS has two topologies: serial and parallel. The latter is the most common way used to connect classifiers [141] so it was adopted in this chapter. MCS methods includes; bagging, boosting, stacking, and voting. Generally, most of them produce slightly different performance degrees [146, 153]. However, many researchers prefer majority voting fusion algorithm due to its simplicity [147, 148]. Therefore, unweighted and weighted majority voting techniques were used in this thesis.

5.4 The Hybrid Feature MCS Proposed Approach

5.4.1 Similarity with the Previous Proposed FS Approach

The algorithm consists of seven steps. Steps 1, 2, 4, and 5 are similar to those discussed in Chapter 4. The main difference here appears in step 6 in which the MCS was proposed instead of a single ANN classifier. Moreover, in step 7, a feature ranking method was presented to be combined with a wrapper feature selection method. In the previous chapter, only one reduced model was chosen as a predictive model. However, instead in this chapter, the variables of all reduced models produced during the model selection process in step 3, were ranked according to their frequency distribution. Then, this ranking is used later to further reduce the number of selected features.

The four similar steps are feature reduction using factor analysis, cross validation and permutation, uncensoring, and iterated nested cross validation. For more details about them

refer to the previous chapter. Each shuffled version of the outer training fold in step 2 will pass through the first stage feature selection procedure (FSFS) in step 3. This stage is divided into two rounds; feature elimination and addition. Afterwards, the uncensoring step (step 4) deals with censoring of subsets generated in FSFS. Next, the generated subsets are re-split again into five more inner nested folds in step 5. Every four inner folds are used for constructing the multiple classifier system based on simple (unweighted) and weighted majority voting which is step 6, while the remaining one is used to test it. The average of the p-value of the log-rank test of the predictions was calculated and chosen as a criterion for feature selection. This procedure is called iterated nested cross validation which produces a stable model and overcome over-fitting that might have occurred later. The three areas of contributions in the proposed approach (feature selection, uncesoring, and classification) are shown in Figure 5.1 along with their interactions.

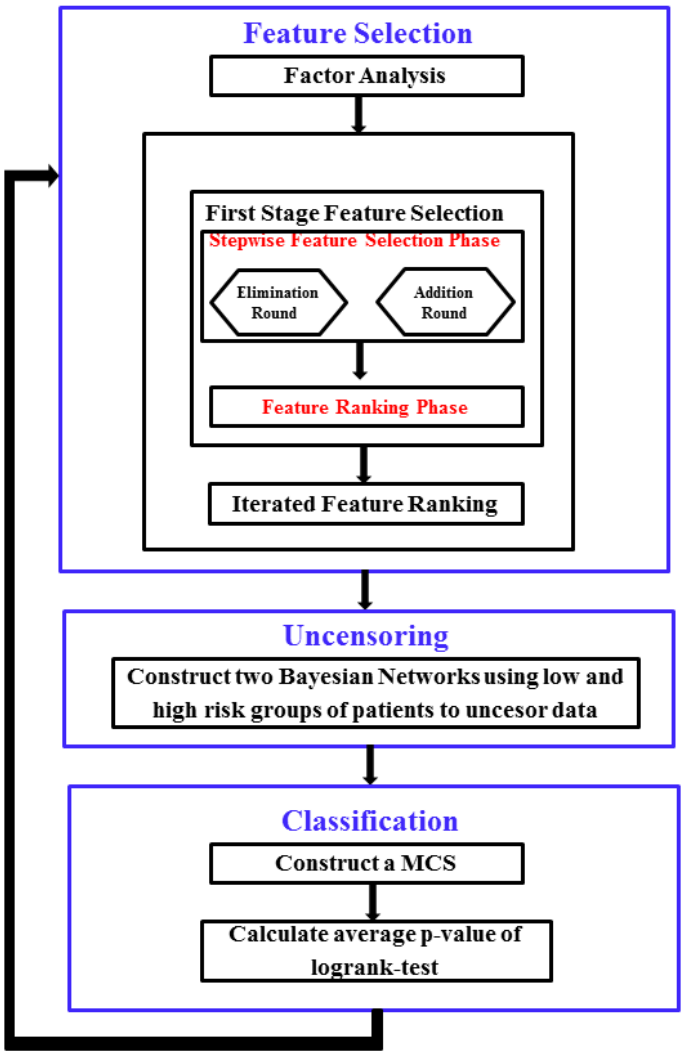


Figure 5.1 The three main contributions of the proposed hybrid feature selection approach (feature selection, uncensoring, and classifications) and their interaction.

5.4.2 The Proposed MCS based on Unweighted and Weighted Majority Voting Methods

The MCS system was constructed using three popular machine learning classifiers which are support vector machine (SVM), multiple layer perceptron (MLP) neural network, and K-nearest neighbor (KNN) classifiers. SVM tries to map the feature space using the kernel function into new feature space which facilities classification. Both SVM and MLP neural networks are well known as being strong classifiers. Moreover, they can detect the complex and high nonlinearity relations available in the datasets [148, 279, 280]. They have been widely used in medical applications [148, 281, 282]. K-nearest neighbour is a simple, straight forward and highly efficient classifier even with noisy data. K refers as to the number of neighbours. In order to perform classification for the current instance to be classified, it starts searching for the nearest K neighbors in the feature space then decides the class value according to the majority of the class label in its K nearest neighbours. Details of KNN classifier could be found in [283]. Though, it is simple, it has been shown to be appropriate and has good performance in medical application [284-288]. Sigmoid function was employed for SVM construction. A three layer MLP-ANN was constructed with 7 hidden and 2 output neurons, and sigmoid activation functions. KNN was built by Euclidean distance function and setting K to 3 which showed good performance.

Predictions were first combined with simple (unweighted) majority voting which simply gives a final decision to the class which has the majority of the votes. This approach usually improves predictive results, however it treats all classifiers equally; it does not put attention to classifiers that has higher impact on classification and generalization. Therefore, an improvement was done to overcome this which is known as weighted majority voting which was used as well in this thesis to enhance the prediction of individual classifiers. It allows each classifier in the pool to have a weight equivalent to its performance. Higher weights are given to those which have greater contribution on prediction results. The total weights should be equal to one in order to construct a proper weight distribution. Prediction of a new instance is made by multiplying the prediction of each classifier by its weights, then adding them to select the class with majority voting using (5.1), where; $c_{i,j}$ is the class value for the i^{th} classifier and j^{th} patient, N is the total number of classifiers, w_i is the weight for the i^{th} classifier,

$$Decision = \sum_{i=1}^{N} c_{i,j} \times w_i. \tag{5.1}$$

The idea here is how to determine the weights given to each classifier. Several methods have been proposed to calculate them which are beyond the focus of this thesis; however the most common approach depends on the training errors of each classifier. The weight is usually the

reciprocal of this error. Though, in this thesis the average of the p-value (P) of the log-rank test for the training data was chosen instead due to censoring nature of the datasets. Since, the average of the p-value for the training sets has value close to zero, their reciprocal will be very large, and therefore, the logarithm of the reciprocal average P is used to calculate the weight of each classifier in the pool as shown in Equation (5.2). These weights are then normalized in order that their sum is equal to one. Note that *"i"* is the classifier order in the pool.

$$w_i = \frac{1}{avg(P_i)}. \tag{5.2}$$

5.4.3 The Proposed Hybrid Feature Selection Approach

Step 3, is the first stage feature selection (FSFS) step which is done in two phases; stepwise feature model selection (SWFMS), and feature ranking (FR). In the former, each outer training fold uses stepwise searching strategy that switches between backward and forward searches to reduce the number of features. It starts with eliminating one variable iteratively. Each eliminated variable is inserted in a subset called 'visited". The features in this subset will be given another chance to re-enter the search space once again. After adding or deleting a variable from every outer training fold, it is shuffled and re-split five times to get the average of the p-value of predictions which is the criterion for feature selection. The model with the smallest average p-value is the one chosen. This is repeated until all variables are visited. Five outer reduced models will be generated at the end of this stage. Usually, in model selection only one model is chosen to win. However choosing one model may not pay attention to the uncertainty in all or some of the candidate models. Therefore, in this thesis all variables appeared in the five models were used in the FR phase and ranked according to their frequency distribution.

Finally, the ranking of the feature ranking phase is used in the iterated feature ranking (IFR) step which is step seven to further reduce the number features used in the predictive model. It starts with the variable of highest score, and then added one variable iteratively in order to enhance predictions. Both FSFS and IFR steps used the minimum p-value of the log rank test as a criterion for selection. Steps 3 and 7 are considered as a hybrid feature selection approach. It combines the advantages of filter and wrapper FS methods. The flowchart which is shown in Figure 5.2 discusses the steps of the proposed hybrid feature selection approach.

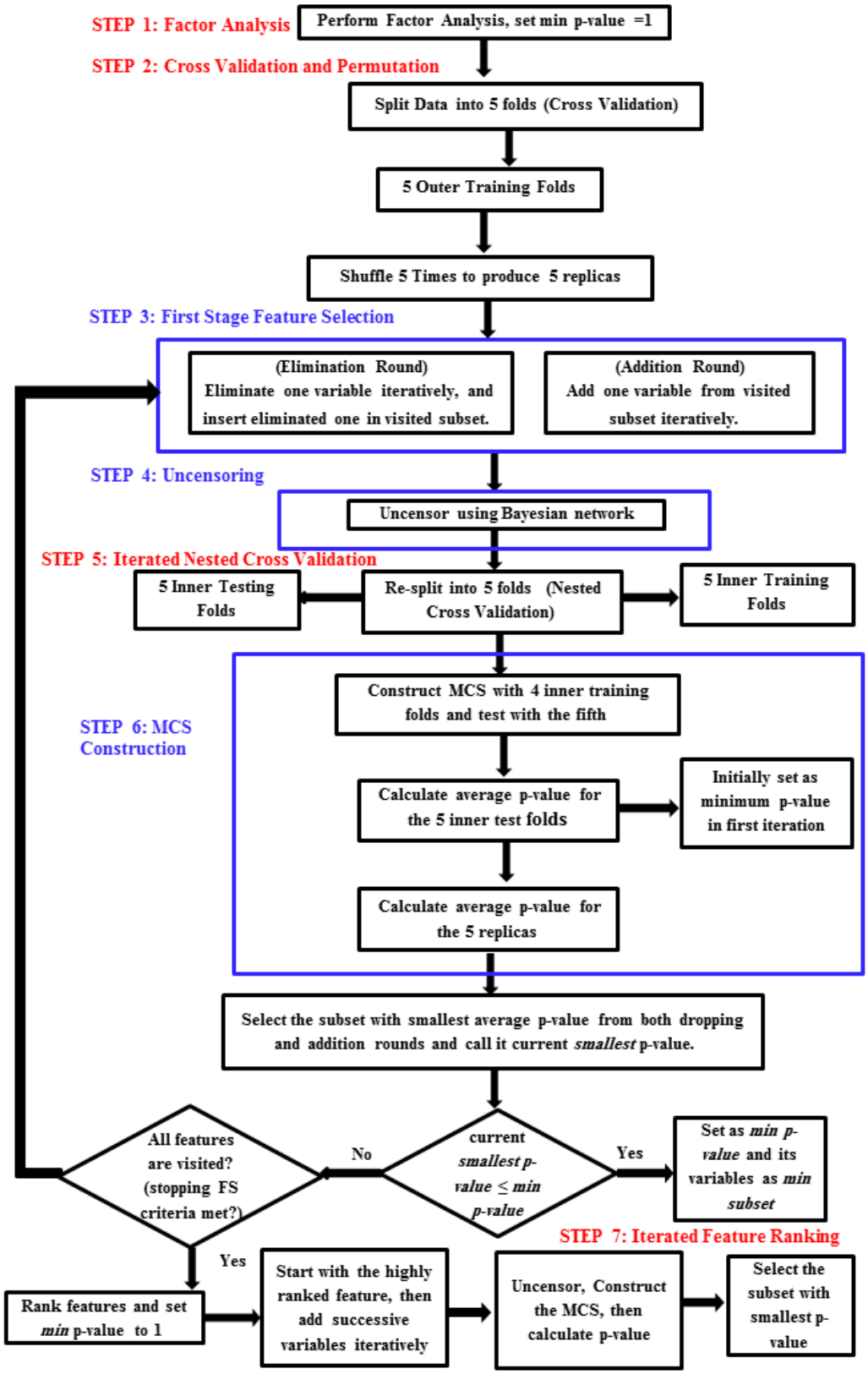


Figure 5.2 A flowchart discussing the seven steps of the proposed hybrid feature selection technique.

The Bayesian networks constructed using patients of the low and high risk groups after the proposed feature selection are shown in Figures 5.3-5.6.

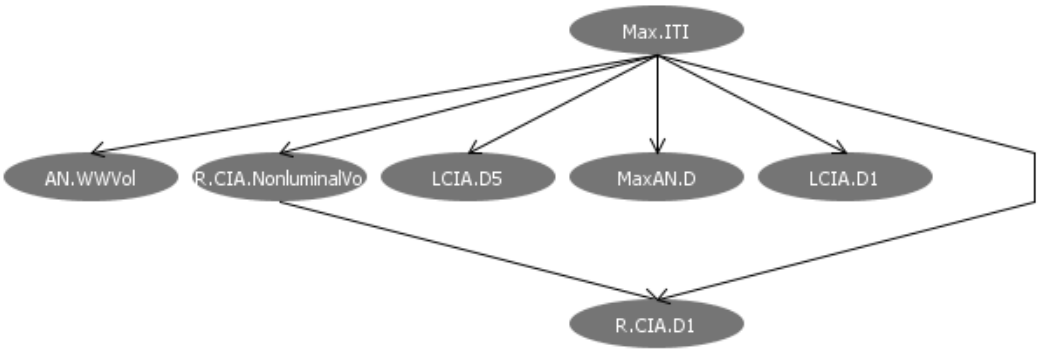


Figure 5.3 The low risk Bayesian Networks constucted using patients of high risk group after the proposed hybrid feature selection constructed on multiple classifier system combined with simple majority voting.

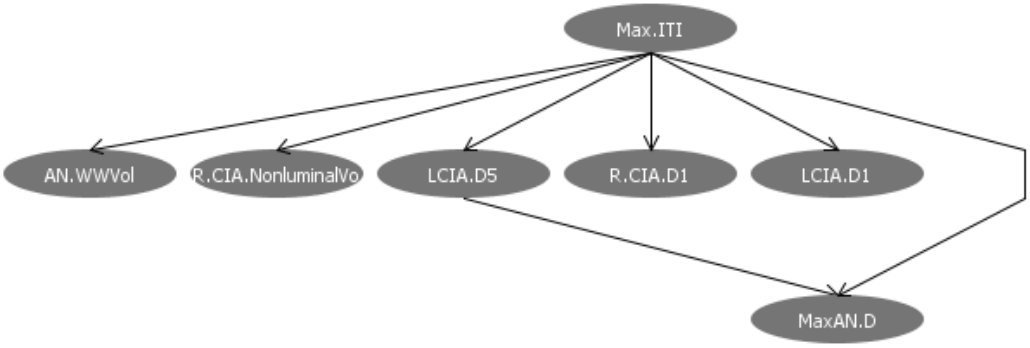


Figure 5.4 The high risk Bayesian Networks constucted using patients of low risk group after the proposed hybrid feature selection based on multiple classifier system combined with simple majority voting.

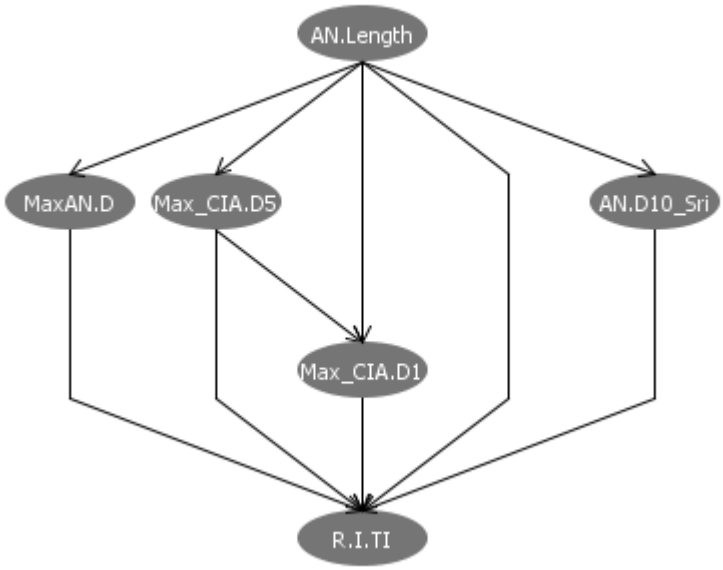


Figure 5.5 The high Bayesian Networks constucted using patients of high risk group after the proposed hybrid feature selection based on multiple classifier system combined with weighted majority voting.

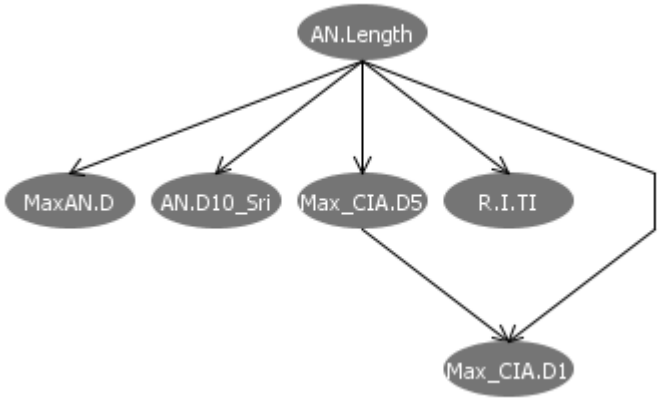


Figure 5.6 The low Bayesian Networks constucted using patients of low risk group after the proposed hybrid feature selection based on multiple classifier system combined with weighted majority voting.

5.5 Results of the Proposed MCS Hybrid Feature-Model Selection

5.5.1 Comparing the Results of the Proposed MCS Hybrid Feature-Model Selection Algorithm with all Features

In this chapter, a fivefold nested cross validation was used for the hybrid feature selection and the stable MCS model construction using Center 1. Center 2 was used to assess the performance of the final reduced model. The results of the MCS hybrid feature selection based on unweighted and weighted majority voting techniques for Center 2 predictions are compared with the full size of the model as shown in Tables 5.1 and 5.2.

Table 5.1 Results of the proposed MCS using unweighted majority voting on the testing set (Center 2) after the two steps of hybrid feature selection. This shows that the proposed method has reduced the number of features and p-value of the log-rank test, and enhanced the concordance index and sensitivity after factor analysis and the two steps of feature selection.

Proposed algorithm	Number of features	p-value (Log rank test)	CI (SD)	Sensitivity
MCS All Features	45	0.0331	0.6599 (0.0635)	0.423
MCS FA step	27	0.0166	0.6630 (0.0571)	0.461
MCS FSFS step	15	0.0075	0.6657 (0.0732)	0.654
MCS IFR step	7	0.00016	0.6793 (0.0556)	0.808

Table 5.1 shows that the proposed multiple classifier system hybrid feature selection technique based on unweighted voting has reduced the number of features from 45 to 27, 15 and 7 after all the steps of proposed approach. Moreover, the concordance index (CI) of the full model is 0.6599 which is increased to 0.6630, 0.6657, and finally 0.6793 after the hybrid FS steps. The p-value of the log-rank test has been reduced as well from 0.0331 to 0.0166, 0.0075 and 0.00016 after all steps of the proposed technique, which indicates an enhancement in the performance of the MCS model with the hybrid FS. Finally, the sensitivity was enhanced during all steps of the hybrid approach from 0.423 to reach finally 0.808.

Table 5.2 shows that the proposed MCS hybrid FS approach based on weighted majority voting has reduced the number of features from 45 to 27, 17 and 6 after all the steps of proposed approach. Moreover, the concordance index (CI) of the full model is 0.6710 which is increased to 0.6762, 0.6793, and finally 0.6808 after the hybrid feature selection steps which are greater than that of the unweighted majority voting in Table 5.1 (0.6599,0.6630, 0.6657, and 0.6793).

The p-value of the log-rank test has been reduced as well from 0.014 to 0.001, 0.0008 and 0.000038 after all steps of the proposed technique, which indicates an enhancement in the performance of the MCS model based on weighted voting with the hybrid FS compared to unweighted majorty voting which has reach a final p-value of 0.00016. In addition, the sensitivity has increased from 0.423 to reach 0.7308.

Table 5.2 Results of the proposed MCS using Weighted Majority Voting on the testing set (Center 2) after the two steps of hybrid feature selection. This shows that the proposed method has reduced the number of features and p-value of the log-rank test, and enhanced the concordance index and sensitivity after factor analysis and the two steps of feature selection.

Proposed algorithm	Number of features	p-value (Log rank test)	CI(SD)	Sensitivity
MCS All Features	45	0.014	0.6710 (0.0572)	0.423
MCS FA step	27	0.0010	0.6762 (0.0643)	0.539
MCS FSFS step	17	0.0008	0.6793(0.0573)	0.615
MCS IFR step	6	0.000038	0.6808 (0.0528)	0.7308

5.5.2 Comparing the Results of the Proposed MCS Hybrid Algorithm with the Performance of the Individual Classifiers

In this section, the performances of the MCS hybrid feature selection algorithm and individual classifiers used to construct it are compared. As shown in Table 5.3 MCS based on unweighted and weighted majority voting and SVM,MLP, and KNN single classifiers have reduced the feature space to 7,6,5,5,6 for MCS based on unweighted majority voting, MCS based on weighted majority voting, SVM, MLP, and KNN models respectively. Predictions of Center 2 are used for comparison as Center 2 was not used in constructing and training the predictive model. The MCS based on weighted majority voting has outperformed the unweighted majority voting in both CI (0.6808 vs. 0.6793) and p-value (0.000038 vs. 0.00016); however, the latter has higher sensitivity (0.808 vs.0.7308). Moreover, the MCS hybrid FS approach using unweighted and weighted majority voting methods outperformed the other individual classifiers in p-value (0.00016 and 0.000038 vs. 0.00085, 0.00073, and 0.0011). Though, the MLP's CI (0.6817) is better than MCS, SVM, and KNN (0.6793 and 0.6808, 0.6776, and 0.6411).

Table 5.3 Performance of the proposed MCS approach compared with individual classifiers after hybrid Feature selection on testing set. This explains that the proposed approach based on weighted and unweighted majority voting has outperforms the individual classifier performances except for the concordance index of the MLP classifier.

Classifier	Number of final features	p-value (Log rank test)	CI (SD)	Sensitivity
MCS Unweighted Majority Voting	7	0.00016	0.6793 (0.0556)	0.808
MCS Weighted Majority Voting	6	0.000038	0.6808 (0.0528)	0.7308
SVM	5	0.00039	0.6776 (0.0499)	0.7308
MLP	5	0.00073	0.6817 (0.0804)	0.7308
KNN	6	0.0011	0.6411 (0.0628)	0.6538

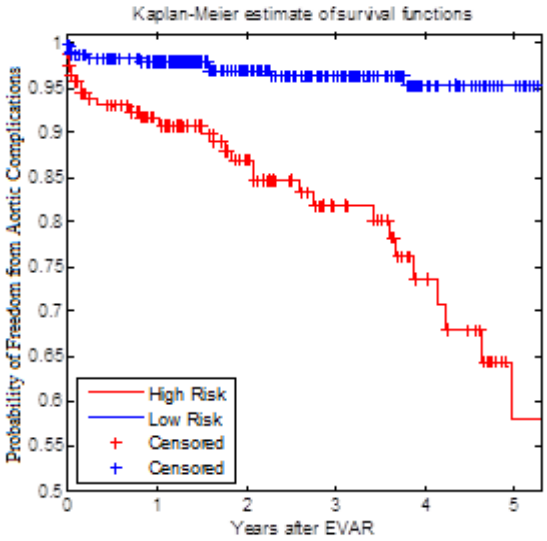
5.5.3 Comparing the results of the proposed MCS hybrid algorithm with performance of Cox's model using AIC, BIC, LASSO

In this section, the results of the MCS hybrid feature selection based on unweighted and weighted majority voting are compared with the state of art variable selection methods using Cox's regression model which are AIC, BIC and LASSO penalized methods. The estimated parameters of the final reduced model are multiplied by each variable to generate a risk score. The same thresholds used in the Chapter 4 were used. The one used for LASSO is 6.7 which is equivalent to the mean of the risk score. As shown in Table 5.4. The number of features of the final MCS model is seven for unweighted majority voting and six for weighted voting which are better than fourteen for AIC and BIC, but equal or smaller than seven of LASSO. For Center 1 prediction, the concordance index of MCS based on weighted majority voting (0.7881) which is higher than unweighted majority voting (0.7521), BIC (0.7624), LASSO (0.738), but smaller than AIC (0.7898). All models have p-value lower than 0.0001 which indicates that they are all capable of separating the two risk groups of Center 1. The sensitivity of MCS model using unweighted majority voting (0.84) and weighted majority voting (0.87) are greater than that of the other methods (0.69, 0.38, and 0.714). Moreover, for Center 2 predictions the proposed MCS technique beats the other techniques in both p-value (0.00016 and 0.000038 vs. 0.034, 0.029, and 0.0068) and CI (0.6793 and 0.6808 vs. 0.6103, 0.630, and 0.6153). The main advantage in the MCS hybrid FS algorithm appears in the sensitivity results (0.808 and 0.7308 vs. 0.35, 0.23, and 0.5) which indicates that it can correctly classify more patients that did the re-intervention (the event of interest in this study). Thus, it is favoured than the other methods.

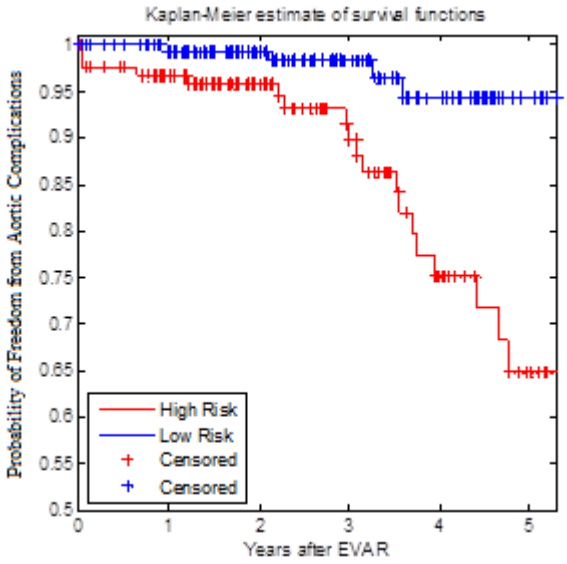
Table 5.4 Results of the proposed MCS after hybrid feature selection compared with Cox's model using AIC, BIC, and LASSO. This explains that the proposed approach based on weighted and unweighted majority voting has outperforms the performances of the methods based on Cox's model for Center 2 predictions.

Technique	Model Size	p-value (Log rank test)		CI (SD)		Sensitivity	
		Center 1	Center 2	Center 1	Center 2	Center 1	Center 2
Unweight Majority Voting MCS Hybrid FS	7	<0.0001	0.00016	0.7521 (0.0332)	0.6793 (0.0556)	0.84	0.808
Weighted Majority Voting MCS Hybrid FS	6	<0.0001	0.000038	0.7881 (0.0337)	0.6808 (0.0528)	0.87	0.7308
AIC Cox FS	14	<0.0001	0.034	0.7898 (0.0408)	0.6103 (0.0725)	0.69	0.35
BIC Cox FS	14	<0.0001	0.029	0.7624 (0.0465)	0.630 (0.0685)	0.38	0.23
LASSO Cox FS	7	<0.0001	0.0068	0.7382 (0.0426)	0.6153 (0.0864)	0.714	0.50

Figures 5.7 and 5.8 show the Kaplan-Meier curves for the two risk groups predictions of both Centers using the MCS hybrid FS technique based on unweighted and weighted majority voting compared with Kaplan-Meier curves for the two risk groups predictions of both centers with AIC (figure 5.9), BIC (Figure 5.10), and LASSO (Figure 5.11) Cox's models. Figure 5.7 indicates that the MCS model based on unweighted model classified 163 and 126 of Center 1 (upper) and Center 2 (lower) patients as high risk which is equivalent to 36 % and 44% of total Center 1 and Center 2 patients. Moreover, Figure 5.8 shows that the MCS model based on unweighted model classified 177 and 101 of Center 1 (upper) and Center 2 (lower) patients as high risk which is equivalent to 38 % and 35% of total Center 1 and Center 2 patients. Prediction results of the MCS model are better than the prediction of the AIC model (104 high risk patients equivalent to 23%) for Center 1 (Figure.5.9 upper) and (41 high risk patients equivalent to 14%) for Center 2 (Figure 5.9 lower), the BIC model in (58 high risk patients equivalent to 13%) for Center 1 (Figure 5.10 upper) and (25 high risk patients equivalent to 9%) for Center 2 (Figure 5.10 lower), and the LASSO model (196 high risk patients equivalent to 43%) for Center 1 (Figure 5.11 upper), and (76 high risk patients equivalent to 26%) for Center 2 (Figure 5.11 lower).

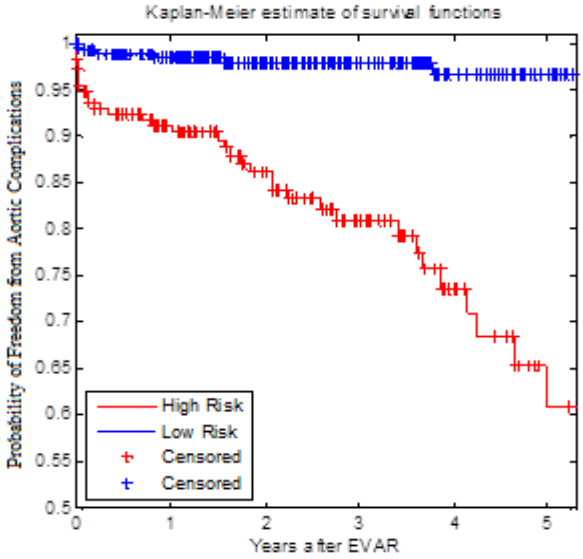


Year		0	1	2	3	4	5
Freedom from Aortic Complications	**Low-risk**	-	97.9%	97%	96.3%	95.3%	95.3%
	High-Risk	-	90.8%	87%	81.8%	73.6%	58.5%
Number at Risk	**Low-risk**	294	243	181	125	78	43
	High-Risk	163	118	81	51	28	9

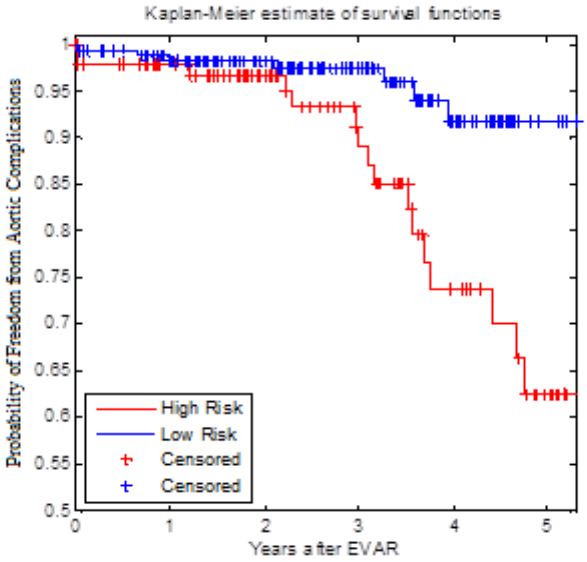


Year		0	1	2	3	4	5
Freedom from Aortic Complications	**Low-risk**	-	99.3%	99.3%	98.3%	94.3%	94.3%
	High-Risk	-	96.7%	95.8%	89.9%	75.2%	64.9%
Number at Risk	**Low-risk**	160	140	106	66	33	5
	High-Risk	126	108	78	54	29	11

Figure 5.7 KM curves for the two risk groups predictions of (upper) Center1 and (lower) Center 2 using the MCS hybrid FS technique based on unweighted voting. This explains that the two risk groups of patients are separable, as the p-values of the log rank test are lower than 0.00001for Center 1 and 0.00016 for Center 2 which is below the significance level of 0.05.

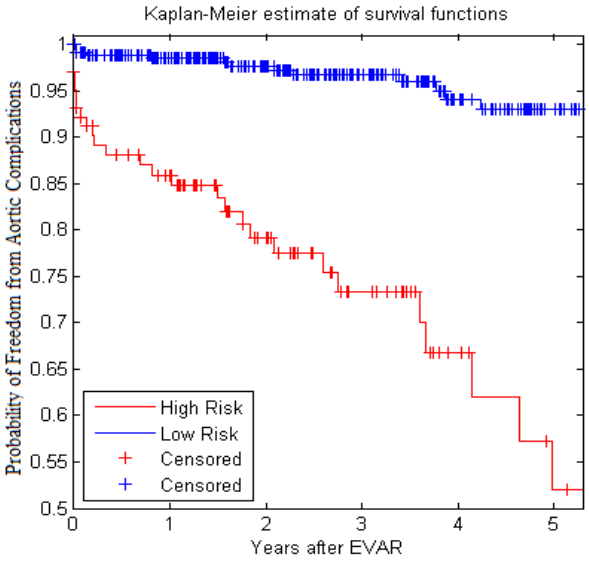


Year		0	1	2	3	4	5
Freedom from Aortic Complications	Low-risk	-	98.5%	97.8%	97.8%	96.8%	96.8%
	High-Risk	-	90.5%	86.2%	81%	73.6%	60.9%
Number at Risk	Low-risk	280	232	169	117	74	39
	High-Risk	177	129	93	59	32	14

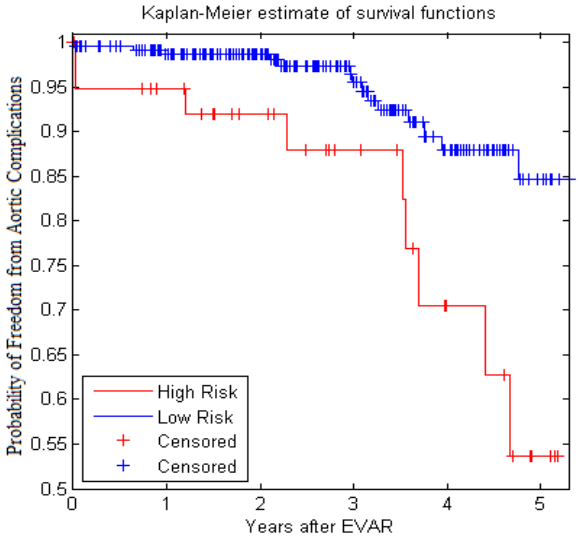


Year		0	1	2	3	4	5
Freedom from Aortic Complications	Low-risk	-	98.3%	98.3%	97.4%	91.8%	91.8%
	High-Risk	-	98%	96.8%	87.2%	73%	62.4%
Number at Risk	Low-risk	185	163	122	77	38	11
	High-Risk	101	85	62	43	24	6

Figure 5.8 KM curves for the two risk groups predictions of (upper) Center1 and (lower) Center2 using the MCS hybrid FS technique based on weighted voting. This explains that the two risk groups of patients are separable, as the p-values of the log rank test are lower than 0.00001for Center 1 and 0.000038 for Center 2 which is below the significance level of 0.05.

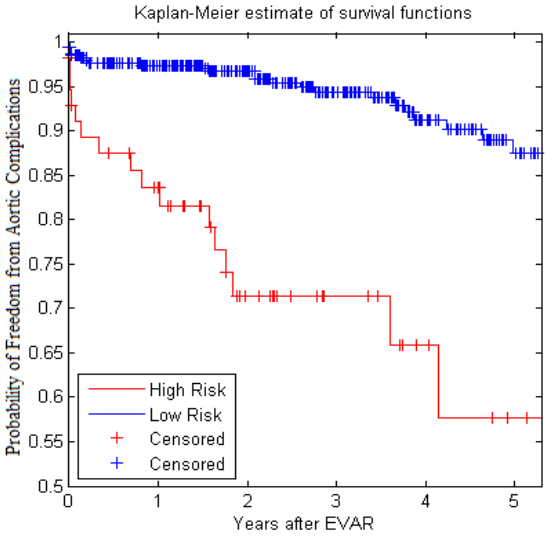


Year		0	1	2	3	4	5
Freedom from Aortic Complications	Low-risk	-	98.5%	97.7%	96.7%	94%	93%
	High-Risk	-	86%	79%	73.4%	66.73%	52%
Number at Risk	Low-risk	353	284	211	144	90	42
	High-Risk	104	77	51	32	16	9

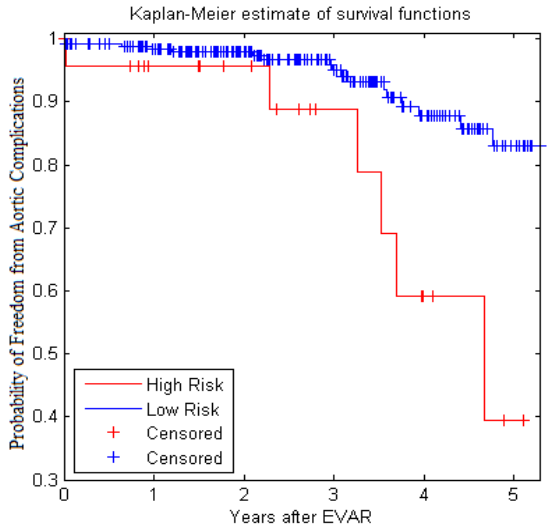


Year		0	1	2	3	4	5
Freedom from Aortic Complications	Low-risk	-	98.7%	98.7%	95.5%	88%	84.6%
	High-Risk	-	94.8%	92%	88%	70.5%	53.7%
Number at Risk	Low-risk	245	214	159	102	53	15
	High-Risk	41	34	25	18	10	1

Figure 5.9 KM curves for the two risk groups predictions of (upper) Center 1 and (lower Center2) using the AIC Cox technique. This explains that the two risk groups of patients are separable, as the p-values of the log rank test are lower than 0.00001for Center 1 and 0.034 for Center 2 which is below the significance level of 0.05.

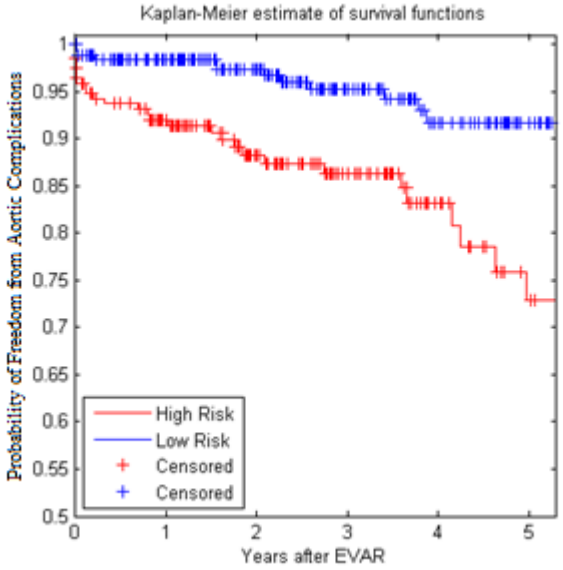


Year		0	1	2	3	4	5
Freedom from Aortic Complications	**Low-risk**	-	98.4%	97.3%	95.2%	91.7%	91.7%
	High-Risk	-	92%	88%	86.3%	83%	73%
Number at Risk	**Low-risk**	261	210	159	106	67	30
	High-Risk	196	151	103	70	39	23

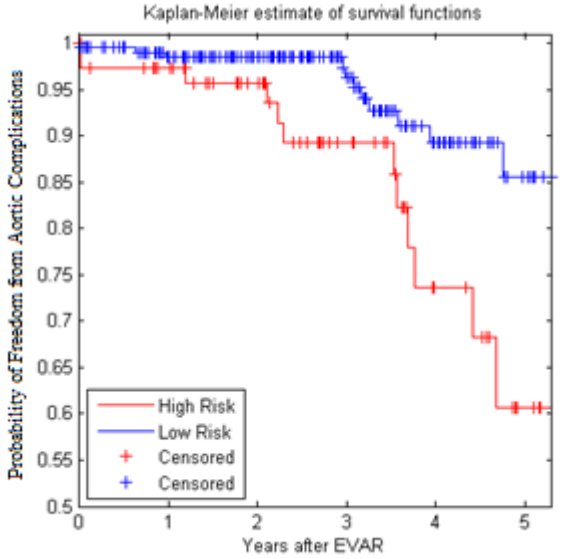


Year		0	1	2	3	4	5
Freedom from Aortic Complications	**Low-risk**	-	98.4%	98%	95%	87.8%	83%
	High-Risk	-	95.6%	95.6%	88.5%	89.2%	39.5%
Number at Risk	**Low-risk**	261	230	169	111	58	15
	High-Risk	25	19	15	9	4	1

Figure 5.10 KM curves for the two risk groups predictions of (upper) Center 1and (lower) Center 2 using the BIC Cox technique. This explains that the two risk groups of patients are separable, as the p-values of the log rank test are lower than 0.00001for Center 1 and 0.029 for Center 2 which is below the significance level of 0.05.



Year		0	1	2	3	4	5
Freedom from Aortic Complications	**Low-risk**	-	98.4%	97.3%	95.2%	91.7%	91.7%
	High-Risk	-	92%	88%	86.3%	83%	73%
Number at Risk	**Low-risk**	261	210	159	106	67	30
	High-Risk	196	151	103	70	39	23



Year		0	1	2	3	4	5
Freedom from Aortic Complications	**Low-risk**	-	98.5%	98.5%	96.3%	89.2%	85.5%
	High-Risk	-	97.3%	95.7%	89.2%	73.5%	60.7%
Number at Risk	**Low-risk**	210	182	134	89	47	15
	High-Risk	76	66	50	31	15	3

Figure 5.11 KM curves for the two risk groups predictions of (upper) Ccenter1 and (lower) Center 2 using the LASSO Cox technique. This explains that the two risk groups of patients are separable, as the p-values of the log rank test are lower than 0.00001for Center 1 and 0.0068 for Center 2 which is below the significance level of 0.05.

5.6 Clinical Findings

Features that were selected using simple (unweighted) majority voting are total aneurysm neck volume, maximum aneurysm neck diameter, diameter of the left common iliac artery 1 and 5 mm below internal iliac ostium, maximum iliac tortuosity index, diameter of the right common iliac artery 1mm below Internal iliac ostium, and right common iliac artery non luminal volume. Moreover, features resulted from weighted voting are maximum common iliac aneurysm area, aneurysm neck diameter 10 mm below lowest renal, aneurysm neck length, common Iliac Artery diameter 1 and 5mm proximal to internal iliac origin, and right iliac tortuosity index. These features were reviewed by the clinical investigators. They confirmed that these variables have good face validity in terms of predicting technically difficult or challenging morphology for endografts currently available. It is well known that hostile sealing zones both proximally (at the aortic neck) or distally (at the common iliac artery) pose considerable technical challenges for durable endograft seal, and therefore it is plausible that the features selected (aortic neck area; and various aspects of iliac morphology) might be predictive of poor long-term clinical performance. Predictions using these features are clinically feasible and make excellent sense. However, weighted majority makes more sense as it includes neck length which is often thought of by surgeons planning the case [278, 289, 290] . Moreover, the concordance index and sensitivity rates are very promising and would have clinical importance if used prospectively. Also, the assignment of most patients to a low risk group counts well with clinical practice in which less patients will have re-intervention over 5 years [291].

5.7 Discussion and Conclusion

In this chapter, a hybrid feature selection approach was proposed which merged filter and wrapper feature selection approaches. Combining several classifiers to construct a multiple classifier system usually enhances prediction results. Therefore, the proposed hybrid approach constructed a multiple classifier system and used it to evaluate features generated during feature selection. This new proposed approach consists of seven steps in which only four steps are similar to the approach previously proposed in Chapter 4. However, the difference in this chapter first appears in step 3 where feature ranking is used to rank features according to their frequency distribution. Then, in step 4 the MCS was used instead of a single ANN classifier. Furthermore, in step 7, iterated filter ranking feature selection method is used to further reduce features.

The proposed approach used a Bayesian network to deal with the highly censored EVAR datasets. It also constructed a multiple classifier system based on unweighted and weighted majority voting for predicting the risk of EVAR re-intervention. The proposed method employed factor analysis reduction method and a hybrid feature selection method to reduce the

number of features used for building the predictive model. Moreover, used cross validation and iterated nested cross validation to generate a stable algorithm, reduce bias, and avoid overfitting. External validation is important to produce a clinically validated predictive model that can be used for medical application. Therefore, the final reduced models were validated using Center 2.

The two datasets used in this thesis were capable of constructing a multiple classifier predictive model to predict the long-term risk of aortic complications after EVAR. The predictive model was combined with the proposed hybrid feature model selection approach to reduce the number of features needed to construct it. Moreover, it may be used for cross-centers prediction as well, as it was constructed and evaluated by patients of two different centers. The model will enable doctors to take decisions about future follow up observation plan for each patient. High risk patients will have to undergo more regular surveillance than low risk patients.

In the proposed technique, the instability that might occur during feature selection, model selection, and multiple classifier system construction was reduced using iterated nested cross validation. The uncensoring issue was solved using Bayesian networks. The multiple classifier system was constructed using three popular machine learning classifiers (SVM, MLP, and KNN) combined with unweighted and weighted majority voting. It was capable of predicting the risk of re-intervention after EVAR. Its performance was compared with both individual classifiers and the statistical Cox's model. Three well-known model selection techniques called AIC, BIC, and LASSO were used with cox's regression model for comparison with the MCS hybrid feature selection approach. The same searching strategy was used for the selection in AIC and BIC.

The results have shown that multiple classifier system using unweighted and weighted voting outperformed both individual classifiers and Cox's model selection methods in both p-values and concordance index expect for that of the MLP for Center 2. It successfully separated between the risks groups for both centers as the p-value of the log rank test was less than 0.0001 for Center1 and 0.00016 and 0.000038 for Center 2 using unweighted and weighted majority voting, In addition the CI has increased from 0.6559 and 0.6710 to reach finally 0.6793 and 0.6808 with sensitivity of 0.808 and 0.7308 which allow it to be used for cross center prediction. Moreover, the proposed technique has higher sensitivity compared to other techniques which make it stronger than the other ones in classifying the long term risk of aortic complications after EVAR for unseen patients. Therefore, it can be used by doctors to facilitate the future follow up decision plan. Patients with high risk prediction will be more closely monitored than other ones which reduces the chance of low risk patients to be exposed to harmful radiations.

Chapter 6: Conclusion, Clinical Findings, and Future Work

6.1 Conclusions

Censoring is the unique characteristic of survival data. It is the main cause why standard machine learning technique cannot be used directly with survival data type. Therefore, it should be handled. Different scenarios were proposed in the literature; however they have some drawbacks and cannot be used with a high level of censoring. This thesis introduced a modified approach that can deal with high levels of censoring. Machine learning techniques are preferred over standard statistical survival models due to various reasons. Therefore, the modified approach used machine learning techniques to handle censoring and to construct a survival predictive model. The model was capable of predicting the risk of re-intervention after EVAR surgery which will help doctors to decide the future follow up plan each patient should undertake.

Feature selection is vital for medical data, though its process becomes complicated with medical survival data due to the presence of censoring. Most of the work done for survival variable selection was based on Cox's proportional hazard survival model. Though, machine learning techniques are more favoured. For this reason, two feature selection approaches that use machine learning classifiers to evaluate the features selected and construct a predictive model were proposed. The two EVAR datasets used to construct and validate the two proposed feature selection approach were capable of validating a reduced predictive model to predict the long-term risk of aortic complications after EVAR. The results were compared with three well known survival variable selection methods based on Cox's model. The two proposed approaches outperformed the other methods in the p-value of log-rank test, concordance index, and sensitivity which makes it more powerful in predicting the long term aortic complication.

In Chapter 3, a new modified approach has been proposed to deal with the high level of censoring existing in the EVAR datasets without deleting, ignoring, weighting censored patients, or considering them as real zero targets through a Bayesian networks approach. Two Bayesian networks were constructed which were called high-risk and low-risk networks and they were used to determine to which risk group each censored patients belongs to. Afterwards, two multilayer perceptron neural networks were employed to predict the risk of EVAR re-intervention and test the efficiency of the algorithm on each dataset separately. Next, a third one was constructed using Center 1 dataset and employed to classify patients of Center 2 into high or low risk categories. The proposed approach successfully increases the AUROC for both

Centers. Moreover; it was able to predict the risk of REINT on censored Center 2 data after five years with the neural network model built with uncensored Center 1 dataset. The performance of the proposed approach was compared with another uncensoring approach called S & DB proposed in [6]. The results showed that the proposed approach prediction was more concordant (0.6340) than S & DB approach (0.593). In addition the predictive model of the proposed approach was capable of distinguishing between the low and high risk groups with a p-value of the log rank test equals to 0.0348 which is not the case for the other approach (p-value of the log rank test equals 0.2316).

In Chapter 4, four challenges motivated the proposal of a new feature selection approach. They are; instability of the model produced during the feature selection and predictions processes, the highly censored datasets, using machine learning methods to deal with this type of data, the construction of a multicenter predictive model that can be used in future cross-center applications. A new feature selection was proposed to deal with these challenges. It used the proposed uncensoring approach discussed in Chapter 3 to handle the highly censored EVAR datasets and predict the risk of EVAR re-intervention. Factor analysis feature reduction method was employed with greedy stepwise feature selection to decrease the complexity of the predictive model and reduced the number of variables. Cross validation was used to split data to perform model selection (choosing the number of hidden neurons of ANN). Nested cross validation was employed to produce stable models during feature selection and generate one final reduced model of only seven features instead of the full model consisting of forty five attributes. Center 1 was used for feature and model selection procedures. Center 1 was also adapted to construct the reduced predictive model. Center 2 was used to test and validate the reduced predictive model which avoids the overoptimistic and biased results that might occur. This model was also able to distinguish successfully between the two risks groups of each center as the p-value of the log rank test was lower than 0.00001 for Center 1 and 0.022 for Center 2. This proves that the model can be used in cross-center predictions.

Two popular survival variable selection techniques based on Cox's model such as; Akaike and Bayesian information criteria (AIC and BIC) were compared to the proposed algorithm. The reduced predictive model constructed using the proposed approach has higher ability in discriminating between the risk groups of patients than the AIC and BIC variable selection methods based on Cox's model, since it has a better p-value of the log-rank test, concordance index, and sensitivity. In addition, the number of patients that were classified as high risk using the proposed method, AIC and BIC Cox's models are 169, 104, and 58 respectively at Center 1 and 136, 25 and 41 at Center 2. This means that the proposed algorithm is better at identifying the risk of EVAR re-intervention. Therefore, it may be preferred by doctors to decide which future surveillance plan each patient should undertake. Clinicians will put a more regular

monitoring schedule in the future follow up for those who have high risk of needing re-intervention, and the lower risk patients can be monitored less regularly. This would help in balancing and developing a cost-effectiveness surveillance system.

In Chapter 5, a hybrid feature selection approach which combines the wrapper and filter variable selection classes was proposed. It consists of seven steps; it has four similar steps to that of the previous feature selection approach proposed in Chapter 4. However, two differences appear in this method. The first one is using an iterative feature ranking method to further reduce the variable selected in the feature selection process. In Chapter 4 only the variable of one model was chosen as a final one. Choosing one model to win does not usually take into consideration the uncertainty in all or some of the candidate models. Therefore in Chapter 5, the variables generated in all models during feature selection were chosen instead of only those of the best model. They were used to build a reduced predictive model. Afterwards, they were ranked according to their frequency distribution, and then used again as a filter method to further reduce the complexity of the predictive model. The second difference is building a multiple classifier system instead of an individual one and merging FS with their construction. It was constructed using three popular machine leaning algorithms which are artificial neural network, support vector machine, and K-nearest neighbor combined using unweighted and weighted majority voting techniques. The performance of the proposed method was also compared with AIC and BIC survival variable selection methods used with Cox's model. It was also compared with LASSO the popular shrinkage variable selection method that is used with Cox's model.

The results of Chapter 5 indicated that MCS using unweighted and weighted voting outperformed both the individual classifiers and Cox's survival variable selection methods in the p-value and concordance index expect for concordance index of MLP for Center 2. In addition, the final model generated by the proposed MCS hybrid feature selection is more powerful in classifying the long term risk of aortic complications after EVAR for new patients not used in training the predictive model compared to the other variable selection methods. This is due its higher prediction's sensitivity (true positive rate). For this reason, it may be preferred by clinicians to help them in planning the future surveillance plan for new patients.

In general, this thesis has suggested a method that can be used as a cost-effective and risk-stratified surveillance system. Censoring is the main feature that differentiates survival data from a standard data. Several methods were suggested to deal with this problem; however, the proposed system offered a new solution that can handle censoring of high level. Some approaches were presented in the literature to reduce the dimension of the data. The proposed system presented a combination of both variable selection and feature reduction methods that

can be used in survival analysis which can lower complexity and reduce the effort done by physicians to collect unimportant variables which might affect prediction. Standard statistical survival methods are commonly used to model survivability; though, this thesis combined the strengths of several machine learning classifiers which is preferred than the statistical approaches to construct a powerful predictive model which would allow clinicians to decide which surveillance plan each patient could undergo. High risk patients with high probability to be re-intervened will have a more regular observation plan. While, other low risk patients with lower probability will be observed less often.

6.2 Clinical Findings

Existing procedures for surveillance following EVAR are experimental but not evidence based which has led to extensive dissimilarity in practice in both UK and globally [18, 292]. This variation includes the imaging protocol and time intervals of scans. Most vascular centers carry on employing CT scans as the first choice of imaging. CT requires the frequent exposure of patients to nephrotoxic contrast [293] and ionising radiation [294] which result in a poor and cost-inefficient EVAR surveillance system. This current study offers an evidence based method that is suitable for EVAR surveillance which avoids the need for costly and harmful imaging taken for patients suffering from aortic aneurysm. It proposed using machine learning techniques such as ANN and MCS for analysing aortic morphology and basic co-morbidity data. Predictions using the proposed models were capable of reporting a more evidence-based method to post-operative surveillance decision plan, in which the regularity of images can be directed to the risk of endograft failure.

The predictive models constructed using the proposed approaches showed promising concordance index and p-values of the log-rank test indicating a great discriminating ability between low and high risk groups. These values would also have clinical importance if used prospectively. Moreover, the sensitivity rates of the models are considered to be more powerful than existing methods as it possibly indicates a rise in correctly classifying the event (EVAR re-intervention in this study) without increasing the cost of both surveillance and the collection of unimportant variables. The results were compared with clinical practice methods which showed the superiority of predictions produced using the proposed approaches.

Features selected using the proposed approaches has been viewed to clinical experts who verified their clinical significance in predicting technically difficult or challenging morphology for endografts available. They mentioned that predictions results using these features are clinically realistic and reasonable. Though, those chosen using the multiple classifier system

combined with weighted majority are more sensible as it contains neck length which is usually considered by surgeons for deciding the future monitoring plan for patients.

6.3 Future Work

In this thesis, patients with missing values were ignored during the construction of the predictive stratified system. This would lead to some loss of information, therefore techniques that deal with data imputation and missing values could be investigated to handle this. These techniques include single and multiple imputation methods. Single imputations methods replace each missing value by a single value. Instead, multiple imputations methods offer a number of possible values that present the uncertainty about the suitable value to impute [295]. Therefore, they are preferred to replace missing values in EVAR datasets to get use of every example available in a dataset.

In Chapter 3, the number of patients used to train the low and high risk Bayesian networks are not exactly equal. Sampling techniques could be used to equalise them. Therefore, they should be investigated. Examples of these methods include random oversampling and synthetic minority oversampling technique (SMOTE). Random oversampling is the simplest sampling technique. It increases the minority class patients by randomly selecting some of them and duplicating them to the training data. SMOTE is another oversampling approach based on the creation of synthetic new examples of the minority class and adding them to the training data instead of data duplication. Based on the quantity of oversampling needed, it randomly uses the K-nearest neighbors to perform oversampling [296].

In Chapters 4 and 5, even though, factor analysis method have reduced the dimension of the data in order to lower the computational cost for feature selection process done later, however, it still have high cost due to the use of the proposed stepwise strategy. Although, this strategy gives another chance of eliminated variables to re-enter the feature selection process once again, but it increases the computational cost for choosing the final variables used for prediction. Fast variable selection approaches could be examined to facilitate and accelerate the choice of a reduced number of variables that enhance the performance of the predictive model and lower its complexity. Examples of these methods are the fast correlation based feature selection (FCFS) and information gain. FCFS method depends on symmetrical uncertainty as a goodness of fit measure which is the ratio between information gain and entropy between two features. FCFS selects a reduced subset of variables by performing a correlation analysis between variables. This analysis also counts for correlation between variables and the class attribute. Irrelevant and redundant variables are eliminated according to a symmetrical uncertainty threshold predetermined by user [297]. An information gain method measures the amount of information gained by each feature which corresponds to how relevant is it to the class attribute. Both FCFS

and information techniques are used with standard supervised data; however as for future work they could be modified to work with survival data containing censoring especially at high level such as the EVAR datasets.

The proposed methods in this thesis used two Centers for constructing a multicenter study. One dataset was used for model construction, feature selection, and model selection. The other Center for the validation of the final reduced predictive model. As for future work, new datasets may be collected to test the performance of the reduced predictive models on a wider scale to be applicable to be used global wise. Moreover, future research can aim to show that surveillance decision plans based on the reduced predictive models resulted in better patient care. It can also show that the number of patients exposed to frequent radiations and contrast nephropathy is reduced and how this will affect their health. Finally, this thesis used EVAR datasets, future project may focus on using the proposed techniques on other medical survival data.

References

[1] P. Burns, Retreived (2015, March 1). *Abdominal Aortic Aneurysm*. Available: http://www.edinburghsurgeon.com/aaa.html.

[2] P. R. Taylor, J. Reidy, and J. E. Scoble, "Endovascular abdominal aortic aneurysm repair and renal function," *Nephrology Dialysis Transplantation*, vol. 21, pp. 2362-2365, 2006.

[3] H. T. Hua, R. P. Cambria, S. K. Chuang, M. C. Stoner, C. J. Kwolek, K. S. Rowell, *et al.*, "Early outcomes of endovascular versus open abdominal aortic aneurysm repair in the National Surgical Quality Improvement Program-Private Sector (NSQIP-PS)," *Journal of Vacsular Surgery*, vol. 41, pp. 382-9, Mar 2005.

[4] P. Burns, Retreived (2015, March 1). *Endovascular Aneurysm Repair*. Available: http://www.edinburghsurgeon.com/aaa.html.

[5] S. Bandyopadhyay, J. Wolfson, D. M. Vock, G. Vazquez-Benitez, G. Adomavicius, M. Elidrisi, *et al.*, "Data mining for censored time-to-event data: a Bayesian network model for predicting cardiovascular risk from electronic health record data," *Data Mining and Knowledge Discovery*, pp. 1-37, 2014.

[6] I. Štajduhar and B. Dalbelo-Bašić, "Uncensoring censored data for machine learning: A likelihood-based approach," *Expert Systems with Applications*, vol. 39, pp. 7226-7234, 2012.

[7] A. F. Taktak, A. Eleuteri, M. S. Aung, P. J. Lisboa, L. Desjardins, and B. Damato, "Survival analysis in cancer using a partial logistic neural network model with Bayesian regularisation framework: a validation study," *International Journal of Knowledge Engineering and Soft Data Paradigms*, vol. 1, pp. 277-295, 2009.

[8] H. Liu and H. Motoda, *Feature selection for knowledge discovery and data mining*. Boston: Kluwer Academic Publishers, 1998.

[9] H. Ishwaran, U. B. Kogalur, E. Z. Gorodeski, A. J. Minn, and M. S. Lauer, "High-dimensional variable selection for survival data," *Journal of the American Statistical Association*, vol. 105, pp. 205-217, 2010.

[10] K. M. Leung, R. M. Elashoff, and A. A. Afifi, "Censoring issues in survival analysis," *Annual Review of Public Health*, vol. 18, pp. 83-104, 1997.

[11] B. Zupan, J. DemšAr, M. W. Kattan, J. R. Beck, and I. Bratko, "Machine learning for survival analysis: a case study on recurrence of prostate cancer," *Artificial intelligence in medicine*, vol. 20, pp. 59-75, 2000.

[12] Y. Liu, U. Aickelin, J. Feyereisl, and L. G. Durrant, "Wavelet feature extraction and genetic algorithm for biomarker detection in colorectal cancer data," *Knowledge-Based Systems*, vol. 37, pp. 502-514, 2013.

[13] H. N. Dezfouli, and Bakar, M. R. A. , "Feed forward neural networks models for survival analysis.," in *2012 International Conference on Statistics in Science, Business, and Engineering (ICSSBE)*, Langkawi, 2012, pp. 1-5.

[14] M. Woźniak, M. Graña, and E. Corchado, "A survey of multiple classifier systems as hybrid systems," *Information Fusion*, vol. 16, pp. 3-17, 2014.

[15] A. Taktak, L. Antolini, M. Aung, P. Boracchi, I. Campbell, B. Damato, *et al.*, "Double-blind evaluation and benchmarking of survival models in a multi-center study," *Computers in biology and medicine*, vol. 37, pp. 1108-1120, 2007.

[16] C. T. Arsene and P. J. Lisboa, "Multicenter study design in survival analysis," in *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2012 IEEE Symposium on*, 2012, pp. 135-143.

[17] N. Hay, F. McCracken, J. Richardson, E. George, and D. Barnett, "Endovascular stent-grafts for the treatment of abdominal aortic aneurysms: NICE technology appraisal guidance," *Heart*, vol. 95, pp. 1798-800, Nov 2009.

[18] A. Karthikesalingam, A. A. Page, C. Pettengell, R. J. Hinchliffe, I. M. Loftus, M. M. Thompson, *et al.*, "Heterogeneity in surveillance after endovascular aneurysm repair in the UK," *Europian Journal of Vascular and Endovascular Surgery*, vol. 42, pp. 585-90, Nov 2011.

[19] R. A. Weerakkody, S. R. Walsh, C. Cousins, K. E. Goldstone, T. Y. Tang, and M. E. Gaunt, "Radiation exposure during endovascular aneurysm repair," *British Journal of Surgery*, vol. 95, pp. 699-702, Jun 2008.

[20] S. R. Walsh, T. Y. Tang, and J. R. Boyle, "Renal consequences of endovascular abdominal aortic aneurysm repair," *Journal of Endovascular Therapy*, vol. 15, pp. 73-82, Feb 2008.

[21] A. Karthikesalingam, P. J. Holt, R. J. Hinchliffe, I. M. Nordon, I. M. Loftus, and M. M. Thompson, "Risk of reintervention after endovascular aortic aneurysm repair," *British Journal of Surgery*, vol. 97, pp. 657-63, May 2010.

[22] L. C. Brown, J. T. Powell, S. G. Thompson, D. M. Epstein, M. J. Sculpher, and R. M. Greenhalgh, "The UK EndoVascular Aneurysm Repair (EVAR) trials: design, methodology and progress," *European Journal of Vascular and Endovascular Surgery*, vol. 27, pp. 372-381, 2004.

[23] D. Chambers, Epstein, D., Walker, S., Fayter, D. ,Paton, F. ,Wright, K., Michaels, J.,Thomas, S.,Sculpher, M.,Woolacott, N., "Endovascular stents for abdominal aortic aneurysms: a systematic review and economic model," *NIHR Health Technology Assessment programme: Executive Summaries*, vol. 13, pp. 1-189, 215-318, iii, Oct 2009.

[24] O. Attallah and X. Ma, "Bayesian neural network approach for determining the risk of re-intervention after endovascular aortic aneurysm repair," *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*, vol. 228, pp. 857-66, Sep 2014.

[25] A. Karthikesalingam, Attallah, Omneya, Ma, Xianghong, Bahia, Sandeep Singh, Thompson, Luke, Vidal-Diez, Alberto, Choke, Edward C, Bown, Matt J, Sayers, Robert D, Thompson, Matt M, Peter J. Holt, "An Artificial Neural Network Stratifies the Risks of Reintervention and Mortality after Endovascular Aneurysm Repair; a Retrospective Observational study," *PloS one*, vol. 10, p. e0129024, 2015.

[26] S. Kul, "The use of survival analysis for clinical pathways," *International Journal of Care Pathways*, vol. 14, pp. 23-26, 2010.

[27] L. Ohno-Machado, "Modeling medical prognosis: survival analysis techniques," *Journal of Biomedical Informatics*, vol. 34, pp. 428-39, Dec 2001.

[28] S. F. Brown, A. J. Branford, and W. Moran, "On the use of artificial neural networks for the analysis of survival data," *IEEE Transactions on Neural Networks*, vol. 8, pp. 1071-7, 1997.

[29] R. Singh and K. Mukhopadhyay, "Survival analysis in clinical trials: Basics and must know areas," *Perspectives in clinical research*, vol. 2, pp. 145-8, 2011.

[30] S. Prinja, N. Gupta, and R. Verma, "Censoring in clinical trials: review of survival analysis techniques," *Indian journal of community medicine: official publication of Indian Association of Preventive & Social Medicine*, vol. 35, pp. 217-21, Apr 2010.

[31] J. P. Klein and M. L. Moeschberger, *Survival analysis : techniques for censored and truncated data*, 2nd ed. New York: Springer, 2003.

[32] M. W. Kattan, K. R. Hess, and J. R. Beck, "Experiments to determine whether recursive partitioning (CART) or an artificial neural network overcomes theoretical limitations of Cox proportional hazards regression," *Computers and Biomedical Research*, vol. 31, pp. 363-73, Oct 1998.

[33] C.-M. Chen, C.-Y. Hsu, H.-W. Chiu, and H.-H. Rau, "Prediction of survival in patients with liver cancer using artificial neural networks and classification and regression trees," in *Seventh International Conference on Natural Computation (ICNC)*, Shanghai, 2011, pp. 811-815.

[34] T.-N. Wang, C.-H. Cheng, and H.-W. Chiu, "Predicting post-treatment survivability of patients with breast cancer using Artificial Neural Network methods," presented at the 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Osaka, 2013.

[35] H. B. Burke, D. B. Rosen, and P. H. Goodman, "Comparing artificial neural networks to other statistical methods for medical outcome prediction," presented at the IEEE World Congress on Computational Intelligence, International Conference on Neural Networks, Orlando, FL, 1994.

[36] Y.-C. Chen, W.-W. Yang, and H.-W. Chiu, "Artificial neural network prediction for cancer survival time by gene expression data," presented at the 3rd International Conference on Bioinformatics and Biomedical Engineering,. ICBBE . , Beijing, 2009.

[37] E. Pesonen, M. Eskelinen, and M. Juhola, "Comparison of different neural network algorithms in the diagnosis of acute appendicitis," *International journal of bio-medical computing*, vol. 40, pp. 227-233, 1996.

[38] B. D. Ripley and R. M. Ripley, "Neural networks as statistical methods in survival analysis," in *in Artificial Neural Networks: Prospects for*

Medicine, E. R. Dybowski and V. Grant, Ed., ed Texas: Landes Biosciences, 1997, pp. 237-255.

[39] M. De Laurentiis and P. M. Ravdin, "A technique for using neural network analysis to perform survival analysis of censored data," *Cancer Letter*, vol. 77, pp. 127-38, Mar 15 1994.

[40] P. M. Ravdin and G. M. Clark, "A practical application of neural network analysis for predicting outcome of individual breast cancer patients," *Breast Cancer Research and Treatment*, vol. 22, pp. 285-93, 1992.

[41] K. Liestbl, P. K. Andersen, and U. Andersen, "Survival analysis and neural nets," *Statistics in medicine*, vol. 13, pp. 1189-1200, 1994.

[42] C.-L. Chi, W. N. Street, and W. H. Wolberg, "Application of artificial neural network-based survival analysis on two breast cancer datasets," *AMIA Annual Symposium Proceedings*, vol. 2007, p. 130, 2007.

[43] E. M. Biganzoli, P. Boracchi, F. Ambrogi, and E. Marubini, "Artificial neural network for the joint modelling of discrete cause-specific hazards," *Artificial intelligence in medicine*, vol. 37, pp. 119-130, 2006.

[44] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American statistical association*, vol. 53, pp. 457-481, 1958.

[45] D. Collett, *Modelling survival data in medical research*, Third edition. ed. Boca Raton: CRC Press, Taylor & Francis Group, 2015.

[46] S. R. Rao and D. A. Schoenfeld, "Survival methods," *Circulation*, vol. 115, pp. 109-113, 2007.

[47] S. L. Spruance, J. E. Reid, M. Grace, and M. Samore, "Hazard ratio in clinical trials," *Antimicrobial agents and chemotherapy*, vol. 48, pp. 2787-2792, 2004.

[48] P. Cheema and R. Burkes, "Overall survival should be the primary endpoint in clinical trials for advanced non-small-cell lung cancer," *Current Oncology*, vol. 20, p. e150, 2013.
[49] H. G. Welch, L. M. Schwartz, and S. Woloshin, "Are increasing 5-year survival rates evidence of success against cancer?," *Jama*, vol. 283, pp. 2975-2978, 2000.
[50] M. K. Goel, P. Khanna, and J. Kishore, "Understanding survival analysis: Kaplan-Meier estimate," *International journal of Ayurveda research*, vol. 1, p. 274, 2010.
[51] R. J. Baatenburg de Jong, J. Hermans, J. Molenaar, J. J. Briaire, and S. le Cessie, "Prediction of survival in patients with head and neck cancer," *Head and Neck*, vol. 23, pp. 718-24, Sep 2001.
[52] T. van der Ploeg, F. Datema, R. B. de Jong, and E. W. Steyerberg, "Prediction of survival with alternative modeling techniques using pseudo values," *PloS one*, vol. 9, p. e100234, 2014.
[53] S. Polterauer, Grimm, C, Hofstetter, G, Concin, N, Natter, C, Sturdza, A, Pötter, R, Marth, C, Reinthaller, A,Heinze, G, "Nomogram prediction for overall survival of patients diagnosed with cervical cancer," *British journal of cancer*, vol. 107, pp. 918-924, 2012.
[54] S. L. Pomeroy, Tamayo, Pablo, Gaasenbeek, Michelle, Sturla, Lisa M, Angelo, Michael, McLaughlin, Margaret E, Kim, John YH, Goumnerova, Liliana C, Black, Peter M, Lau, Ching, "Prediction of central nervous system embryonal tumour outcome based on gene expression," *Nature*, vol. 415, pp. 436-442, 2002.
[55] J. S. Michaelson, Silverstein, Melvin, Wyatt, John, Weber, Griffin, Moore, Richard, Halpern, Elkan, Kopans, Daniel B, Hughes, Kevin, "Predicting the survival of patients with breast carcinoma using tumor size," *Cancer*, vol. 95, pp. 713-723, 2002.
[56] S. M. Kim, Kim, Moon Jin, Jung, Hyun, Kim, Kihyun, Kim, Seok Jin, Jang, Jun Ho, Kim, Won Seog, Jung, Chul Won, "Comparison of the Freiburg and Charlson Comorbidity Indices in Predicting Overall Survival in Elderly Patients with Newly Diagnosed Multiple Myeloma," *BioMed research international*, vol. 2014, 2014.
[57] H. W. Farber, D. P. Miller, M. D. McGoon, A. E. Frost, W. W. Benton, and R. L. Benza, "Predicting outcomes in pulmonary arterial hypertension based on the 6-minute walk distance," *The Journal of Heart and Lung Transplantation*, vol. 34, pp. 362-368, 2014.
[58] I. S. Lossos, Czerwinski, Debra K, Alizadeh, Ash A, Wechser, Mark A, Tibshirani, Rob, Botstein, David, Levy, Ronald, "Prediction of survival in diffuse large-B-cell lymphoma based on the expression of six genes," *New England Journal of Medicine*, vol. 350, pp. 1828-1837, 2004.
[59] J.-Y. Yang, Yoshihara, Kosuke, Tanaka, Kenichi, Hatae, Masayuki, Masuzaki, Hideaki, Itamochi, Hiroaki, Takano, Masashi, Ushijima, Kimio, Tanyi, Janos L, Coukos, George, "Predicting time to ovarian carcinoma recurrence using protein markers," *The Journal of clinical investigation*, vol. 123, p. 3740, 2013.
[60] Z. F. Tokatli, M. Türe, İ. K. Ömürlü, R. Ç. Alas, and M. C. Uzal, "Developing and comparing two different prognostic indexes for predicting disease-free survival of nonmetastatic breast cancer patients," *Turkish Journal of Medical Sciences*, vol. 41, pp. 769-780, 2011.
[61] A. Abadi, F. Amanpour, C. Bajdik, and P. Yavari, "Breast Cancer Survival Analysis: Applying the Generalized Gamma Distribution under Different Conditions of the Proportional Hazards and Accelerated Failure Time

Assumptions," *International journal of preventive medicine,* vol. 3, pp. 644-51, Sep 2012.
[62] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society. Series B,* vol. 34, pp. 187–220, 1972.
[63] D. R. Cox, "Partial likelihood," *Biometrika,* vol. 62, pp. 269-276, 1975.
[64] X. Li, Y. Zhang, Y. Zhang, J. Ding, K. Wu, and D. Fan, "Survival prediction of gastric cancer by a seven-microRNA signature," *Gut,* vol. 59, pp. 579-85, 2010.
[65] J. S. Lee, Chu, I. S., Heo, J., Calvisi, D. F., Sun, Z., Roskams, T., Durnez, A., Demetris, A. J.,Thorgeirsson, S. S., "Classification and prediction of survival in hepatocellular carcinoma by gene expression profiling," *Hepatology,* vol. 40, pp. 667-76, Sep 2004.
[66] K. L. Lee, Pryor, D. B., Harrell, F. E., Jr.,Califf, R. M., Behar, V. S. Floyd, W. L., Morris, J. J., Waugh, R. A.,Whalen, R. E.,Rosati, R. A., "Predicting outcome in coronary disease. Statistical models versus expert clinicians," *The American Journal of Medicine,* vol. 80, pp. 553-60, Apr 1986.
[67] W. C. Levy, Mozaffarian, D., Linker, D. T., Sutradhar, S. C., Anker, S. D., Cropp, A. B., Anand, I., Maggioni, A., Burton, P., Sullivan, M. D., Pitt, B., Poole-Wilson, P. A., Mann, D. L.,Packer, M., "The Seattle Heart Failure Model: prediction of survival in heart failure," *Circulation,* vol. 113, pp. 1424-33, 2006.
[68] T. A. MacKenzie, Malenka, David J, Olmstead, Elaine M, Piper, Winthrop D, Langner, Craig, Ross, Cathy S, O'Connor, Gerald T, Northern New England Cardiovascular Disease Study Group, "Prediction of survival after coronary revascularization: modeling short-term, mid-term, and long-term survival," *The Annals of thoracic surgery,* vol. 87, pp. 463-472, 2009.
[69] K. Bambha, Pierce, C., Cox, C., French, A. L., Tien, P. C., Sharp, G. B.,Augenbraun, M., Glesby, M. J., Villacres, M. C., Plankey, M., Strickler, H. D.,Gange, S. J.,Peters, M. G., "Assessing mortality in women with hepatitis C virus and HIV using indirect markers of fibrosis," *AIDS,* vol. 26, pp. 599-607, 2012.
[70] H. C. van Houwelingen, T. Bruinsma, A. A. Hart, L. J. van't Veer, and L. F. Wessels, "Cross-validated Cox regression on microarray gene expression data," *Statistics in medicine,* vol. 25, pp. 3201-3216, 2006.
[71] W. N. Van Wieringen, D. Kun, R. Hampel, and A.-L. Boulesteix, "Survival prediction using gene expression data: a review and comparison," *Computational statistics & data analysis,* vol. 53, pp. 1590-1603, 2009.
[72] K. H. Lee, S. Chakraborty, and J. Sun, "Bayesian variable selection in semiparametric proportional hazards model for high dimensional survival data," *The International Journal of Biostatistics,* vol. 7, pp. 1-32, 2011.
[73] J. Kalpathy-Cramer, W. Hersh, J. S. Kim, C. R. Thomas, and S. J. Wang, "Survival prediction models for estimating the benefit of post-operative radiation therapy for gallbladder cancer and lung cancer," *Symposium Proceedings,* vol. 2008, pp. 348-352, 2008.
[74] J. Cui, Forbes, Andrew,Kirby, Adrienne, Marschner, Ian,Simes, John, Hunt, David, West, Malcolm,Tonkin, Andrew, "Semi-parametric risk prediction models for recurrent cardiovascular events in the LIPID study," *BMC medical research methodology,* vol. 10, p. 27, 2010.
[75] E. S. Sheets, L. E. Duncan, A. S. Bjornsson, L. W. Craighead, and W. E. Craighead, "Personality pathology factors predict recurrent major depressive disorder in emerging adults," *Journal of Cinical Psychology,* vol. 70, pp. 536-545, 2014.

[76] T. Emura, Y.-H. Chen, and H.-Y. Chen, "Survival prediction based on compound covariate under cox proportional hazard models," *PloS one,* vol. 7, p. e47627, 2012.
[77] W. Zhang, T. Ota, V. Shridhar, J. Chien, B. Wu, and R. Kuang, "Network-based survival analysis reveals subnetwork signatures for predicting outcomes of ovarian cancer treatment," *PLoS computational biology,* vol. 9, p. e1002975, 2013.
[78] G. Zhao, *Nonparametric and parametric survival analysis of censored data with possible violation of method assumptions*. USA: ProQuest, 2008.
[79] Z. Ali, M. Hosseini, M. Mahmoodi, K. Mohammad, H. Zeraati, and K. H. Naieni, "A Comparison between Accelerated Failure-time and Cox Proportional Hazard Models in Analyzing the Survival of Gastric Cancer Patients," *Iranian journal of public health,* vol. 44, p. 1095, 2015.
[80] J. Qi, "Comparison of proportional hazards and accelerated failure time models," Master, University of Saskatchewan Saskatoon, Canada, 2009.
[81] F. Nakhaee and M. Law, "Parametric modelling of survival following HIV and AIDS in the era of highly active antiretroviral therapy: data from Australia," *East Mediterr Health J,* vol. 17, pp. 231-7, Mar 2011.
[82] K. J. Carroll, "On the use and utility of the Weibull model in the analysis of survival data," *Controlled Clinical Trials,* vol. 24, pp. 682-701, 2003.
[83] S. J. Wang, J. Kalpathy-Cramer, J. S. Kim, C. D. Fuller, and C. R. Thomas, "Parametric survival models for predicting the benefit of adjuvant chemoradiotherapy in gallbladder cancer," *AMIA Annual Symposium Proceedings* vol. 2010, pp. 847-51, 2010.
[84] P. Royston and M. K. Parmar, "Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects," *Statistics in medicine,* vol. 21, pp. 2175-2197, 2002.
[85] A. Biglarian, E. Hajizadeh, and A. Kazemnejad, "Comparison of Artificial Neural Network and Parametric Regression Models in Survival Prediction of Patients with Gastric Cancer," *Iranian Journal of Epidemiology,* vol. 6, pp. 22-27, 2010.
[86] L. Claret, Girard, Pascal, Hoff, Paulo M, Van Cutsem, Eric, Zuideveld, Klaas P, Jorga, Karin, Fagerberg, Jan, Bruno, René, "Model-based prediction of phase III overall survival in colorectal cancer on the basis of phase II tumor dynamics," *Journal of Clinical Oncology,* vol. 27, pp. 4103-4108, 2009.
[87] P. Newcombe, Ali, H Raza,Blows, FM, Provenzano, E, Pharoah, PD, Caldas, C,Richardson, S, "Weibull regression with Bayesian variable selection to identify prognostic tumour markers of breast cancer survival," *Statistical methods in medical research,* p. 0962280214548748, 2014.
[88] I. Yahav and G. Shmueli, "Predicting potential survival rates of kidney transplant candidates from databases with existing allocation policies," in *Proceedings of the 5th INFORMS workshop on data mining and health informatics (DM-HI 2010), Austin, TX,* 2010.
[89] C.-Y. Fan, P.-C. Chang, J.-J. Lin, and J. Hsieh, "A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification," *Applied Soft Computing,* vol. 11, pp. 632-644, 2011.
[90] A. Eleuteri, Tagliaferri, R,Milano, L, Sansone, G, D'Agostino, D, De Placido, S, De Laurentiis, M, "Survival analysis and neural networks," presented at the Proceedings of the International Joint Conference on Neural Networks, Portland, OR, 2003.

[91] A. G. Karegowda, A. Manjunath, and M. Jayaram, "Comparative Study if Attribute Selection using Gain Ratio and Correlation based Feature Selection," *International Journal of Information Technology,* vol. 2, pp. 271-277, 2010.

[92] S. Gupta, D. Kumar, and A. Sharma, "Data mining classification techniques applied for breast cancer diagnosis and prognosis," *Indian Journal of Computer Science and Engineering (IJCSE),* vol. 2, pp. 188-195, 2011.

[93] C. J. Desilva, P. Choong, and Y. Attikiouzel, "Artificial neural networks and breast-cancer prognosis," *Australian Computer Journal,* vol. 26, pp. 78-81, 1994.

[94] C. J. S. d. P. L. Choong, J. Taran, P. Heenan, and H. Dawkins "Survival analysis using artificial neural networks," in *1st Australia and New Zealand Conference on Intelligent Information. Systems.*, Perth, Australia, 1993, pp. 283–287.

[95] J. M. Jerez-Aragonés, J. A. Gómez-Ruiz, G. Ramos-Jiménez, J. Muñoz-Pérez, and E. Alba-Conejo, "A combined neural network and decision trees model for prognosis of breast cancer relapse," *Artificial intelligence in medicine,* vol. 27, pp. 45-63, 2003.

[96] B. Baesens, T. Van Gestel, M. Stepanova, D. Van den Poel, and J. Vanthienen, "Neural network survival analysis for personal loan data," *Journal of the Operational Research Society,* vol. 56, pp. 1089-1098, 2005.

[97] M. Krętowska and L. Bobrowski, "Artificial neural networks in identifying areas with homogeneous survival time," presented at the 7th International Conference on Artificial Intelligence and Soft Computing-ICAISC 2004, Zakopane, Poland,, 2004.

[98] B. Taşdelen, S. Helvaci, H. Kaleağasi, and A. Özge, "Artificial neural network analysis for prediction of headache prognosis in elderly patients," *Turkish Journal of Medical Sciences,* vol. 39, pp. 5-12, 2009.

[99] Y. Pu, Baad, M. J., Jiang, Y., and Chen, Y. , "Application of artificial neural network and multiple linear regression models for predicting survival time of patients with non-small cell cancer using multiple prognostic factors including FDG-PET measurements," in *International Joint Conference on Neural Networks*, Beijing, China, july 2014, pp. 225-230.

[100] M. Fornili, F. Ambrogi, P. Boracchi, and E. Biganzoli, "Piecewise Exponential Artificial Neural Networks (PEANN) for Modeling Hazard Function with Right Censored Data," in *Computational Intelligence Methods for Bioinformatics and Biostatistics*, ed: Springer, 2014, pp. 125-136.

[101] M. R. Gohari, A. Biglarian, E. Bakhshi, and M. A. Pourhoseingholi, "Use of an artificial neural network to determine prognostic factors in colorectal cancer patients," *Asian Pacific Journal of Cancer Prevention,* vol. 12, pp. 1469-72, 2011.

[102] B. Ahmed, B. Dimitrov, A. Perna, G. Remuzzi, and M. Nahas, "Five Years Survival on Hemodialysis Predicted by Artificial Neural Network Model," *Arab Journal of Nephrology and Transplantation,* vol. 2, pp. 10-19, 2009.

[103] S. Belciug, Gorunescu, F., Salem, A. B., & Gorunescu, M. , "Clustering-based approach for detecting breast cancer recurrence," in *10th international conference on Intelligent Systems Design and Applications* Cairo, Egypt, 2010, pp. 533-538.

[104] S. Belciug and F. Gorunescu, "A hybrid neural network/genetic algorithm applied to breast cancer detection and recurrence," *Expert Systems,* vol. 30, pp. 243-254, 2013.

[105] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods," *Artificial Intelligence in Medicine,* vol. 34, pp. 113-127, 2005.
[106] A. Bellaachia and E. Guven, "Predicting breast cancer survivability using data mining techniques," *Age,* vol. 58, pp. 10-110.
[107] W. N. Street, "A Neural Network Model for Prognostic Prediction," presented at the The Fifteenth International Conference on Machine Learning, Madison, Wisconson, USA, 1998.
[108] E. Biganzoli, P. Boracchi, L. Mariani, and E. Marubini, "Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach," *Statistics in Medicine,* vol. 17, pp. 1169-1186, 1998.
[109] P. J. Lisboa, H. Wong, P. Harris, and R. Swindell, "A Bayesian neural network approach for modelling censored data with an application to prognosis after surgery for breast cancer," *Artificial Intelligence in Medicine,* vol. 28, pp. 1-25, 2003.
[110] A. R. L. U. H. Eleuteri, Liverpool ; Aung, M.S.H. ; Taktak, A.F.G. ; Damato, B. , "Continuous and discrete time survival analysis: Neural network approaches," presented at the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, lyon,France, 2007.
[111] B. Damato, A. Eleuteri, A. C. Fisher, S. E. Coupland, and A. F. Taktak, "Artificial neural networks estimating survival probability after treatment of choroidal melanoma," *Ophthalmology,* vol. 115, pp. 1598-1607, 2008.
[112] D. Ansari, J. Nilsson, R. Andersson, S. Regnér, B. Tingstedt, and B. Andersson, "Artificial neural networks predict survival from pancreatic cancer after radical surgery," *The American Journal of Surgery,* vol. 205, pp. 1-7, 2013.
[113] J. M. Jerez, Molina, Ignacio, García-Laencina, Pedro J, Alba, Emilio, Ribelles, Nuria, Martín, Miguel, Franco, Leonardo, "Missing data imputation using statistical and machine learning methods in a real breast cancer problem," *Artificial Intelligence in Medicine,* vol. 50, pp. 105-115, 2010.
[114] P. J. Lisboa, Etchells, Terence A, Jarman, Ian H, Aung, MS Hane, Chabaud, Sylvie, Bachelot, Thomas, Perol, David, Gargi, Thérèse, Bourdès, Valérie, Bonnevay, Stéphane, "Time-to-event analysis with artificial neural networks: An integrated analytical and rule-based study for breast cancer," *Neural networks,* vol. 21, pp. 414-426, 2008.
[115] P. A. Ana S. Fernandes, Ian H. Jarman, Terence A. Etchells, José M. Fonseca, Paulo J. G. Lisboa, "A Clinical Decision Support System for Breast Cancer Patients," in *Doctoral Conference on Computing, Electrical and Industrial Systems, DoCEIS 2010*, Caparica, Portugal, 2010, pp. 122-129.
[116] A. Taktak, Eleuteri, A., Aung, M., Lisboa, P., Desjardins, L., and Damato, B, "External Validation of a Bayesian Neural Network Model in Survival Analysis. In Machine Learning and Applications," in *Seventh International Conference on Machine Learning and Applications ICMLA'08,* San Diego, CA, 2008, pp. 607 - 612.
[117] L. Spelt, J. Nilsson, R. Andersson, and B. Andersson, "Artificial neural networks–A method for prediction of survival following liver resection for colorectal cancer metastases," *European Journal of Surgical Oncology (EJSO),* vol. 39, pp. 648-654, 2013.
[118] H. Hamdan and J. M. Garibaldi, "A framework for automatic modelling of survival using fuzzy inference," in *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), ,* 2012, pp. 1-8.

[119] L. Yan, Verbel, D., and Saidi, O., "Predicting prostate cancer recurrence via maximizing the concordance index," in *Proceedings of the Tenth International Conference on Knowledge Discovery and Data Mining*, Seattle, WA, USA 2004, pp. 479-485.

[120] F. E. Harrell, Jr., K. L. Lee, and D. B. Mark, "Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors," *Statistics in Medicine*, vol. 15, pp. 361-87, Feb 28 1996.

[121] J. Kalderstam, P. Edén, P.-O. Bendahl, C. Strand, M. Fernö, and M. Ohlsson, "Training artificial neural networks directly on the concordance index for censored data using genetic algorithms," *Artificial Intelligence in Medicine*, vol. 58, pp. 125-132, 2013.

[122] H. S. Vikas C. Raykar, Balaji Krishnapuram, Cary Dehing-oberije, Philippe Lambin, "On ranking in survival analysis: bounds on the concordance index," presented at the Advances in Neural Information Processing Systems 20: 21st Annual Conference on Neural Information Processing Systems 2007, Vancouver, British Columbia, Canada., 2007.

[123] D. Faraggi and R. Simon, "A neural network model for survival data," *Statistics in Medicine*, vol. 14, pp. 73-82, 1995.

[124] L. Mariani, Coradini, D, Biganzoli, E, Boracchi, P, Marubini, E, Pilotti, S, Salvadori, B, Silvestrini, R, Veronesi, U, Zucali, R, "Prognostic factors for metachronous contralateral breast cancer: a comparison of the linear Cox regression model and its artificial neural network extension," *Breast Cancer Research and Treatment*, vol. 44, pp. 167-178, 1997.

[125] A. Xiang, P. Lapuerta, A. Ryutov, J. Buckley, and S. Azen, "Comparison of the performance of neural network methods and Cox regression for censored survival data," *Computational Statistics & Data analysis*, vol. 34, pp. 243-257, 2000.

[126] R. M. Ripley, A. L. Harris, and L. Tarassenko, "Non-linear survival analysis using neural networks," *Statistics in Medicine*, vol. 23, pp. 825-842, 2004.

[127] F. V. Jensen, *An introduction to Bayesian networks*. New York: Springer, 1996.

[128] P. Sebastiani, Nolan, Vikki G, Baldwin, Clinton T, Abad-Grau, Maria M, Wang, Ling, Adewoye, Adeboye H, McMahon, Lillian C, Farrer, Lindsay A, Taylor IV, James G, Kato, Gregory J, "A network model to predict the risk of death in sickle cell disease," *Blood*, vol. 110, pp. 2727-2735, 2007.

[129] J. P. Choi, T. H. Han, and R. W. Park, "A hybrid Bayesian network model for predicting breast cancer prognosis," *Journal of Korean Society of Medical Informatics*, vol. 15, pp. 49-57, 2009.

[130] M. Gerstung, M. Baudis, H. Moch, and N. Beerenwinkel, "Quantifying cancer progression with conjunctive Bayesian networks," *Bioinformatics*, vol. 25, pp. 2809-2815, 2009.

[131] I. Štajduhar and B. Dalbelo-Bašić, "Learning Bayesian networks from survival data using weighting censored instances," *Journal of Biomedical Informatics*, vol. 43, pp. 613-622, 2010.

[132] S. Steele, Bilchik, Anton, Eberhardt, John, Kalina, Philip, Nissan, Aviram, Johnson, Eric, Avital, Itzhak, Stojadinovic, Alexander, "Using machine-learned Bayesian belief networks to predict perioperative risk of clostridium difficile infection following colon surgery," *Interactive Journal of Medical Research*, vol. 1, 2012.

[133] P. M. Rancoita, M. Zaffalon, E. Zucca, F. Bertoni, and C. P. de Campos, "Bayesian network data imputation with application to survival tree analysis," *Computational Statistics & Data Analysis*, vol. In Press, 2015.
[134] D. H. Wolpert, "The supervised learning no-free-lunch theorems," presented at the Proceedings of the 6th Online World Conference on Soft Computing and Industry, 2001.
[135] M. Woźniak and B. Krawczyk, "Combined classifier based on feature space partitioning," *International Journal of Applied Mathematics and Computer Science*, vol. 22, pp. 855-866, 2012.
[136] K. Woods, W. P. Kegelmeyer Jr, and K. Bowyer, "Combination of multiple classifiers using local accuracy estimates," *IEEE Transactions on Pattern Analysis and Machine Learning*, vol. 19, 1997.
[137] A. Tiwari, Knowles, Joshua, Avineri, Erel, Dahal, Keshav, Roy, Rajkumar, *Applications of Soft Computing: Recent Trends*. Berlin, Germany: Springer 2006.
[138] C.-h. Chen, L.-F. Pau, and P. S.-p. Wang, *Handbook of Pattern Recognition and Computer Vision* vol. 27. Singapore,: World Scientific, 2010.
[139] K. Tumer and J. Ghosh, "Analysis of decision boundaries in linearly combined neural classifiers," *Pattern Recognition*, vol. 29, pp. 341-348, 1996.
[140] T. G. Dietterich, "Ensemble methods in machine learning," in *1st International Workshop on Multiple Classifier Systems*. vol. 1857, E. J. Kitler and F. Roli, Ed., ed: Springer, 2000, pp. 1-15.
[141] L. I. Kuncheva, *Combining pattern classifiers : methods and algorithms*. Hoboken, NJ: J. Wiley, 2004.
[142] L. Lam, "Classifier combinations: implementations and theoretical issues," in *Proceedings of the First International Workshop on Multiple Classifier Systems, MCS '00*, Cagliari, Italy, 2000, pp. 77–86.
[143] A. F. R. Rahman and M. C. Fairhurst, "Serial combination of multiple experts: A unified evaluation," *Pattern Analysis & Applications*, vol. 2, pp. 292-311, 1999.
[144] P. I. Fumera G. , and Roli F., "A two-stage classifier with reject option for text categorization," in *Proceeding of the 5th International Workshop on Statistical Techniques in Pattern Recognition (SPR 2004)*, Lisbon, Portugal, 2004, pp. 771-779.
[145] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, pp. 119-139, 1997.
[146] G. Wang, J. Hao, J. Ma, and H. Jiang, "A comparative assessment of ensemble learning for credit scoring," *Expert Systems with Applications*, vol. 38, pp. 223-230, 2011.
[147] C.-F. Tsai, Y.-F. Hsu, and D. C. Yen, "A comparative study of classifier ensembles for bankruptcy prediction," *Applied Soft Computing*, vol. 24, pp. 977-984, 2014.
[148] M. Sattlecker, R. Baker, N. Stone, and C. Bessant, "Support vector machine ensembles for breast cancer type prediction from mid-FTIR micro-calcification spectra," *Chemometrics and Intelligent Laboratory Systems*, vol. 107, pp. 363-370, 2011.
[149] Y. Freund, and Schapire, R, "Experiments with a new boosting algorithm.," in *Proceedings of the Thirteenth International Conference on Machine Learning*, Bari, Italy, 1996, pp. 148–156.
[150] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, pp. 241-259, 1992.

[151] S. Džeroski and B. Ženko, "Is combining classifiers with stacking better than selecting the best one?," *Machine Learning,* vol. 54, pp. 255-273, 2004.
[152] K. Takeuchi and N. Collier, "Bio-medical entity extraction using support vector machines," *Artificial Intelligence in Medicine,* vol. 33, pp. 125-137, 2005.
[153] B. Twala, "Multiple classifier application to credit risk assessment," *Expert Systems with Applications,* vol. 37, pp. 3326-3336, 2010.
[154] J. Kittler, M. Hatef, R. P. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 20, pp. 226-239, 1998.
[155] M. Zhu and G. Fan, "Variable selection by ensembles for the Cox model," *Journal of Statistical Computation and Simulation,* vol. 81, pp. 1983-1992, 2011.
[156] A. Mayr and M. Schmid, "Boosting the concordance index for survival data--a unified framework to derive and evaluate biomarker combinations," *PLoS One,* vol. 9, p. e84483, 2014.
[157] M. Schmid and T. Hothorn, "Flexible boosting of accelerated failure time models," *BMC Bioinformatics,* vol. 9, p. 269, 2008.
[158] B. A. Johnson and Q. Long, "Survival ensembles by the sum of pairwise differences with application to lung cancer microarray studies," *The Annals of Applied Statistics,* vol. 5, p. 1081, 2011.
[159] I. Bou-hamad, D. Larocque, H. Ben-Ameur, L. C. Mâsse, F. Vitaro, and R. E. Tremblay, "Discrete-time survival trees," *Canadian Journal of Statistics,* vol. 37, pp. 17-32, 2009.
[160] F. R. Datema, Moya, Ana, Krause, Peter, Bäck, Thomas, Willmes, Lars, Langeveld, Ton, de Jong, Baatenburg, Robert, J, Blom, Henk M, "Novel head and neck cancer survival analysis approach: Random survival forests versus cox proportional hazards regression," *Head & Neck,* vol. 34, pp. 50-58, 2012.
[161] X. Chen and H. Ishwaran, "Pathway hunting by random survival forests," *Bioinformatics,* vol. 29, pp. 99-105, 2013.
[162] E. Hsich, E. Z. Gorodeski, E. H. Blackstone, H. Ishwaran, and M. S. Lauer, "Identifying important risk factors for survival in patient with systolic heart failure using random survival forests," *Circulation: Cardiovascular Quality and Outcomes,* vol. 4, pp. 39-45, 2011.
[163] M. Banerjee, Y. Ding, and A.-M. Noone, "Identifying representative trees from ensembles," *Statistics in Medicine,* vol. 31, pp. 1601-1616, 2012.
[164] Y. Chen, Z. Jia, D. Mercola, and X. Xie, "A gradient boosting algorithm for survival analysis via direct optimization of concordance index," *Computational and Mathematical Methods in Medicine,* vol. 2013, 2013.
[165] S. Floyd, Ruiz, C., Alvarez, S. A., Tseng, J., Whalen, G. , "Prediction of Pancreatic Cancer Survival through Automated Selection of Predictive Models," in *Third International Joint Conference on Biomedical Engineering Systems and Technologies BIOSTEC 2010*, Valencia, Spain, 2010, pp. 29-43.
[166] N.-C. Hsieh, L.-P. Hung, C.-C. Shih, H.-C. Keh, and C.-H. Chan, "Intelligent postoperative morbidity prediction of heart disease using artificial intelligence techniques," *Journal of Medical Systems,* vol. 36, pp. 1809-1820, 2012.
[167] J. A. Koziol, Feng, Anne C, Jia, Zhenyu, Wang, Yipeng, Goodison, Seven, McClelland, Michael, Mercola, Dan, "The wisdom of the commons: ensemble tree classifiers for prostate cancer prognosis," *Bioinformatics,* vol. 25, pp. 54-60, 2009.

[168] H. Ishwaran, E. H. Blackstone, C. Apperson-Hansen, and T. W. Rice, "A novel approach to cancer staging: application to esophageal cancer," *Biostatistics,* p. kxp016, 2009.
[169] X. Zhou, J. Liu, and J. Xiong, "Predicting distant metastasis in breast cancer using ensemble classifier based on context-specific miRNA regulation modules," in *2012 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Philadelphia, USA, 2012, pp. 1-6.
[170] N. Rathore, and Agarwal, S. , "Predicting the survivability of breast cancer patients using ensemble approach.," in *2014 International Conference on Issues and Challenges in Intelligent Computing Techniques*, Ghaziabad,India, 2014, pp. 459-464.
[171] R. Al-Bahrani, Agrawal, A., and Choudhary, A. , "Colon cancer survival prediction using ensemble data mining on SEER data," in *2013 IEEE International Conference on Big Data*, Silicon Valley, CA,USA, 2012, pp. 9-16.
[172] A. Agrawal, S. Misra, R. Narayanan, L. Polepeddi, and A. Choudhary, "Lung cancer survival prediction using ensemble data mining on SEER data," *Scientific Programming,* vol. 20, pp. 29-42, 2012.
[173] A. Kalinli, F. Sarikoc, H. Akgun, and F. Ozturk, "Performance comparison of machine learning methods for prognosis of hormone receptor status in breast cancer tissue samples," *Computer Methods and Programs in Biomedicine,* vol. 110, pp. 298-307, 2013.
[174] N. Mantel, "Evaluation of survival data and two new rank order statistics arising in its consideration," *Cancer Chemotherapy Reports. Part 1,* vol. 50, pp. 163-170, 1966.
[175] J. K. Taylor and C. Cihon, *Statistical techniques for data analysis*. Florida, USA: CRC Press, 2004.
[176] L. I. e. K. Han, "Model selection and model averaging for neural networks," PhD, Department of Statistics, Carnegie Mellon University Pittsburgh, 1999.
[177] A. Jain and D. Zongker, "Feature selection: Evaluation, application, and small sample performance," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 19, pp. 153-158, 1997.
[178] C.-F. Tsai, W. Eberle, and C.-Y. Chu, "Genetic algorithms in feature and instance selection," *Knowledge-Based Systems,* vol. 39, pp. 240-247, 2013.
[179] Y. C. Lim, T. S. Tan, S. H. S. Salleh, and D. K. Ling, "Application of Genetic Algorithm in unit selection for Malay speech synthesis system," *Expert Systems with Applications,* vol. 39, pp. 5376-5383, 2012.
[180] H. Lian, J. Li, and X. Tang, "SCAD-penalized regression in additive partially linear proportional hazards models with an ultra-high-dimensional linear part," *Journal of Multivariate Analysis,* vol. 125, pp. 50-64, 2014.
[181] M. Blachnik, "Comparison of various feature selection methods in application to prototype best rules," in *Computer Recognition Systems 3*. vol. 57, ed New York: Springer, 2009, pp. 257-264.
[182] D. Oakes and A. K. Manatunga, "A new representation of Cox's score statistic and its variance," *Statistics & Probability Letters,* vol. 14, pp. 107-110, 1992.
[183] D. G. Beer, Kardia, Sharon LR, Huang, Chiang-Ching, Giordano, Thomas J, Levin, Albert M, Misek, David E, Lin, Lin, Chen, Guoan. Gharib, Tarek G, Thomas, Dafydd G, "Gene-expression profiles predict survival of patients with lung adenocarcinoma," *Nature Medicine,* vol. 8, pp. 816-824, 2002.
[184] T. Ueda, Volinia, Stefano, Okumura, Hiroshi, Shimizu, Masayoshi, Taccioli, Cristian, Rossi, Simona, Alder, Hansjuerg, Liu, Chang-gong, Oue, Naohide, Yasui, Wataru, "Relation between microRNA expression and progression and

prognosis of gastric cancer: a microRNA expression analysis," *The Lancet Oncology*, vol. 11, pp. 136-146, 2010.

[185] X. Hu, Schwarz, Julie K, Lewis, James S, Huettner, Phyllis C, Rader, Janet S, Deasy, Joseph O, Grigsby, Perry W, Wang, Xiaowei, "A microRNA expression signature for cervical cancer prognosis," *Cancer Research*, vol. 70, pp. 1441-1448, 2010.

[186] S. Roessler, H.-L. Jia, A. Budhu, M. Forgues, Q.-H. Ye, J.-S. Lee, *et al.*, "A unique metastasis gene signature enables prediction of tumor relapse in early-stage hepatocellular carcinoma patients," *Cancer research*, vol. 70, pp. 10202-10212, 2010.

[187] Q. Tan, M. Thomassen, K. M. Jochumsen, O. Mogensen, K. Christensen, and T. A. Kruse, "Gene selection for predicting survival outcomes of cancer patients in microarray studies," in *Advances in Computer and Information Sciences and Engineering*, ed Netherlands: Springer, 2008, pp. 405-409.

[188] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proceedings of the National Academy of Sciences*, vol. 98, pp. 5116-5121, 2001.

[189] C. Yau, L. Esserman, D. H. Moore, F. Waldman, J. Sninsky, and C. C. Benz, "A multigene predictor of metastatic outcome in early stage hormone receptor-negative and triple-negative breast cancer," *Breast Cancer Research*, vol. 12, p. R85, 2010.

[190] R. Blanco, I. Inza, M. Merino, J. Quiroga, and P. Larrañaga, "Feature selection in Bayesian classifiers for the prognosis of survival of cirrhotic patients treated with TIPS," *Journal of Biomedical Informatics*, vol. 38, pp. 376-388, 2005.

[191] P. K. Mankoo, R. Shen, N. Schultz, D. A. Levine, and C. Sander, "Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles," *PLoS One*, vol. 6, p. e24709, 2011.

[192] T. Wu, W. Sun, S. Yuan, C.-H. Chen, and K.-C. Li, "A method for analyzing censored survival phenotype with gene expression data," *BMC Bioinformatics*, vol. 9, p. 417, 2008.

[193] H. Neuvirth, Michal Ozery-Flato, Jianying Hu, Jonathan Laserson, Martin S. Kohn, Shahram Ebadollahi, and Michal Rosen-Zvi, "Toward personalized care management of patients at risk: the diabetes case study," presented at the 17th ACM SIGKDD San Diego, CA,USA, 2011.

[194] Y.-C. Chen, W.-C. Ke, and H.-W. Chiu, "Risk classification of cancer survival using ANN with gene expression data from multiple laboratories," *Computers in Biology and Medicine*, vol. 48, pp. 1-7, 2014.

[195] D. M. Witten and R. Tibshirani, "Survival analysis with high-dimensional covariates," *Statistical Methods in Medical Research*, 2009.

[196] S.-L. Yu, Chen, Hsuan-Yu, Chang, Gee-Chen, Chen, Chih-Yi, Chen, Huei-Wen, Singh, Sher, Cheng, Chiou-Ling, Yu, Chong-Jen, Lee, Yung-Chie, Chen, Han-Shiang, "MicroRNA signature predicts survival and relapse in lung cancer," *Cancer Cell*, vol. 13, pp. 48-57, 2008.

[197] I. Choi, B. J. Wells, C. Yu, and M. W. Kattan, "An empirical approach to model selection through validation for censored survival data," *Journal of Biomedical Informatics*, vol. 44, pp. 595-606, 2011.

[198] D. Y. Heng, Xie, Wanling, Regan, Meredith M, Warren, Mark A, Golshayan, Ali Reza, Sahi, Chakshu, Eigl, Bernhard J, Ruether, J Dean, Cheng, Tina, North, Scott, "Prognostic factors for overall survival in patients with metastatic renal cell carcinoma treated with vascular endothelial growth factor–targeted agents:

results from a large, multicenter study," *Journal of Clinical Oncology,* vol. 27, pp. 5794-5799, 2009.

[199] S. Adabag, Rector, Thomas S, Anand, Inder S, McMurray, John J, Zile, Michael, Komajda, Michel, McKelvie, Robert S, Massie, Barry, Carson, Peter E, "A prediction model for sudden cardiac death in patients with heart failure and preserved ejection fraction," *European Journal of Heart Failure,* vol. 16, pp. 1175-1182, 2014.

[200] I. Thöm, Andritzky, Birte, Schuch, Gunter, Burkholder, Iris, Edler, Lutz, Johansen, Julia S, Bokemeyer, Carsten, Schumacher, Udo, Laack, Eckart, "Elevated pretreatment serum concentration of YKL-40-An independent prognostic biomarker for poor survival in patients with metastatic nonsmall cell lung cancer," *Cancer,* vol. 116, pp. 4114-4121, 2010.

[201] M. Naesens, Kuypers, DRJ, De Vusser, Katrien, Vanrenterghem, Yves, Evenepoel, Pieterm, Claes, Kathleen, Bammens, Bert, Meijers, Björn, Lerut, Evelyne, "Chronic histological damage in early indication biopsies is an independent risk factor for late renal allograft failure," *American Journal of Transplantation,* vol. 13, pp. 86-99, 2013.

[202] S. Eichinger, G. Heinze, L. M. Jandeck, and P. A. Kyrle, "Risk Assessment of Recurrence in Patients With Unprovoked Deep Vein Thrombosis or Pulmonary Embolism The Vienna Prediction Model," *Circulation,* vol. 121, pp. 1630-1636, 2010.

[203] A. Khosla, Cao, Y., Lin, C. C. Y., Chiu, H. K., Hu, J., Lee, H. , "An integrated machine learning approach to stroke prediction," in *16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Washington DC,USA, 2010, pp. 183-192.

[204] V. Lagani and I. Tsamardinos, "Structure-based variable selection for survival data," *Bioinformatics,* vol. 26, pp. 1887-1894, 2010.

[205] H. Pang, S. L. George, K. Hui, and T. Tong, "Gene selection using iterative feature elimination random forests for survival outcomes," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB),* vol. 9, pp. 1422-1431, 2012.

[206] V. Van Belle, Pelckmans, K., Suykens, J. A., Van Huffel, S. , "Feature selection in survival least squares support vector machines with maximal variation constraints," in *10th International Work-Conference on Artificial Neural Networks, IWANN 2009*, Salamanca, Spain,, 2009, pp. 65-72.

[207] M. J. Donovan, Kotsianti, Angeliki, Bayer-Zubek, Valentina, Verbel, David, Teverovskiy, Mikhail, Cordon-Cardo, Carlos, Costa, Jose, Greco, F Anthony, Hainsworth, John D, Parums, Dinah V, "A systems pathology model for predicting overall survival in patients with refractory, advanced non-small-cell lung cancer treated with gefitinib," *European Journal of Cancer,* vol. 45, pp. 1518-1526, 2009.

[208] M. Ahdesmäki, Lancashire, Lee, Proutski, Vitali, Wilson, Claire, Davison, Timothy S, Harkin, D Paul, Kennedy, Richard D, "Model selection for prognostic time-to-event gene signature discovery with applications in early breast cancer data," *Statistical Applications in Genetics and Molecular Biology,* vol. 12, pp. 619-635, 2013.

[209] A. Stebbins, Mehta, Rajendra H, Armstrong, Paul W, Lee, Kerry L, Hamm, Christian, Van de Werf, Frans, James, Stefan, Toftegaard-Nielsen, Torsten, Seabra-Gomes, Ricardo, White, Harvey D, "A Model for Predicting Mortality in Acute ST-Segment Elevation Myocardial Infarction Treated With Primary Percutaneous Coronary Intervention Results From the Assessment of

Pexelizumab in Acute Myocardial Infarction Trial," *Circulation: Cardiovascular Interventions,* vol. 3, pp. 414-422, 2010.

[210] K. M. Ting, & Quek, R. J. Y., "Model stability: a key factor in determining whether an algorithm produces an optimal model from a matching distribution," in *Third IEEE International Conference on Data Mining, ICDM 2003.* , Melbourne, Florida, USA., 2003, pp. 653 - 656.

[211] D. Posada and T. R. Buckley, "Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests," *Systematic Biology,* vol. 53, pp. 793-808, 2004.

[212] K. P. Burnham, D. R. Anderson, and K. P. Burnham, *Model selection and multimodel inference : a practical information-theoretic approach*, 2nd ed. New York: Springer, 2002.

[213] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics,* vol. 6, pp. 461-464, 1978.

[214] J. Huang, L. Liu, Y. Liu, and X. Zhao, "Group selection in the cox model with a diverging number of covariates," *Statistica Sinica,* vol. 24, pp. 1787-1810, 2014.

[215] W. Sauerbrei, A. Buchholz, A. L. Boulesteix, and H. Binder, "On stability issues in deriving multivariable regression models," *Biometrical Journal,* 2014.

[216] R. Xu, F. Vaida, and D. P. Harrington, "Using profile likelihood for semiparametric model selection with application to proportional hazards mixed models," *Statistica Sinica,* vol. 19, pp. 819-842, 2009.

[217] M. Donohue, R. Overholser, R. Xu, and F. Vaida, "Conditional Akaike information under generalized linear and proportional hazards mixed models," *Biometrika,* vol. 98, pp. 685-700, 2011.

[218] S. Luo, J. Xu, and Z. Chen, "Extended Bayesian information criterion in the Cox model with a high-dimensional feature space," *Annals of the Institute of Statistical Mathematics,* pp. 1-25, 2014.

[219] B. Hofner, T. Kneib, W. Hartl, and H. Küchenhoff, "Building Cox-type structured hazard regression models with time-varying effects," *Statistical Modelling,* vol. 11, pp. 3-24, 2011.

[220] Z. He, W. Tu, S. Wang, H. Fu, and Z. Yu, "Simultaneous variable selection for joint models of longitudinal and survival outcomes," *Biometrics,* vol. 71, pp. 178-187, 2014.

[221] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological),* pp. 267-288, 1996.

[222] R. Tibshirani, "The lasso method for variable selection in the Cox model," *Statistics in Medicine,* vol. 16, pp. 385-395, 1997.

[223] S. Lee, Rahnenführer, Jörg, Lang, Michel, De Preter, Katleen, Mestdagh, Pieter, Koster, Jan, Versteeg, Rogier, Stallings, Raymond L, Varesio, Luigi, Asgharzadeh, Shahab, "Robust Selection of Cancer Survival Signatures from High-Throughput Genomic Data Using Two-Fold Subsampling," *PloS One,* vol. 9, p. e108818, 2014.

[224] K. Kammers, M. Lang, J. G. Hengstler, M. Schmidt, and J. Rahnenführer, "Survival models with preclustered gene groups as covariates," *BMC Bioinformatics,* vol. 12, p. 478, 2011.

[225] Z. Liu and F. Jiang, "Gene identification and survival prediction with Lp Cox regression and novel similarity measure," *International Journal of Data Mining and Bioinformatics,* vol. 3, pp. 398-408, 2009.

[226] Z. Liu, R. B. Gartenhaus, X.-W. Chen, C. D. Howell, and M. Tan, "Survival prediction and gene identification with penalized global AUC maximization," *Journal of Computational Biology,* vol. 16, pp. 1661-1670, 2009.
[227] G. Ambler, S. Seaman, and R. Omar, "An evaluation of penalised survival methods for developing prognostic models with rare events," *Statistics in Medicine,* vol. 31, pp. 1150-1161, 2012.
[228] T. T. Wu, H. Gong, and E. M. Clarke, "A transcriptome analysis by lasso penalized Cox regression for pancreatic cancer survival," *Journal of Bioinformatics and Computational Biology,* vol. 9, pp. 63-73, 2011.
[229] S. Ma, J. Huang, M. Shi, Y. Li, and B.-C. Shia, "Semiparametric prognosis models in genomic studies," *Briefings in Bioinformatics,* vol. 11, pp. 385-393, 2010.
[230] T. Cai, J. Huang, and L. Tian, "Regularized estimation for the accelerated failure time model," *Biometrics,* vol. 65, pp. 394-404, 2009.
[231] F. Liu, D. Dunson, and F. Zou, "High-Dimensional Variable Selection in Meta-Analysis for Censored Data," *Biometrics,* vol. 67, pp. 504-512, 2011.
[232] H. H. Zhang and W. Lu, "Adaptive Lasso for Cox's proportional hazards model," *Biometrika,* vol. 94, pp. 691-703, 2007.
[233] R. I. Garcia, J. G. Ibrahim, and H. Zhu, "Variable selection in the cox regression model with covariates missing at random," *Biometrics,* vol. 66, pp. 97-104, 2010.
[234] I. Sohn, J. Kim, S.-H. Jung, and C. Park, "Gradient lasso for Cox proportional hazards model," *Bioinformatics,* vol. 25, pp. 1775-1781, 2009.
[235] J. J. Goeman, "L1 penalized estimation in the cox proportional hazards model," *Biometrical Journal,* vol. 52, pp. 70-84, 2010.
[236] L. C. Bergersen, I. K. Glad, and H. Lyng, "Weighted lasso with data integration," *Statistical Applications in Genetics and Molecular Biology,* vol. 10, pp. 1-29, 2011.
[237] J. G. Ibrahim, M. H. Chen, and S. N. MacEachern, "Bayesian variable selection for proportional hazards models," *Canadian Journal of Statistics,* vol. 27, pp. 701-717, 1999.
[238] D. Faraggi and R. Simon, "Bayesian variable selection method for censored survival data," *Biometrics,* pp. 1475-1485, 1998.
[239] K. E. Lee, Y. Kim, and R. Xu, "Bayesian variable selection under the proportional hazards mixed-effects model," *Computational Statistics & Data Analysis,* vol. 75, pp. 53-65, 2014.
[240] N. Sha, M. G. Tadesse, and M. Vannucci, "Bayesian variable selection for the analysis of microarray data with censored outcomes," *Bioinformatics,* vol. 22, pp. 2262-2268, 2006.
[241] F. C. Stingo, Y. A. Chen, M. G. Tadesse, and M. Vannucci, "Incorporating biological information into linear models: A Bayesian approach to the selection of pathways and genes," *The Annals of Applied Statistics,* vol. 5, 2011.
[242] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky, "Bayesian model averaging: a tutorial," *Statistical Science,* pp. 382-401, 1999.
[243] A. Annest, R. E. Bumgarner, A. E. Raftery, and K. Y. Yeung, "Iterative bayesian model averaging: A method for the application of survival analysis to high-dimensional microarray data," *BMC Bioinformatics,* vol. 10, p. 72, 2009.
[244] I. Bichindaritz, and Annest, A. , "Case based reasoning with Bayesian model averaging: an improved method for survival analysis on microarray data," in *18th International Conference on Case-Based Reasoning, ICCBR 2010,* Alessandria, Italy, 2010, pp. 346-359.

[245] S. Sarini, J. McGree, and K. Mengersen, "Bayesian variable selection and modelling for metastatic breast cancer data," *ANZIAM Journal,* vol. 55, pp. 168-181, 2014.
[246] M. Colson, J. Baglin, S. Bolsin, and M. Grocott, "Cardiopulmonary exercise testing predicts 5 yr survival after major surgery," *British Journal of Anaesthesia,* p. aes263, 2012.
[247] C. T. Arsene, and Lisboa, P. J. , "Multicenter study design in survival analysis. ," in *2012 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, San Diego, CA,USA, 2012, pp. 135-143.
[248] C. M. Bishop, *Neural networks for pattern recognition*. Oxford, New York: Clarendon Press ; Oxford University Press, 1995.
[249] A. Ghaffari, H. Abdollahi, M. Khoshayand, I. S. Bozchalooi, A. Dadgar, and M. Rafiee-Tehrani, "Performance comparison of neural network training algorithms in modeling of bimodal drug delivery," *International Journal of Pharmaceutics,* vol. 327, pp. 126-138, 2006.
[250] H. Yu, and Wilamowski, B. M. , "Levenberg-marquardt training," in *Industrial Electronics Handbook, 2nd edition*. vol. 3, ed Boca Raton: CRC Press., 2011, pp. 12-1-12-15.
[251] L. Gong, C. Liu, Y. Li, and F. Yuan, "Training feed-forward neural networks using the gradient descent method with the optimal stepsize," *Journal of Computational Information Systems,* vol. 8, pp. 1359-1371, 2012.
[252] N. Qian, "On the momentum term in gradient descent learning algorithms," *Neural networks,* vol. 12, pp. 145-151, 1999.
[253] Ö. KISI and E. Uncuoglu, "Comparison of three back-propagation training algorithms for two case studies," *Indian Journal of Engineering & Materials Sciences,* vol. 12, pp. 434-442, 2005.
[254] R. M. Zur, Y. Jiang, L. L. Pesce, and K. Drukker, "Noise injection for training artificial neural networks: A comparison with weight decay and early stopping," *Medical Physics,* vol. 36, pp. 4810-4818, 2009.
[255] D. Svozil, V. Kvasnicka, and J. Pospichal, "Introduction to multi-layer feed-forward neural networks," *Chemometrics and Intelligent Laboratory Systems,* vol. 39, pp. 43-62, 1997.
[256] C. M. Bishop, "Regularization and Complexity Control in Feed-Forward Networks," in *Proceedings International Conference on Artificial Neural Networks ICANN'95*, 1995, pp. 141–148.
[257] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural networks,* vol. 2, pp. 359-366, 1989.
[258] S. Haykin and R. Lippmann, "Neural Networks, A Comprehensive Foundation," *International Journal of Neural Systems,* vol. 5, pp. 363-364, 1994.
[259] A. Bhatt, H. B. Helle, and B. Ursin, "Application of committee machines in reservoir characterisation while drilling: a novel neural network approach in log analysis," in *Proceedings of the Nordic Symposium on Petrophysics, Trondheim, Norway*, 2001, pp. 1-15.
[260] S.-I. Amari, N. Murata, K.-R. Muller, M. Finke, and H. H. Yang, "Asymptotic statistical theory of overtraining and cross-validation," *IEEE Transactions on Neural Networks,* vol. 8, pp. 985-996, 1997.
[261] I. Štajduhar, B. Dalbelo-Bašić, and N. Bogunović, "Impact of censoring on learning Bayesian networks in survival modelling," *Artificial Intelligence in Medicine,* vol. 47, pp. 199-217, 2009.
[262] I. H. Witten and E. Frank, *Data mining : practical machine learning tools and techniques*, 2nd ed. Amsterdam ; Boston, MA: Morgan Kaufman, 2005.

[263] A. Karthikesalingam, Holt, P. J., Vidal-Diez, A., Choke, E. C., Patterson, B. O., Thompson, L. J., Ghatwary, T., Bown, M. J., Sayers, R. D., Thompson, M. M., "Predicting aortic complications after endovascular aneurysm repair," *British Journal of Surgery,* vol. 100, pp. 1302-11, Sep 2013.

[264] P. G. Johnson, C. R. Chipman, S. S. Ahanchi, J. H. Kim, D. J. Dexter, and J. M. Panneton, "A case-matched validation study of anatomic severity grade score in predicting reinterventions after endovascular aortic aneurysm repair," *Journal of Vascular Surgery,* vol. 58, pp. 582-588, 2013.

[265] H.-J. Kim, "Common factor analysis versus principal component analysis: choice for symptom cluster research," *Asian Nursing Research,* vol. 2, pp. 17-24, 2008.

[266] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics,* vol. 2, pp. 433-459, 2010.

[267] I. T. Jolliffe, *Principal component analysis*, 2nd ed. New York: Springer, 2002.

[268] H. Abdi, "Factor rotations in factor analyses," in *Encyclopedia of Social Science Research Methods.* , ed Thousand Oaks, CA: SAGE Publications, Inc., 2003, pp. 792-795.

[269] A. M. Molinaro, R. Simon, and R. M. Pfeiffer, "Prediction error estimation: a comparison of resampling methods," *Bioinformatics,* vol. 21, pp. 3301-3307, 2005.

[270] C. D. Dziuban and E. C. Shirkey, "When is a correlation matrix appropriate for factor analysis? Some decision rules," *Psychological Bulletin,* vol. 81, p. 358, 1974.

[271] S. Tobias and J. E. Carlson, "Brief report: Bartlett's test of sphericity and chance findings in factor analysis," *Multivariate Behavioral Research,* vol. 4, pp. 375-377, 1969.

[272] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *IJCAI'95 Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Montreal, Canada, 1995, pp. 1137-1145.

[273] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Statistics Surveys,* vol. 4, pp. 40-79, 2010.

[274] L. Brown, R. Greenhalgh, J. Powell, and S. Thompson, "Use of baseline factors to predict complications and reinterventions after endovascular repair of abdominal aortic aneurysm," *British Journal of Surgery*, vol. 97, pp. 1207-1217, 2010.

[275] N. Peppelenbosch, J. Buth, P. L. Harris, C. van Marrewijk, G. Fransen, and E. Collaborators, "Diameter of abdominal aortic aneurysm and outcome of endovascular aneurysm repair: does size matter? A report from EUROSTAR," *Journal of Vascular Surgery,* vol. 39, pp. 288-297, 2004.

[276] B. W. Lo, R. L. Macdonald, A. Baker, and M. A. Levine, "Clinical outcome prediction in aneurysmal subarachnoid hemorrhage using bayesian neural networks with fuzzy logic inferences," *Computational and Mathematical Methods in Medicine,* vol. 2013, pp. 1-11, 2013.

[277] T. R. Wyss, F. Dick, L. C. Brown, and R. M. Greenhalgh, "The influence of thrombus, calcification, angulation, and tortuosity of attachment sites on the time to the first graft-related complication after endovascular aneurysm repair," *Journal of Vascular Surgery,* vol. 54, pp. 965-971, 2011.

[278] P. W. Stather, J. B. Wild, R. D. Sayers, M. J. Bown, and E. Choke, "Endovascular aortic aneurysm repair in patients with hostile neck anatomy," *Journal of Endovascular Therapy,* vol. 20, pp. 623-637, 2013.

[279] J. Peláez, J. Doña, J. Fornari, and G. Serra, "Ischemia classification via ECG using MLP neural networks," *International Journal of Computational Intelligence Systems*, vol. 7, pp. 344-352, 2014.
[280] H. Yýldýrým, Altýnsoy, Hasan Baki, Barýpçý, Necaattin, Ergün, Uçman, Oğur, Erkin, Hardalaç, Fırat, Güler, İnan, "Classification of the frequency of carotid artery stenosis with MLP and RBF neural networks in patients with coroner artery disease," *Journal of Medical Systems*, vol. 28, pp. 591-601, 2004.
[281] J. Xie and C. Wang, "Using support vector machines with a novel hybrid feature selection method for diagnosis of erythemato-squamous diseases," *Expert Systems with Applications*, vol. 38, pp. 5809-5815, 2011.
[282] G.-Z. Li, T.-Y. Liu, and V. S. Cheng, "Classification of brain glioma by using SVMs bagging with feature selection," in *Data Mining for Biomedical Applications*, ed: Springer, 2006, pp. 124-130.
[283] Y. Yu, Z. Yu, and J. Zhao, "Computer science for environmental engineering and ecoinformatics," in *Proceedings of International Workshop on Communications in Computer and Information Science*, Kunming, China, 2011, pp. 158-159.
[284] H. Kim and M. Bredel, "Feature selection and survival modeling in The Cancer Genome Atlas," *International Journal of Nanomedicine*, vol. 8, p. 57, 2013.
[285] K. Shinjo, Okamoto, Yasuyuki, An, Byonggu, Yokoyama, Toshihiko, Takeuchi, Ichiro, Fujii, Makiko, Osada, Hirotaka, Usami, Noriyasu, Hasegawa, Yoshinori, Ito, Hidemi, "Integrated analysis of genetic and epigenetic alterations reveals CpG island methylator phenotype associated with distinct clinical characters of lung adenocarcinoma," *Carcinogenesis*, vol. 33, pp. 1277-1285, 2012.
[286] N. Tomašev, M. Radovanović, D. Mladenić, and M. Ivanović, "Hubness-based fuzzy measures for high-dimensional k-nearest neighbor classification," *International Journal of Machine Learning and Cybernetics*, vol. 5, pp. 445-458, 2014.
[287] W. A. Chaovalitwongse, Y.-J. Fan, and R. C. Sachdeo, "On the time series k-nearest neighbor classification of abnormal brain activity," *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, vol. 37, pp. 1005-1016, 2007.
[288] D. Lowsky, Ding, Y, Lee, DKK, McCulloch, CE, Ross, LF, Thistlethwaite, JR, Zenios, SA, "AK-nearest neighbors survival probability prediction method," *Statistics in Medicine*, vol. 32, pp. 2062-2069, 2013.
[289] W. D. Jordan, Ouriel, Kenneth, Mehta, Manish, Varnagy, David, Moore, William M, Arko, Frank R, Joye, James, de Vries, Jean-Paul PM, "Outcome-based anatomic criteria for defining the hostile aortic neck," *Journal of Vascular Surgery*, 2015.
[290] G. A. Antoniou, G. S. Georgiadis, S. A. Antoniou, G. Kuhan, and D. Murray, "A meta-analysis of outcomes of endovascular abdominal aortic aneurysm repair in patients with hostile and friendly neck anatomy," *Journal of Vascular Surgery*, vol. 57, pp. 527-538, 2013.
[291] A. Karthikesalingam, S. Markar, P. Holt, and R. Praseedom, "Meta-analysis of randomized controlled trials comparing laparoscopic with open mesh repair of recurrent inguinal hernia," *British Journal of Surgery*, vol. 97, pp. 4-11, 2010.
[292] A. Patel, R. Edwards, and S. Chandramohan, "Surveillance of patients post-endovascular abdominal aortic aneurysm repair (EVAR). A web-based survey of practice in the UK," *Clinical Radiology*, vol. 68, pp. 580-587, 2013.

[293] S. R. Walsh, T. Y. Tang, and J. R. Boyle, "Renal consequences of endovascular abdominal aortic aneurysm repair. ," *Journal of Endovascular Therapy,* pp. 73-82, 2008.
[294] R. A. Weerakkody, S. R. Walsh, C. Cousins, K. E. Goldstone, and M. E. Gaunt, "Radiation exposure during endovascular aneurysm repair.," *British Journal of Surgery,* vol. 95, pp. 699-702, 2008.
[295] J. L. Schafer and J. W. Graham, "Missing data: our view of the state of the art," *Psychological Methods,* vol. 7, p. 147, 2002.
[296] M. M. Rahman and D. Davis, "Addressing the class imbalance problem in medical datasets," *International Journal of Machine Learning and Computing,* vol. 3, pp. 224-228, 2013.
[297] L. Yu, Liu, Huan, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *Proceedings of The Twentieth International Conference on Machine Learning (ICML),* Washington, DC, 2003, pp. 856-863.
[298] J. Cheng , Greiner, R., "Learning Bayesian Belief Network Classifiers: Algorithms and System," in *In Advances in Artificial Intelligence ,14th Biennial Conference of the Canadian Society for Computational Studies of Intelligence, AI,* Ottawa, Canada, 2001, pp. 141-151.
[299] T. A. Stephenson "An introduction to Bayesian network theory and usage," IDIAP Research Report 00-032000.
[300] N. Friedman, I. Nachman, and D. Peér, "Learning bayesian network structure from massive datasets: the «sparse candidate «algorithm," in *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, Stockholm, Sweden 1999, pp. 206-215.
[301] L. M. De Campos, "A scoring function for learning bayesian networks based on mutual information and conditional independence tests," *The Journal of Machine Learning Research,* vol. 7, pp. 2149-2187, 2006.
[302] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Machine Learning,* vol. 29, pp. 131-163, 1997.
[303] G. F. Cooper and E. Herskovits, "A Bayesian method for the induction of probabilistic networks from data," *Machine Learning,* vol. 9, pp. 309-347, 1992.
[304] Z. Liu, B. Malone, and C. Yuan, "Empirical evaluation of scoring functions for Bayesian network model selection," *BMC Bioinformatics,* vol. 13, p. S14, 2012.
[305] B. B. Perry, "A genetic algorithm for learning Bayesian network adjacency matrices from data," PhD, Kansas State University, 2003.
[306] S. Baghdadi, G. Gravier, C.-h. Demarty, and P. Gros, "Structure learning in a bayesian network-based video indexing framework," in *IEEE International Conference onMultimedia and Expo,* Hannover,Germany, 2008, pp. 677-680.
[307] M. Janžura and J. Nielsen, "A simulated annealing-based method for learning Bayesian networks from statistical data," *International Journal of Intelligent Systems,* vol. 21, pp. 335-348, 2006.
[308] J. Cheng and R. Greiner, "Comparing Bayesian network classifiers," in *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, Stockholm, Sweden, 1999, pp. 101-108.
[309] A. M. Hassan, "A probabilistic relaxation framework for learning Bayesian network structures from data," Master of Science Faculty of Engineering, Cairo University 2007.
[310] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Transactions on Neural Networks,* vol. 10, pp. 988-999, 1999.
[311] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning,* vol. 20, pp. 273-297, 1995.

[312] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121-167, 1998.
[313] B. Scholkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. Cambridge, MA, USA: MIT press, 2001.
[314] H.-T. Lin and C.-J. Lin, "A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods," National Taiwan University, Technical Report,March 2003.
[315] B. O. Patterson, R. J. Hinchliffe, P. J. Holt, I. M. Loftus, and M. M. Thompson, "Importance of aortic morphology in planning aortic interventions," *Journal of Endovascular Therapy*, vol. 17, pp. 73-7, Feb 2010.
[316] B. O. Patterson, P. J. Holt, R. Hinchliffe, I. M. Nordon, I. M. Loftus, and M. M. Thompson, "Existing risk prediction methods for elective abdominal aortic aneurysm repair do not predict short-term outcome following endovascular repair," *Journal of Vascular Surgery*, vol. 52, pp. 25-30, Jul 2010.
[317] T. Ghatwary, Karthikesalingam, A., Patterson, B., Hinchliffe, R., Morgan, R., Loftus, I., Salem, A., Thompson, M. M., Holt, P. J., "St George's Vascular Institute Protocol: an accurate and reproducible methodology to enable comprehensive characterization of infrarenal abdominal aortic aneurysm morphology in clinical and research applications," *Journal of Endovascular Therapy*, vol. 19, pp. 400-14, Jun 2012.
[318] D. G. Altman, *Practical statistics for medical research*. USA: CRC press, 1990.

Appendix A: Machine learning Classifiers

A.1 Bayesian Networks

It is a probabilistic network that uses probability theory to calculate joint probabilities between variables. It is also considered as a graphical model, as it looks like a graph representing relations between numbers of nodes (also called vertices) which correspond to random variables. These nodes may be connected to each other by arcs or not. When any couple of nodes is not connected, this indicates a conditional independency between the two variables standing for them and vice versa. Hence, this type of network presents a group of conditional dependencies and independencies between all variables of a dataset used to create the network itself [127].

Bayesian network has two parts, qualitative and quantitative. The qualitative part is related to its structure being a graphical model, how vertices represent variables, and how arcs show the relations between these variables. On the other hand, the quantitative part explains the strength of relationship between variables by using conditional probability tables.

A.1.1 Bayesian Network Structure

A Bayesian network *B(G,Θ)* follows a particular structure of graphical models known as directed acyclic graph (DAG given symbol *G*). The symbol *Θ* refers to the group of parameters of this network. DAG *G=(V,A)* consists of a number of nodes (vertices) *V*={$V_1, V_2, V_3, ...V_n$ }and arcs $A \subseteq V \times V$ connecting them which represent the joint probability distribution and dependency between the variables. Each node represents a specific random variable and is drawn as a circle with its name on it. Arcs connecting the nodes are drawn as arrows and must be directed in only one direction. they do not follow a cyclic direction which means that when an arc leaves a node, it does not return to it again [298]. Absence of arcs between two nodes indicates the presence of conditional independence between these nodes (variables).

The idea of Bayesian network as a graphical model can be easily illustrated by the simple following example. Consider, *A, B, C, and D* as four random variables. As shown in Figure A.1, the arcs connecting nodes *AC* and *AB* mean that variables B and C are conditionally dependent on variable A. In other words, node A influences nodes B and C. The probability of them depend only on *A* (*P*(*B/A*) and *P*(*C/A*)) respectively). Notice that, there is no arc joining *B* and *C*; this indicates that they are conditionally independent. The probability of node *B* does not depend on that of node *C* [299]. While for node *D*, there are two arcs connecting it by nodes *B* and *C*, this means that the variable *D* conditionally dependent on both variables *B* and *C* which are their parents. The probability of *D* depend on *B* and *C* as well (*P(D/B,C)*).

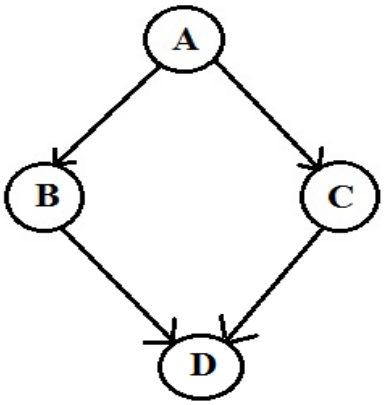


Figure A.1 A Simple Bayesian Network Structure. It shows relation between variables A, B,C,and D which represent the nodes in Bayesian network. The Arcs are used to show if two variables are related or not.

A.1.2 Learning Bayesian Networks

Techniques used for learning Bayesian networks are divided into two tasks; learning the network structure and its parameters (conditional probability tables (CPT)).

A.1.2.1 Learning Bayesian Network Structure

There are two groups of methods used for learning the structure of a Bayesian network. The first one is the search and score based algorithms. A number of networks may be generated from a dataset. Hence, the scoring algorithm task is giving a score to the network that fits this data. Afterwards, optimizing is needed to search and select the network that best fits the data which has the highest score [300]. The second group is the constraint based algorithms. Constraint based methods have two basic concerns that may lead to unreliable results which are the complexity and number of independent tests. Therefore score and search methods may be favored.

A.1.2.1.1 Score and Search Based Algorithms

The learning process of Bayesian network structure can be defined as; given a dataset *D* and Bayesian network with a DAG structure *G*, search to select the best *G* (denoted as *G**) that fits the network. These methods depend on scoring function that gives each candidate network a score equivalent to the goodness of fit between the network and data. This is done using Equation (A.1) where; $Score(G, D)$ is a score given to each generated network *G* from the dataset *D*, and G_n is the set of generated networks (DAGs) with *n* nodes. Scoring methods are combined with a searching strategy to evaluate the goodness of each candidate network from a pool of all candidates.

$$G^* = \arg\max_{G \in G_n} Score(G, D). \tag{A.1}$$

Scoring Functions

Many score and search based methods have been used for training Bayesian networks which can be classified into two classes; information and Bayesian measures. In this section the most used techniques will be discussed.

- **Minimum Description Length (MDL) scoring function** is the common used and well known information scoring measure which uses the principle of data compression in

order to learn a Bayesian network [301]. According to information theory which indicates the coding of a message (a network in our case) tries to minimize the number of elements used to represent this message. Regular messages will therefore be given shorter codes and less regular ones will have longer codes. In the same manner an MDL tries to find out a network model which expresses the minimum description length of the network model and the minimum description length of the dataset given the model. Since Bayesian networks provide a probability distribution P_B among instances of a dataset *D*, an encoding scheme can be constructed using this distribution which gives shorter code words to more likely instances. MDL then, select a network *B* which has minimum combined length of both network description and the encoded data [302]. The MDL scoring function given a training data can be defined using equation A.4 where; $|B|$ is the number of parameters in the network *B*, ½ log N are bits used for each parameter, *LL* is the log likelihood of *B* given *D* defined in Equation (A.5).

$$MDL(B \setminus D) = \frac{\log N}{2} |B| - LL(B \setminus D), \tag{A.2}$$

$$LL(B \setminus D) = \sum_{i=1}^{N} \log(P_B(u_i)). \tag{A.3}$$

- **Bayesian Scoring Functions** calculate the posterior probability distribution with a prior probability distribution of all generated networks given data *D*. The network which has the higher posterior probability is the one selected. Bayesian Dirichlet (BD) and its extensions; Bayesian Dirichlet score equivalence (BDe) and Bayesian Dirichlet score equivalence and uniform priors (BDeu) are the most well-known Bayesian scoring function. The BD calculates the joint probability between variables of a dataset. It was introduced by Cooper and Herskovits [303]. The main drawback in BD metric is a parameter for each parent-child combination should be specified by the user. In addition, identical network structures are not given the same score. Therefore, BDe metric was presented which introduced a hyper-parameter known as equivalent sample size. All needed parameters are estimated from this hyperparameter and the prior probability distribution of all generated networks *D*. it is a complex task, hence, BDeu was proposed which assumes the generated networks have the same prior probability distribution. However, this assumption is not always true. BDeu calculation is very sensitive to the hyperparamter, so when the network density is unknown, then choosing proper hyperparameter will be a complicated task [304].
- **Other metrics** depends on the asymptotic performance of models with adequately large datasets. AIC and BIC are examples of these metrics [304].

Searching Methods

As mentioned before searching is combined with a scoring function in order to find the best network structure that fits the data. The size of the space of all structure is exponentially increasing with the number of nodes (variables). Therefore, the exhaustive search of all possible structure will be very difficult. Alternatively heuristic search methods are more preferred which search structure space by iteratively making a small modification in the network structure. These searching methods will be discussed below along with simulated annealing meta-heuristic search strategy.

- **Hill Climbing Searching Method**

This is a heuristic greedy search algorithm where the search process can start with one of three graph states. An initial graph is suggested which may be to either have no initial nodes, randomly created, or produced with the help of an algorithm [305]. It is the most widely used searching strategy to learn the structure of BN. The hill climbing method starts searching the structure of the network by checking with neighbour nodes for the possibility of adding,

removing or reversing an arc to produce a network with DAG structure *G*. Afterwards, a set of networks G_n will be generated with different scores, the one that improves the scoring metric is chosen. This process is repeated several times. Finally, the procedure is terminated when there is no further improvement made on the scoring metric by changes done to any single arc [131].

- **K2 Searching Method**

This is a heuristic greedy algorithm that makes three assumptions to perform the search. These assumptions are; the previous ordering on the variables is known, the prior distributions of the all network structures are equal and the number of maximum parents of a node is available. It begins by supposing that all nodes have no parents, and then it starts with the first node. Afterwards, all subsequent nodes are added iteratively to check all possibilities for constructing the network. Next, each possibility is given a score using a scoring function. Finally, the combination with highest score is selected. The procedure is iterated until the addition of a new node does not improve the scoring metric. The disadvantage of this algorithm is that the initial ordering of the nodes may reflect on the structure of the generated network. In other words, it depends mainly on the initial nodes order [306].

- **Simulated Annealing Searching Method**

The idea of simulated annealing was first used in the process of annealing of metals in which the solid was heated in order to melt then left to cool with gradual decrease in temperature to produce a new optimal shape with lower defects [307]. Similarly, this idea was adopted in machine learning for learning a Bayesian network structure. The searching process of simulated annealing may progress in an unfavourable worse path to get out of local optima. For example, suppose a network structure with an initial temperature T_o and score. For search iteration, new network structure is formed and the variation in the score achieved is Δ. If Δ has improved the score, then this network is accepted (favourable case). Otherwise, it may be also accepted (unfavourable case) but with a probability equals $e^{-\frac{\Delta}{T}}$ to reduce chances of local optima; where *T* is the annealing temperature which decreases regularly with time [305]. Network's structure selection using a simulated annealing algorithm depends mainly on *T*. Initially, *T* begins with a large value indicating higher chances of accepting unfavourable worse solutions while searching the structure space. This will help for not falling to local optima. As the searching process continues, *T* decreases and unfavorable structures accepted probability decreases too [305]. This procedure terminates when reaching a predefined temperature.

- **Tree Augmented Naïve Bayesian (TAN) Searching Method**

This is a technique used for learning Bayesian networks. First, it learns a network as a tree structure by connecting the class node (parent node) to all other nodes (children nodes). Then, it starts adding an arc between each couple of these children nodes and see whether it is possible to have an extra parent to this node. Next, it calculates a score for each possibility using a scoring metric. The procedure is iterated until the addition of an arc does not improve the score [308].

A.1.2.1.2 Constraint Based Algorithms

These are the type of algorithms that consider some constraints that must be satisfied in order to learn Bayesian network structure. First, they figure out dependency and independency relations between attributes of a dataset. Then, they use tests in order to discover these dependencies to form a structure that fits the data. These tests may be statistical based like X^2 test or information theory based like mutual information test. The drawback of these algorithms is that repeated

tests for dependencies are subject to failure and the high complexity of these tests [309]. Therefore, they were not used in this work.

A.1.2.2 Learning Bayesian Network Parameters

For learning the parameters of a Bayesian network, the structure may be known already or learned using one of the previously mentioned algorithms. The target of learning the network parameter is to figure out the values of the parameters which maximize an estimation using the training data. The estimator used in parameter learning for Bayesian networks constructed in this thesis is maximum likelihood. Learning parameters of a given Bayesian network *G* with nodes $X_1, X_{2,} X_3, ...X_n$ is calculated with a joint probability distribution between all nodes using the Equation (A.4), where; *pa* denotes the parents of node X_i, θ is a parameter for each possible value of X_i and pa_{X_i}, and *i* is the variable number [309] .

$$P(X_1,....,X_n) = \prod_{i=1}^{n} P(X_i \setminus pa(X_i)) = \prod_{i=1}^{n} \theta_{X_i \setminus pa_{X_i}}. \quad (A.4)$$

As mentioned before, the target of learning here is to get the most possible parameter value $\hat{\theta}$ that best fits *D*. The log likelihood function L_D can be used to measure this using Equation (A.5), where; $pa_i[x_i(m)]$ represents the i^{th} parent of variable $x_i(m)$ and *m* is the number of examples in the training data.

$$L_D(\theta) = \log\{\prod_{m=1}^{N} P(x_1[m],.....x_n[m]:\theta)\} = \log\{\prod_{i=1}^{n}\prod_{m=1}^{N} P(x_i[m] \setminus pa_{i(}[x_i(m)]:\theta)\}. \quad (A.5)$$

The maximum likelihood method is used to get $\hat{\theta}$ using Equation (A.6).

$$\hat{\theta} = \arg\max_{\theta} L_D(\theta). \quad (A.6)$$

A.2 Support Vector Machine

The support vector machine (SVM) is a popular machine learning classifier [281]. It was first introduced by Vapnik as a binary classifier, and then extended for regression and multiclass classification. The basic idea behind the SVM is to transform the input vector which is not linearly separable using a hyper plane into a higher dimensional feature space capable of linearly separating between classes of input data to facilitate classification. This is done with the help of a kernel function which maps the similarity between the input vector and the new higher dimension feature space.

A.2.1 Statistical theory

In this section, the basic necessary underlying theory of the two classes SVM will be introduced briefly. Their details can be found in [310, 311]. Suppose an *N* training examples *x* with class values $y\ \{x_1, y_1\},.....,\{x_N, y_N\}$, where $x_i \in R^m$ which is the *m* dimensional input feature vector corresponding to the i^{th} training example and $y \in \{+1,-1\}$. Training of the SVM is made in order to find the optimal hyper-plane that separates the two classes of input data. The hyper-plane can be defined using Equation (A.7) where *w* is the weight vector to the hyper-plane and *b* is a bias.

$$w^T \cdot x + b = 0. \quad (A.7)$$

Note that, the nearest training examples to the hyper-plane are called support vectors. The distance between the hyper-plane and support vectors is called the margin. When the training data is linear, it will be easy to find the pair (*w*,*b*) corresponding to the optimal hyper-plane using (A.8) and (A.9). Therefore, training SVM will be a very simple task and the optimal hyper plane can be found without any training errors, and it will maximize the margin.

Minimize: $$\frac{1}{2}\|w\|^2, \quad \text{(A.8)}$$

Subject to: $$y_i(w^T \cdot x_i + b) \geq 1, i = 1,2,....,N. \quad \text{(A.9)}$$

Other cases where data are not linear, there is no direct linear hyper-plane that could be found. The classification task become more complex, therefore in these cases, training of the SVM employs kernel function *K* which transforms the original input vectors into new higher dimension feature space that facilitate separating between the two classes of data linearly and find the optimal linear hyper-plane. In other words it maps input vector by nonlinearly mapping function $\Phi(x)$ to a higher order feature space. The new optimization problem is solved using Equations (A.10) and (A.11) where; *C* is the penalty parameter of the error term which is used to balance the width of the margin and training error. ξ is the slack variables which are introduced to allow for some training errors as in case of noisy data firm separation between the two classes is not needed.

Minimize: $$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i, \quad \text{(A.10)}$$

Subject to: $$y_i(w^T \cdot \Phi(x_i) + b) \geq 1, i = 1,2,....,N. \quad \text{(A.11)}$$

The optimization problems discussed above are constrained optimization problems solved by Lagrange multipliers that transform them into quadratic programming problem. Details of Lagrange multipliers α can be found in [312]. Classification decision is done using the function (A.12); the predicted class label is determined by the sign of this function. If it is positive, the predicted class is +1 and vice versa.

$$f(x) = w^T \cdot \Phi(x) + b = \sum_{i=1}^{N}\alpha_i y_i K(x_i, x_j) + b, \quad \text{(A.12)}$$

where; *K* is the kernel function and $\Phi(x)$ is the nonlinearly mapping function.

A.2.2 Kernel Functions

Several kernel functions can be used for feature mapping, however the most common ones are linear, polynomial, radial basis, and sigmoid. Details about kernel function could be found in [313].The choice of kernel function is dependent on the problem and type of data. There is no one specific kernel that can be used for all applications and problems. Usually, several kernels functions are tested and the one which has high impact on classification is the one chosen. Kernel parameters are usually done using cross validation method. In this thesis, the sigmoid function was the one chosen. The kernel parameter α is commonly determined using $1/M$, where *M* is the number of features [314].

Appendix B: Datasets Description and Analysis

B.1 Dataset Description

Patients that experienced EVAR surgery in two different vascular centers located in the UK with median age of 75 were observed by physicians during the period 2004 till 2010. These centers were located in St George Vascular Institute in London and Leicester Cardiovascular Biomedical Research Unit in Leicester. The major ending point was the progress of aortic complications and the need of EVAR re-intervention. The thoracic inlet to the level of the common femoral artery bifurcation was captured using CT three- dimensional computed tomography (CT). The CT images have a slice thickness of 0.625 mm or 1.25 mm. Morphological features were collected for these patients and used in this work as they have greater effect on aortic complications than physiology features [21, 315, 316]. Both centers have 45 variables. Patients containing missing values were removed. The number of patients after removal is 457 and 286 for center 1 and 2 respectively. Patients that actually experienced the EVAR during the observations are 42 and 26 for center 1 and 2 respectively. A full detailed definition and description of each variable and how it was collected can be found in [317]. Table B.1 gives a brief definition of each variable.

Table B.1 The definitions of morphology variables included in the endovascular aortic aneurysm repair dataset.

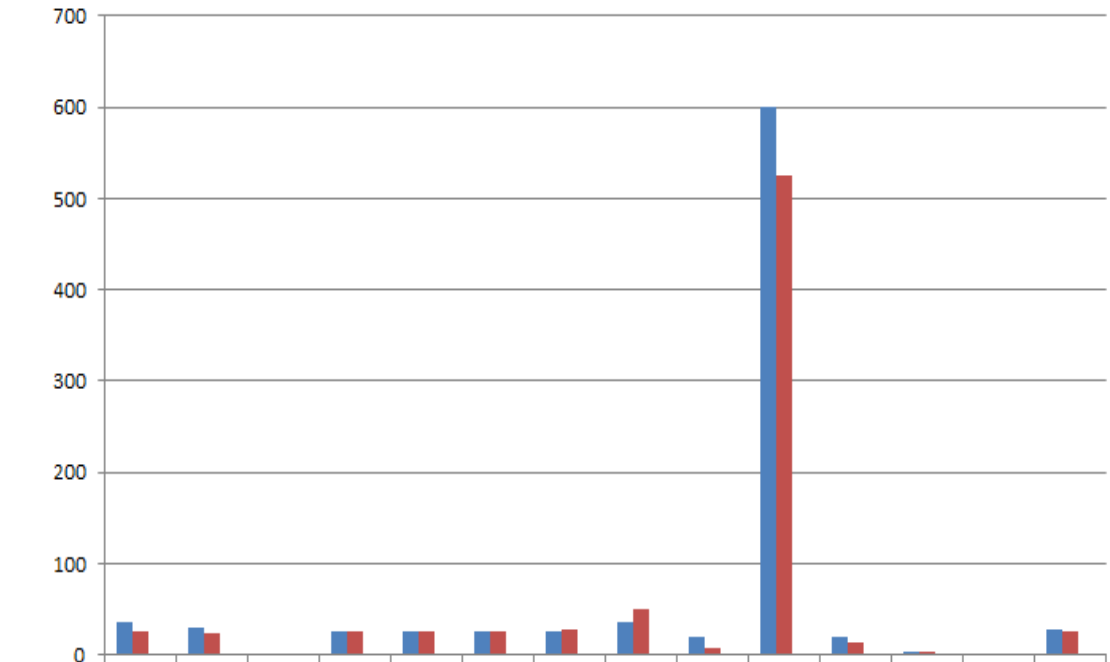
Variable	Symbol	Definition
Aneurysm Neck Length L1	AN.L1	aneurysm neck length, in mm from lowest renal artery to start of aneurysm
Aneurysm Neck Length L2	AN.L2	Distance in mm between the two ends of the neck
Tortuosity Index (L1/L2)	Neck.TI	The ratio between aneurysm neck length L1 and L2
Neck Diameter 1mm below lowest renal	AN D1	The diameter of the aneurysm in mm from 1mm below lowest renal
Neck Diameter 5mm below lowest renal	AN D5	The diameter of the aneurysm in mm from 5 mm below lowest renal
Neck Diameter 10 mm below lowest renal	AN D10	The diameter of the aneurysm in mm from 10 mm below lowest renal
Neck Diameter 15 mm below lowest renal	AN D15	The diameter of the aneurysm in mm from 15 mm below lowest renal
Maximum Aneurysm Neck Diameter	Max AN.D	Maximum diameter of the aneurysm among all measurement's
Maximum Neck Area	Max AN.A	The area of the maximum diameter of the aneurysm
Neck Angulation	AN.Gamma	Aneurysm neck angulation from centerline at renal arteries, to centerline at commencement of aneurysm.
Neck Distal Angle	AN.Alpha	Distal neck angle between aneurysm neck and sac
Total Neck Volume	AN.WWVol	Total volume of the aortic neck from external wall to the other wall
Neck Thrombus Volume	AN.NonlumnialVol	Volume of the thrombus to the wall of the

		aneurysm
Calcification Index	AN.CircumCalcification	percentage of circumferential calcium deposition in the aneurysm neck segment
Max Sac Diameter	MaxAA.D	The maximum diameter of the sac
Maximum Sac Area	MaxAA.A	External wall to external wall area for maximum diameter
Aneurysm sac length	AA.Length	Distance in mm between the aortic aneurysm neck and the aortic bifurcation
Sac Thrombus Volume	AA.NonluminalVol	The volume of the area between the thrombus and the wall of the aneurysm
Sac Volume	AA.WWVol	The volume of the aortic aneurysm sac
Sac Tortuosity Index	AA.TI	The ratio between the distance between the aortic aneurysm neck and the aortic bifurcation and the straight between these points
Suprarenal Aortic Diameter	SR.A.D	Maximum external wall to external wall diameter 15 mm above the aneurysm neck
Maximum Common Iliac Artery Diameter, 1mm proximal to Internal Iliac Origin	Max.Cia.D1	The largest diameter of the common iliac artery 1 mm away from internal iliac
Maximum Common Iliac Artery Diameter, 5mm proximal to Internal Iliac Origion	Max.Cia.D5	The largest diameter of the common iliac artery 5 mm away from internal iliac
Maximum Common Iliac Artery Diameter, 10mm proximal to Internal Iliac Origin	Max.Cia.D10	The largest diameter of the common iliac artery 5 mm away from internal iliac
Maximum Common Iliac Artery Area	Max.Cia.A	The area of the maximum diameter amoung measured ones.
Maximum Common Iliac Thrombus Volume	Max.Cia.NonliminalVol	Maximum volume of the thrombus and the wall of the common iliac artery
Maximum Common Iliac Artery Volume	Max.CiaWWVol	Largest total volume of the common iliac artery
Maximum Iliac Tortuosity Index	Max.ITI	The ratio between centerline length and the straight-line length between the artery otium and bifurcation.
Right iliac artery Tortuosity Index	R.I.TI	The ratio between centerline length and the straight-line length between the aortic and femoral bifurcation.
Diameter of the right common iliac artery 1mm below Internal iliac ostium	R.CIA.D1	The diameter in mm of the right common iliac artery, measured 1 mm proximal to the Internal iliac ostium
Diameter of the right common iliac artery 5mm below Internal iliac ostium	R.CIA.D5	The diameter in mm of the right common iliac artery, measured 5 mm proximal to the Internal iliac ostium
Diameter of the right common iliac artery 10 mm below Internal iliac ostium	R.CIA.D10	The diameter in mm of the right common iliac artery, measured 10 mm proximal to the Internal iliac ostium
Maximum area of the right common iliac artery	MaxR.CIA.A	largest right common iliac artery area

Right common iliac artery volume	R.CIA.Vo	The volume of the right common iliac artery from the distal point of the ostium
Right common iliac artery non luminal volume	R.CIA.NonluminalVo	The volume of the thrombus from the wall of the right common iliac artery
Maximum right iliac angulation	R.IliacMaxAngulation	The largest angulation of the right iliac artery
Left iliac artery Tortuosity Index	L.I.TI	The ratio between centerline length and the straight-line length between the aortic and femoral bifurcation of the left iliac artery.
Diameter of the left common iliac artery 1mm below Internal iliac ostium	L.CIA.D1	The diameter in mm of the left common iliac artery, measured 1 mm proximal to the Internal iliac ostium
Diameter of the left common iliac artery 5mm below Internal iliac ostium	L.CIA.D5	The diameter in mm of the left common iliac artery, measured 5 mm proximal to the Internal iliac ostium
Diameter of the left common iliac artery 10 mm below Internal iliac ostium	L.CIA.D10	The diameter in mm of the left common iliac artery, measured 10 mm proximal to the Internal iliac ostium
Maximum the left common iliac artery	L.Max.CIA.A	Largest left common iliac artery area
left common iliac artery volume	L.CIA.Vo	The volume of the left common iliac artery from the distal point of the ostium
Right common iliac artery non luminal volume	L.CIA.NonluminalVo	The volume of the thrombus from the wall of the left common iliac artery
Maximum Left iliac angulation	LIliacMaxAngulation	The largest angulation of the left iliac artery
Maximum iliac angulation	Max_IliacAngulation	The largest angulation of the iliac artery

B.2 Datasets Analysis

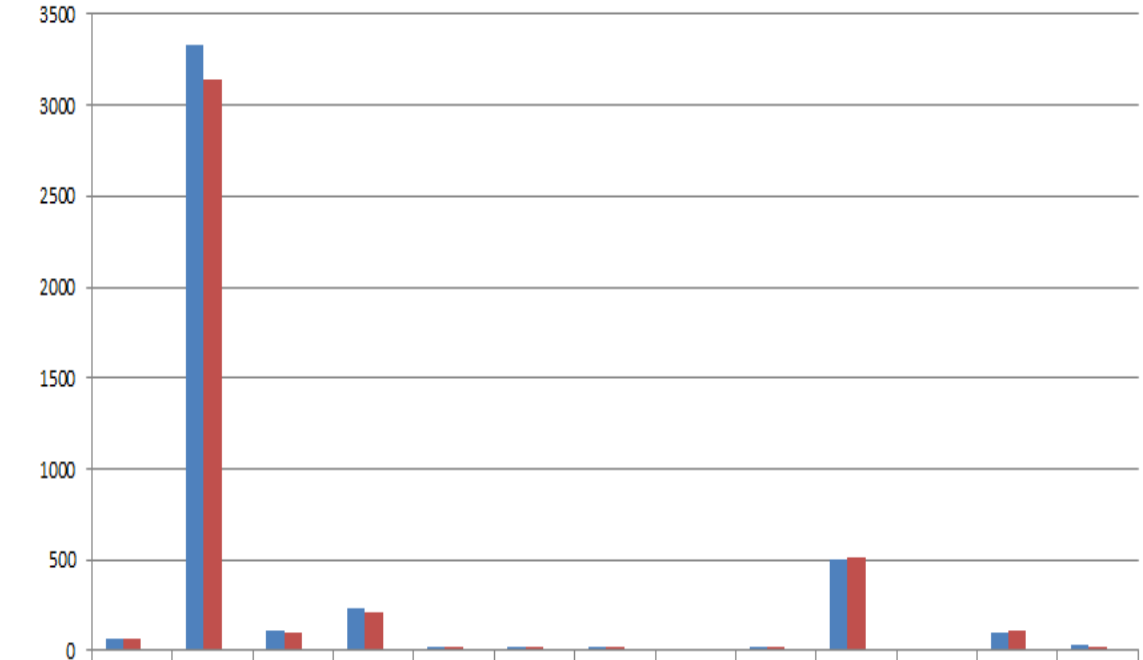
Univariate analyses were done for centers, using Chi-Squared or Fisher's exact test for dichotomous data, Mann-Whitney U-test for non-normally distributed continuous data and student's t-test for normally distributed continuous data to show a descriptive analysis of patients of both Centers. These tests examine the null hypotheses which states there no significant difference between the same variable included in two different datasets [318]. Even though there were statistical differences in aneurysm morphology between Centers 1 and 2 as shown in Figures B.1-B.4, these dissimilarities have no major clinical significance [263]. Figure B.1 shows that the mean neck lengths L1 and L2 was shorter in center 2 (35.87 vs. 26.47 mm, $p<0.001$ and 30.098 vs. 22.97 mm, $p<0.001$). Mean of neck tortuosity index was almost the same in the two centers (1.224 vs. 1.168, $p=0.44$), Neck mean diameters within 1mm difference are slightly bigger in center 2 (25.13 vs. 25.32 mm $p=0.071$, 25.07 vs. 25.5 mm $p=0.011$, 25.20 vs. 26.22 mm, $p=0.012$, 26.70 vs. 27.26 mm, $p=0.024$ at 1, 5, 10 and 15mm below the lowest renal artery. Mean values of the neck Angulation, distal angle, maximum area, and total volume are different all with p values lower than 0.001. (34.9 vs. 50.7 , 18.969 vs. 7.922, 600.39 vs. 524.3, and 20.414 vs.14.15). While, mean neck thrombus volume and calcification index are nearly the same (3.48 vs. 3.02, $p=0.078$ and 0.143 vs. 0.183, $p=0.793$) as well as, neck maximum diameter (27.158 vs. 26.70, $p=0.787$).



	AN. L1	AN.L2	Neck.TI	AN.D1	AN.D5	AN.D10	AN.D15	An.Gamma	AN.Alpha	Max AN.A	AN.WWVol	AN.Nonluminal.Vol	AN.CircumCalcification	Max AN.D
Center 1	35.87	30.098	1.224	25.13	25.07	25.2	26.7	34.9	18.97	600.39	20.414	3.48	0.143	27.158
Center 2	26.47	22.97	1.168	25.32	25.5	26.22	27.26	50.7	7.922	524.3	14.15	3.02	0.183	26.7
P-value	0.001	0.001	0.44	0.071	0.011	0.012	0.024	0.001	0.001	0.001	0.001	0.078	0.0793	0.787

Figure B.1 The mean values of the first groups of variables of both centers along with the p-value results from the statistical tests testing similarity among variables.

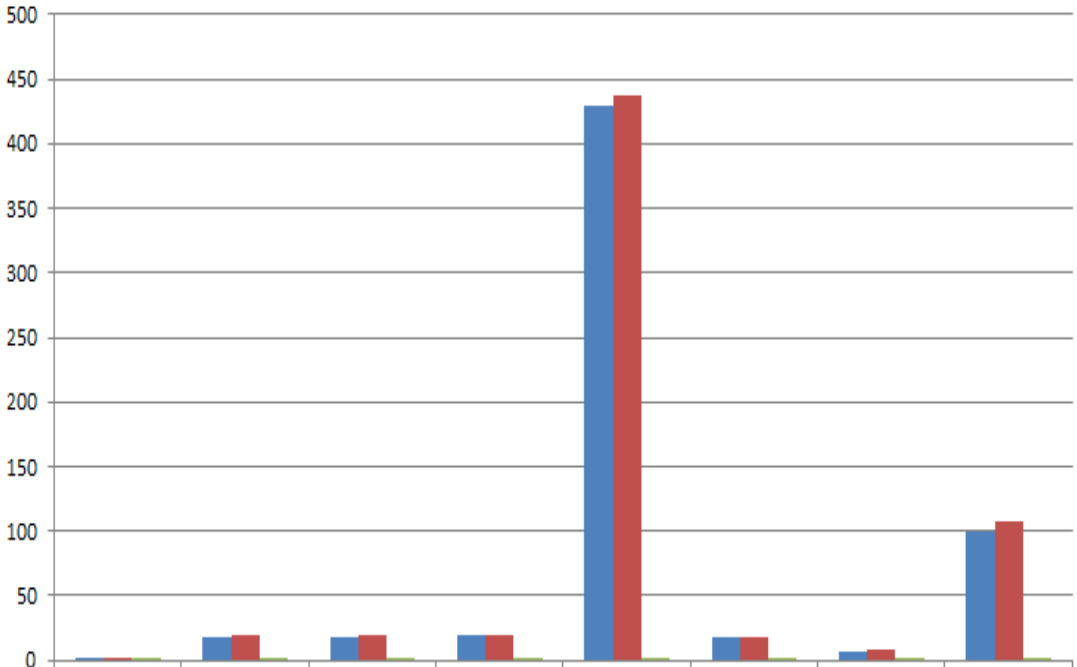
Figure B.2 shows that the mean value of maximum sac aneurysm diameter, area, and thrombus volume are (66.2 vs. 64.6 mm, p=0.233, and 3333.3 vs. 3141, p=0.067, and 115.1 vs. 103.5, p=0.51). Sac total volume is nearly the same for the two centers (233.7 vs. 209., p=0.141).while, the maximum common iliac artery diameter was larger in center 2 (20.01 vs. 18.88 mm, p< 0.001, 20.7 vs. 19.8 mm, p < 0.001,and 22.19 vs. 20.55 mm, p=0.002, at 1,5, and 10mm proximal to internal iliac origin). The Maximum Common Iliac Thrombus Volume, artery area, and aneurysm sac length are bigger in center 2 as well (9.78 vs. 9, p< 0.001, 515.823 vs. 498, p=0.185, and 107 vs. 104.3, p=0.015). Suprarenal Aortic Diameter is close for both centers (27.90 vs. 26.857 mm, p=0.055).



	Max AA.D	Max AA.A	AA.NonluminalVol	AA.WWVol	Max.Cia.D1	Max.Cia.D5	Max.Cia.D10	Max.Cia.Nonluminal Vol	Max.CiaWWvol	Max.Cia.A	Max.ITI	AA.Length	SR.A.D
Center 1	66.2	3333.3	115.1	233.7	18.88	19.8	20.55	9	21.3	498	1.56	104.3	27.9
Center 2	64.6	3141	103.5	209.8	20.1	20.7	22.19	9.78	21.55	515.8	1.45	107	26.857
P-value	0.233	0.067	0.51	0.141	< 0.001	< 0.001	0.002	<0.001	0.622	0.185	< 0.001	0.015	0.055

Figure B.2 The mean values of the second group of variables of both centers along with the p-value results from the statistical tests testing similarity among variables.

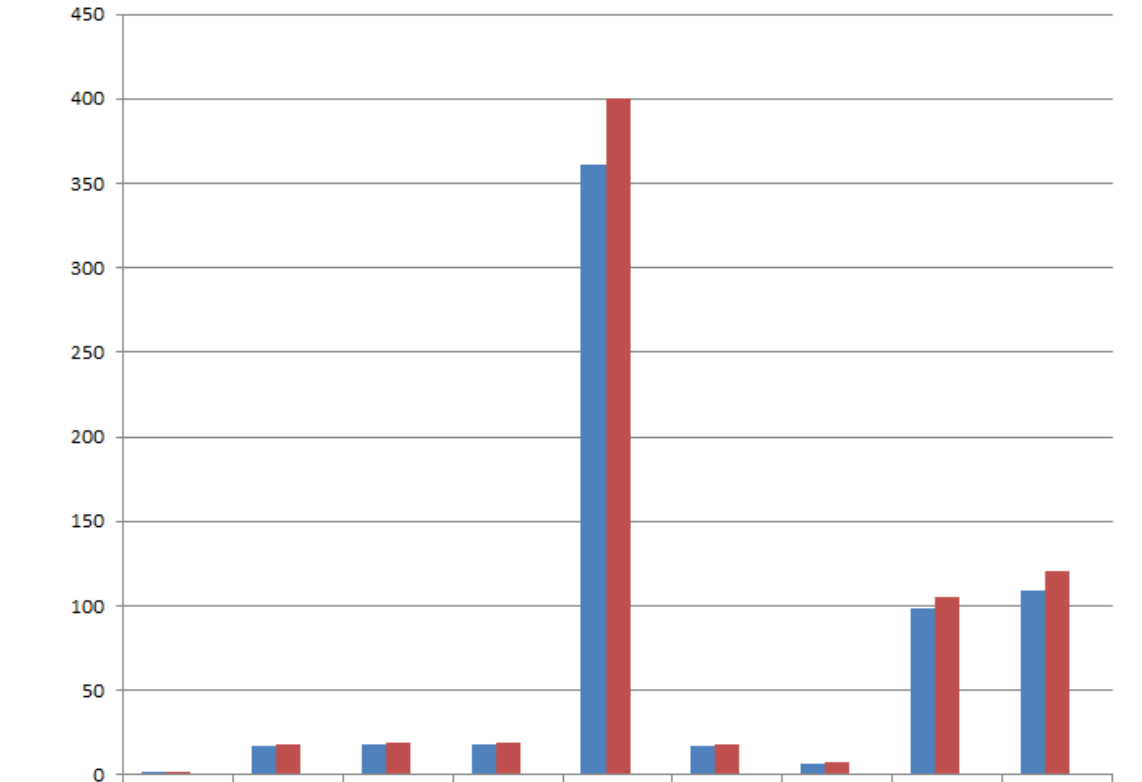
Figure B.3 shows that the mean of the right iliac artery tortuosity index is larger in center 1 (1.46 vs.1.386, p< 0.001), while diameters of the right common iliac artery are slightly bigger in center 2 (18.913 vs. 17.415, p=0.3936, 19.3 vs. 18.188, p=0.5899, and 19.8 vs. 18.839,p= 0.1351 at 1mm,5mm, and 10 mm proximal Internal iliac ostium). Maximum area of the right common iliac artery, right common iliac artery non luminal volume, and maximum right iliac angulation are slightly bigger as well for center 2 (437.007 vs. 429.679, p=0.122, 7.661 vs. 6.869, p=0.892, and 108.069 vs. 99.375, p=0.4889). The mean of the right common iliac artery volume is nearly equal in both centers (17.31vs. 17.495 , p< 0.001).



	RITI	R.CIA.D1	R.CIA.D5	R.CIA.D10	Max.R.CIA.A	R.CIA.Vo	R.CIA.NonluminalVo	R.IliacMAxAngulation
Center 1	1.46	17.415	18.188	18.839	429.68	17.31	6.869	99.375
Center 2	1.386	18.913	19.3	19.8	437	17.49	7.661	108.07
P-value	< 0.001	0.3936	0.5899	0.1351	0.122	< 0.001	0.892	0.4889

Figure B.3 The mean values of the third group of variables of both centers along with the p-value results from the statistical tests testing similarity among variables.

Figure B.4 shows that the mean of the left iliac artery Tortuosity Index is slightly larger in center 1 (1.455 vs. 1.401, p< 0.001), But diameters of the left common iliac artery are marginally bigger in center 2 (18.548 vs. 17.288, p=0.565, 19.057 vs. 17.942, p=0.802, and 19.378 vs. 18.35, p= 0.906 at 1mm,5mm, and 10 mm proximal Internal iliac ostium). Maximum area of the left common iliac artery, left common iliac artery non luminal volume, left common iliac artery volume and maximum left iliac angulation, and maximum iliac angulation are bigger as well for center 2 (399.989 vs. 360.632, p=0.862, 7.48 vs. 6.514, p=0.573, 17.629 vs. 16.934, p-0.03, 105.489 vs. 98.12, p=0.908, 121.069 vs. 109.584, p= 0.1683).



	L.ITI	L.CIA.D1	L.CIA.D5	L.CIA.D10	L.CIA.A	L.CIA.WWVo	L.CIA.NonluminalVo	L.IliacMaxAngulation	Max_IliacAngulation
Center 1	1.455	17.288	17.942	18.35	360.63	16.934	6.514	98.12	109.58
Center 2	1.401	18.548	19.057	19.057	399.9	17.63	7.48	105.48	121.07
P-value	< 0.001	0.0565	0.802	0.802	0.862	0.03	0.573	0.9081	0.1683

Figure B.4 The mean values of the third group of variables of both centers along with the p-value results from the statistical tests testing similarity among variables.