A Data Mining Approach for Predicting Delirium After Cardiac Surgery

By

Hani N. Mufti

Submitted in partial fulfillment of the requirements

for the degree of Master of Health Informatics

At

Dalhousie University

Halifax, Nova Scotia

December 2014

DEDICATION PAGE

In memory of my late aunt, Mona Mufti, you left fingerprints of grace on my life. You shall not be forgotten.

To my three mothers; my mother, Thanaa, my godmother, Abdia, and my late aunt, Mona. Thank you for inspiring me to ask many questions. Your love and kindness will always guide me.

To my father, Nabeel, and my grandfather, Mohammed. Thank you for raising me to become a man of high quality and value. Your words of encouragement and push for tenacity ring in my ears.

To my sisters; Salha and Hanaa. Thank you for your support and love. I will always love you.

To my son Mohammed, thank you for all the hugs and kisses.

To the love of my life, my wife Lama, at my darkest moments you came into my life and gave me hope. You showed me a love that was real and changed my life forever…I could never love anyone but you… Thank you for being my angel…I love you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Of particular concern to patients is the effect of surgery upon brain functions following a surgical intervention. Indeed, post-operative neurocognitive complications occur in up to 60% of patients. These include: stroke, seizures, and delirium. Delirium is a temporary disturbance of consciousness, attention, cognition, and/or perception, which occurs frequently among hospitalized patients. It develops over a short period and tends to fluctuate. Several risk factors predispose patients to post-operative delirium, including: medications, age, male gender, major surgery (cardiac and orthopedic), and others. Delirium occurs relatively frequently (10% to 15%) among patients who undergo cardiac surgery. Patients who experience delirium after cardiac surgeries are at higher risk of multiple adverse outcomes (e.g.: infections, prolonged hospitalization, and death). Identification of patients at risk will allow targeted personalized preventive strategies that might improve the patient transition through the process of care. This thesis demonstrates the development of several predictive models, using a data mining approach, to predict the development of delirium in patients undergoing cardiac surgery. The developed models were derived from a large contemporary registry, and their performance was evaluated on an independent dataset. This work also addresses the issue of class imbalance and its effect on model performance. The findings of this research suggest that, applying machine learning and data mining techniques on complex medical data is capable of achieving superior results in comparison to standard statistical approaches. With increased adaptation of electronic health records, data mining techniques offer novel approaches to aid in the prediction of complex relationships, a typical property of adverse medical events. These models will aid the recovery of high-risk patients by enabling a more proactive approach, initiating preventive measures in a timely fashion.

LIST OF ABBREVIATIONS USED

95% CI	95% Confidence Interval
A-Fib	Atrial Fibrillation
ACEI	Angiotensin Converting Enzyme Inhibitors
ACS	Acute Coronary Syndrome
AF	Atrial Fibrillation
AIHW	Australian Institute of Health and Welfare
AMT	Abbreviated Mental Test
ANN	Artificial Neural Networks
ANN-1 Hidden	Artificial Neural Network with 1 Hidden Layer
ANN-2 Hidden	Artificial Neural Network with 2 Hidden Layers
ARB	Angiotensin Receptor Blockers
ASA	Aspirin
AUC	Area Under the Curve
AVR	Aortic Valve Replacement
BB	Beta Blockers
BBN	Bayesian Belief Networks
BBN-1 Parent	Bayesian Belief Network with 1 Parent
BBN-2 Parents	Bayesian Belief Network with 2 Parents
BGD	Batch Gradient Descent
BPNN	Back-Propagation Neural Networks
CABG	Coronary Arteries Bypass Graft Surgery
CAM	Confusion Assessment Method
CAM-ICU	Confusion Assessment Method in the Intensive Care Unit
CCS	Canadian Cardiovascular Society
CDHA	Capital District Health District Health Authority
CHF	Congestive Heart Failure
COPD	Chronic Obstructive Pulmonary Disease
CPB	Cardio-Pulmonary Bypass
CPT	Conditional Probability Table
CVD	Cerebrovascular Disease
DAG	Directed Acyclic Graph
DLP	Dyslipidemia
DM	Data Mining
DNA	Deoxyribonucleic Acid
DRIP	Data Rich, Information Poor
DSM-5	*Diagnostic and Statistical Manual of Mental Disorders 5th Edition*
ECC	Extra Corporeal Circulation
EM	Expectation-Maximization

ER	Emergency Room
EURO-Score	European System for Cardiac Operative Risk Evaluation Score
EUROII	European System for Cardiac Operative Risk Evaluation II
FN	False Negative
FP	False Positive
GDP	Gross Domestic Product
HR	Hazard Ratio
HRQL	Health Related Quality of Life
HTN	Hypertension
IABP	Intra-aortic Balloon Pump
ICU	Intensive Care Unit
IHD	Ischemic Heart Disease
IV-Hep	Intravenous Heparin
KDD	Knowledge Discovery from Databases
LAGD	Look Ahead in a Good Direction
LR	Logistic Regression
MHC	Maritime Heart Center
MI	Myocardial Infarction
ML	Machine Learning
MMSE	Mini-Mental State Examination
MV-Rep	Mitral Valve Repair
MVR	Mitral Valve Replacement
NYHA	New York Heart Association
OR	Odds Ratio
PCI	Per-cutaneous Intervention
PHC	Personalized health care
PHTN	Pulmonary Hypertension
PRIND	Prolonged Reversible Ischemic Neurologic Deficits
PVD	Peripheral Vascular Disease
QEII-HSC	Queen Elizabeth II Health Sciences Center
RASS	Richmond Agitation Sedation Scale
REB	Research Ethics Board
ROC	Receiver Operator Characteristics
ROC-AUC	Receiver Operator Characteristics-Area Under the Curve
SAS	Statistical Analysis Software
SD	Standard Deviation
SGD	Stochastic Gradient Descent
SMOTE	Synthetic Minority Over-sampling Technique
STS	Society of Thoracic Surgery
TEE	Trans-esophageal Echocardiography

TIA	Transient Ischemic Attacks
TN	True Negative
TNR	True Negative Rate
TP	True Positive
TPR	True Positive Rate
UK	United Kingdom
USA	United States of America
WEKA	Waikato Environment for Knowledge Acquisition

ACKNOWLEDGMENTS

I would never have been able to finish my dissertation without the guidance of my committee members, help from friends, and support from my wife and family.

First and foremost, I would like to sincerely thank my supervisor, Dr. Syed Sibte Raza Abidi, for his excellent guidance, care, help, support, and his patience in correcting my writing and providing me with the opportunity to pursue my research endeavors.

I would like to express my deepest appreciation to my other supervisor, Dr. Gregory Hirsch, who allowed me to indulge in the big data science research, fought for me, facilitated obtaining the data, supported my research, and believed in me. Without his guidance and support, this research would not exist.

In addition, I would also like to thank Dr. Samina Abidi, for her guidance and support to my research for the past year and helping me to develop a solid background in health systems and research methods.

A special thanks goes to Dr. Samuel Alan Stewart, who taught me the principles of statistics and was willing to give me his guidance and support throughout this work.

Special thank Ms. Karen Buth who was always willing to help and give her best suggestions when it came to dealing with the data.

A special thank you to my sponsor; King Saud bin Abdulaziz University for Health Sciences, National Guard Health Affairs of Saudi Arabia, and the Saudi Arabian Cultural Bureau in Canada for believing in the importance of research and its huge significance in building a better future for humanity.

My love and gratitude to my parents and sisters, I could not have completed this without your support. I love you all.

Finally, I would like to thank my wife, Lama Khoshaim, for always being here, cheering me up, and standing by me through good and bad times.

I doubt that I will ever be able to convey my appreciation fully to any of you.

CHAPTER 1: INTRODUCTION

1.1. Predicting Post-operative Complications

A 16th-century French surgeon, Ambroise Paré, stated that to perform surgery is: "to eliminate that which is superfluous, restore that which has been dislocated, separate that which has been united, join that which has been divided and repair the defects of nature." Surgical techniques have been developed since the ancient times to treat injuries and traumas, and to preserve limb and life. The oldest operation for which evidence exists is trepanation [1]. Trepanation, making a burr hole, is a surgical procedure in which a hole is drilled into the human skull, exposing the dura mater with the intention of treating health problems related to intracranial diseases. The term may also refer to any hole created through other body surfaces with the intention to relief pressure. Although surgeons were considered experts with vast knowledge in life sciences and delicate hands, making them able to relieve suffering and end misery; surgery was always feared because of its association with pain and the potentially horrific experience. Surgeons therefore had to be swift and perform hasty procedures. Until the discovery of anesthetic agents, surgeons were largely restricted to amputations and external growth removals.

It was not until the middle of the 19th century, when the American surgeon, Crawford Long, demonstrated the use of "ether" as a general anesthetic agent, which enabled surgeons to perform surgery while minimizing the patient pain and awareness of the procedure[2]. In enhancing the surgeon's ability to perform more extensive operations without making the patient suffer, this discovery revolutionized surgery and, therefore, medical care. With advances in the practice of surgery, such as sterility, anesthesia, and post-operative care, the likelihood of dying because of surgery declined. Therefore, the concerns of patients shifted somewhat, from focusing entirely on the possibility of death to thinking more about their recovery: when they would be able to return to work and other activities, their level of independence, and their overall quality of life[3-6].

Because neurocognitive complications after any medical intervention can cause a big burden on the patient, medical team and society, every effort should be made to prevent them. Post-operative neurocognitive complications occur in up to 60% of patients[7]; these include stroke (permanent or transient), seizures, and delirium. Stroke, defined as the rapid loss of brain function resulting from any disturbance in the blood supply to the brain, is considered to be a major complication after surgery – especially after cardiac surgery, where it occurs in up to 6% of patients. Delirium or acute confusion is a temporary mental disorder that occurs frequently among hospitalized patients[8]. The incidence of delirium varies from 1-52% among reports[7], reflecting an inconsistency that may be attributed to the lack both of a standard definition and of good detection tools. Delirium is certainly an adverse neurocognitive outcome after cardiac surgery; especially in vulnerable patients and it influences additional complications[9, 10].

Delirium symptoms range from a disturbance in consciousness (e.g., coma, disorders related to concentration, and attention) to cognitive disorders involving disorientation and hallucinations. There is also a motor component, and presentation ranges from a depression-like inactive state, to an agitated hyperactive state. This diversity of possible presentations, along with its sudden onset and unpredictable course, makes early detection difficult. According to a paper by Royston and Cox, from the patient's point of view, delirium, and subsequent cognitive decline is among the most feared adverse events following surgery[3].

Patients undergoing cardiac surgery are considered to be at higher risk of developing delirium because of several factors. In several studies, delirium – especially after cardiac surgery – was linked to increased morbidity and mortality[10-13]. The effects of delirium can extend beyond the initial hospitalization; a retrospective review by Martin et.al discovered that patients with delirium following coronary arteries bypass graft surgery (CABG) exhibit an increased long-term risk of death and stroke[14]. In another study, time to discharge was 11 days longer for patients who developed delirium after elective cardiac surgery[8]. Bakker et. al found that patients who developed post-operative delirium had lower preoperative Mini-Mental State Examination (MMSE) score, higher creatinine level, longer extra corporeal circulation (ECC) time, and a significantly higher mortality at 30 days from surgery[15]. Type of surgery, symptomatic cerebrovascular

disease, advanced age, and diabetes mellitus are some of the proposed pre-operative factors that contribute to post-operative neurocognitive complications, especially delirium[16]. Delirium following cardiac surgery has also been associated with a higher European System for cardiac operative risk evaluation score (EURO-Score) [8, 17, 18]. One study identified delirium as a predictor of post-CABG sepsis[9]. Others have demonstrated a strong association between delirium and post-operative infections in cardiac surgery patients[14, 18-20].

One of the most significant predisposing factors for delirium is the patient age at surgery[8, 12, 15, 19, 21-24]. Some authors have also linked delirium to frailty[25-29]. Frailty is a common geriatric syndrome that is associated with steep declines in health and functioning among older adults[30-32]. Recent studies indicate a substantial, and alarming, increase in the number of elderly frail patients undergoing cardiac surgery[33, 34]. The dramatic increase in the numbers of elderly patients who are undergoing cardiac interventions can be attributed to the advances in pre-, intra-, and post-operative techniques. Almost 45% of patients over the age of 60 years undergoing cardiac procedures develop delirium[18]. Several studies have shown that preventive interventions can decrease the incidence of delirium and improve outcomes[24, 35-37]. Therefore, prevention or early recognition of delirium is essential.

Fortunately, there is no lack of data in health care. Regrettably, the medical community has recognized the “data rich, information poor” (DRIP) syndrome since the early 1990 s[38]. The DRIP syndrome refers to the abundance of data, but the data does not inform practice and decisions because it is not presented in the right context with relevant comparisons [38]. Predictive analytics deal with the abstraction of information from data and using this information to uncover obscured trends and hidden patterns, which can be used to mitigate risks and exploit opportunities for the future. Predictive analytics can enhance health care performance by reducing readmissions and providing preventive care. Providing preventive care by identifying patients with existing conditions that may lead to undesirable consequences, and acting to avert such outcomes.

There is a growing interest in health care frameworks that focuses on prevention, health promotion, and the use of novel technologies to allow health care professionals to take a

more active role in reducing the burden on the system by targeting treatments to those who will most likely gain the maximum benefit and least harm from them[39-41]. Providing health care professionals with the right information at the right time, will improve the quality of care[42-44]. This information can also be used to enhance patient education, compliance, and involvement through automated prompts for specific tasks[45-47]. Predictive models that are embedded in clinical decision support systems will ensure the delivery of appropriate and comprehensive care based on up-to-date practice guidelines and personalized plans[48-50].

1.2. Research Objectives

Several pre-, intra-, and post-operative factors have been linked to the development of post-operative delirium (e.g., frailty, type of surgery, gender, post-operative transfusion, post-operative renal failure)[7-11, 15, 17, 20-23, 25, 26, 51]. Further, patients that experience post-operative delirium are at increased risk of developing infectious complications, especially wound infections, and pneumonia[14, 18, 19, 52]. The negative consequences of delirium on post-operative outcomes, especially after cardiac surgery, are well documented[7, 10, 12-14, 21, 25]; but anticipating its development is not well documented in the literature[8, 11, 15, 17]. So far, models that predict delirium in adult cardiac surgery patients published in the medical literature have mainly relied on conventional statistical approaches, primarily logistic regression (LR). LR produces a linear combination of the attributes with weights that illustrates the attribute's statistical significance[53, 54]. Creating a simplified representation of how a subset of attribute influences the result. However, accurate models are complex[55].

A concise model is more likely to miss valuable information about ambiguous, yet important, relationships. A predictive model will inform future decisions by clarifying the process that leads to a specific result or an outcome. Clarifying the process will highlight key attributes that influence the final result, identify areas of potential obstacles and initiate appropriate interventions to improve future results. The final model will aid health care professionals in implementing customized treatment plans by identifying patients

who are prone to adverse events, which will trigger appropriate early preventive interventions.

The ultimate goal is to develop a model that will improve early detection; and a complex problem, like delirium, will require a complex solution to achieve better results. The proposed solution was to develop a predictive model that is capable of distinguishing vulnerable cardiac surgery patients that are prone to post-operative delirium. To develop this solution, two main objectives were formulated (Table 1-1):

1. **Identify key attributes/features that contribute to the development of postoperative delirium after cardiac surgery.** Two main methods were utilized. First, a conventional manual selection based on statistical significance and domain knowledge. Then, machine learning feature selection methods were used to identify interesting attributes that were not picked up by the conventional approach.
2. **Develop several predictive models that are capable of identifying patients who are prone to post operative delirium and compare their performance**. The primary goal was to verify if using a complex data-mining model would improve the predictive power when compared to the traditional statistical approach. Hence, it correctly identifies patients at risk so preventive measures can be initiated and negative consequences can be mitigated.

Table 1-1 Thesis research objectives, methods and expected outcomes

Objective	Methods	Expected Outcome
Identify key attributes/features	• Conventional approach (Statistical significance and domain knowledge) • Machine learning feature selection methods	• Identify key attributes • Discover hidden attributes that can be identified by feature selection methods
Develop predictive models	• Traditional statistical method: Logistic regression (Reference model) • Data-mining methods: Artificial Neural Networks and Bayesian Belief Networks	• Evaluate the predictive performance • Discern if complex data-mining methods will outperform traditional methods
Mitigate the impact of outcome class imbalance on performance and model stability	• Several imbalance class manipulation techniques are explored	• Manipulating the class imbalance might improve classifier performance • Some models might be more stable than others

Several modifiable pre-operative features can be tackled in an effort to optimize the patient medical and physical condition prior to surgery (e.g.: frailty and pre-operative rehabilitation program, diabetic control, nutritional education, pre-operative hemoglobin level, smoking cessation…etc.). This information can also be used to alarm the care team; so more attentive monitoring will be enforced (e.g.: confusion assessment method) with appropriate resources devoted for these patients. In this thesis, I will demonstrate that adapting data mining approaches to flag cardiac surgery patients that are at higher risk of developing post-operative delirium will yield better predictions when compared to traditional methods.

1.3. Research Tasks

Predictive models utilize all available information, past and present, to forecast the future to avoid threats, boost revenue, and help attain positive results. The methodology of this thesis is based on a data-mining pipeline, which involves the analysis of a large data set to discover new and useful patterns that might help define a particular problem and suggest novel solutions. Here, the problem we wanted to address was post-operative delirium, a dichotomous (binary) outcome of (yes/no) in cardiac surgery patients. To accomplish our objectives, we went through the following steps:

1. **Detecting key features:** The first step in detecting useful patterns is identifying key features. Several approaches will be used, including: conventional statistical analysis (univariate and multivariate analysis), cluster analysis, and machine learning feature selection methods. Feature selection is a task of choosing the features that are necessary and sufficient to delineate the target concept.
2. **Build predictive models:** In this work we used 5 different models: LR, artificial neural networks with 1 hidden layer (ANN-1 Hidden), artificial neural networks with 2 hidden layers (ANN-2 Hidden), Bayesian belief networks with 1 parent (BBN-1 Parent), and Bayesian belief networks with 2 parents (BBN-2 Parents). These models were built using the training set. LR is a well-known statistical data model that is extensively used in medicine. ANN is distinguished machine learning algorithmic approach that is inspired by the human brain architecture of connected neurons. BBN provides a graphical model of causal relationships that is supplemented by the probabilistic power of Bayesian statistics.
3. **Models performance evaluation:** The LR, ANN-1 Hidden, ANN-2 Hidden, BBN-1 parent, and BBN-2 parents models performance were evaluated on an test set that was not seen before by the models. There is a significant imbalance of the outcome distribution (delirium: 11.4% positive cases) in the final dataset. The use of predictive accuracy as an evaluation measure of the models will lead to false conclusions; the receiver operator characteristics-area under the curve (ROC-AUC) is a more appropriate measure and it was used to evaluate general

performance[56-59]. Other measures were used to compare the core performance of each model and its ability to recognize positive cases (e.g.: Kappa statistics, F1-score, precision, recall, and specificity) with the primary focus on the F1-score[29, 57, 60-68].

4. **The effect of outcome class imbalance on model performance, and proposed solutions:** Like most real life scenarios, we are attentive to the fact that the dataset is describing an infrequent but important event, delirium. It has been reported that one of causes of a poor classifier performance is related to class imbalance in which observations in training data from one class outnumber the other class. This encouraged us to explore the effect of applying several data mining techniques that overcome the class imbalance effects on the models performance (LR, ANN-1 hidden, ANN-2 hidden, BBN-1 parent, and BBN-2 parents) and examine the their stability (e.g.: data level resampling and applying cost)[56-59, 65, 69, 70].

1.4. Thesis Organization

The remainder of the thesis is organized as follows: Chapter 2 presents the background, research motivation, and core concepts related to the research question. Chapter 3 outlines the data source, the preparation process, feature selection, and the description of the final dataset. Chapters 4 describes the evaluation measures of a classification task in the presence of imbalanced class, how to deal with imbalance, the applied classification methods that were used to develop a predictive model for delirium, and the results of the performed experiments. Finally, Chapter 5 includes a discussion of the major findings and limitations of this research, potential directions for future research, and concluding remarks.

CHAPTER 2: BACKGROUND

2.1.Neurocognitive Disorders

The moment that a patient requires a medical intervention, whether it's something simple, like an x-ray or a prescription, or something more complex, like a course of chemotherapy or a surgical procedure, he or she is at risk for the complications and side effects associated with that intervention.

While death because of surgery is a genuine concern among patients, the failure to regain normal brain function is an equally important – and somewhat more realistic – worry[3]. People are more familiar with this worrisome possibility for several reasons, including media attention and greater access to mass information (varying in quality and validity). The revolution of pre-, intra-, and post-operative care has also contributed to these fears in an indirect manner, making deaths related to anesthesia and surgery much less of a concern and thus shifting the focus of patients to their post-intervention quality of life. Royston and Cox stated, "From the patients point of view, delirium, and subsequent cognitive decline is among the most feared adverse events following surgery"[3]. In the mid-1980s, the cardiac surgical community started to notice that some of the patients who are undergoing open heart surgery demonstrated a deficit in cognitive function that extended up to the 8 weeks after surgery[71].

Post-operative deficits in brain function represent a type of acquired brain injury, which includes any type of brain damage or neurological disruption occurring after birth. The *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5)* has recently switched to the term 'neurocognitive disorder,' which is on the same spectrum of impairments with causes that include trauma, vascular disease, Alzheimer disease, and infection. The term neurocognitive disorder provides a diagnosis for people experiencing cognitive symptoms alone, without memory or physical impairments[21]. This means that many individuals who are not currently receiving recognition or services (due to the lack

of memory or physical impairments) will now qualify. Due to the complexity and localization of the various functions overseen by the brain, organic damage can cause a wide range of effects that are not limited to cognitive deficits[21]. Sight, hearing, and movement can all be affected by damage to specific areas of the brain. Therefore, people with neurocognitive disorders have a higher proportion of co-morbid disabilities compared to the general population[21].

Neurocognitive disorders are linked to mental health disorders. A neurocognitive disorder is defined by a shift in abilities and functioning, potentially leading to mental health issues. The Australian Institute of Health and Welfare (AIHW) states that over 40% of people with an acute brain injury have a co-morbid mental health issue[72]. In the same report, 96% of people with acute brain injury aged 65 years or over suffered a physical disability. In comparison, those less than 65 years were more likely to suffer more psychiatric and intellectual disability[21].

There are several major consequences of post-operative neurocognitive complications, the most direct of which is prevention of patient recovery, delaying hospital discharge and, therefore, a return to normal life activities. From a policy point of view, this resumption of daily life is a major concern. A prime example of this concern is a patient's ability to resume operating their automobile. If a patient's career is dependent on their ability to commute or travel from one location to another and they suffer a neurocognitive complication that will impair their ability to drive, this will have a major effect on the patient's life.

For elderly patients, the possibility of losing their independence because of a surgical complication is a larger concern than mortality. In a study assessing the perception of death and health status in elderly patients (mean age of 70 years) with heart failure in Sweden, Strömberg and Jaarsma found that many patients were concerned about suffering and losing their independence, and the possibility of not receiving good care[73].

With life expectancy relatively unaffected, a neurocognitive complication has the potential to adversely affect individuals during the remainder of their lives. The incidence of neurocognitive complications rises as people age, which will increase the incidence of fall, dependence, and the burden on the health care system[13, 26, 33, 74, 75]. Careful

detection of and early intervention for neurocognitive complications in this group of patients is necessary in order to mitigate lasting and secondary effects.

2.2. Delirium

2.2.1. Definition, Etiology, and General Information

Delirium is a disturbance of consciousness, attention, cognition, and perception. The disturbance develops over a short period of time (usually from hours to days) and tends to fluctuate during the course of the day[25]. According to DSM-5, delirium represents a sudden and significant decline from a previous level of functioning that cannot be better accounted for by a preexisting or evolving dementia. There is usually evidence from the patient history, physical examination, or laboratory tests that the delirium is a direct physiological consequence of a general medical condition, substance intoxication or withdrawal, use of a medication, toxin exposure, or a combination of these factors[21].

The medical community has always been aware of delirium's wide range of presentations, from extremely dangerous agitation to depression-like isolation. Nonetheless, the latest updates of the DSM-5 and Geriatric Psychiatry, Fifth Edition were the first place to formally establish 3 distinct subclasses based on presentation: hyperactive, hypoactive, and mixed[21]. A new entity, attenuated delirium syndrome, was also added as a diagnosis for the presence of some but not all of the diagnostic criteria for delirium[21].

The occurrence of delirium is linked to many types of factors, including: systemic illness (e.g., infection, electrolytes imbalance, hypoxia, renal dysfunction, liver dysfunction, heart failure, neurological pathology, etc), medications (e.g., analgesics, steroids, sedatives, antidepressants, anti-Parkinsonism and others), and numerous other risk factors (age above 60, male gender, major surgery, dehydration, substance abuse, functional independence, depression, admission to the intensive care unit (ICU), anemia, sleep deprivation, anxiety, uncontrolled pain, and others)[21, 25, 76]. Reade and Finfer recently delineated an ICU triad of pain, agitation, and delirium. The notion of a triad emphasizes the complexity of delirium and other related problems, which highlights how difficult it is

to find a single intervention that can be used as a preventive measure without negative consequences[76] (Figure 2-1[76]).

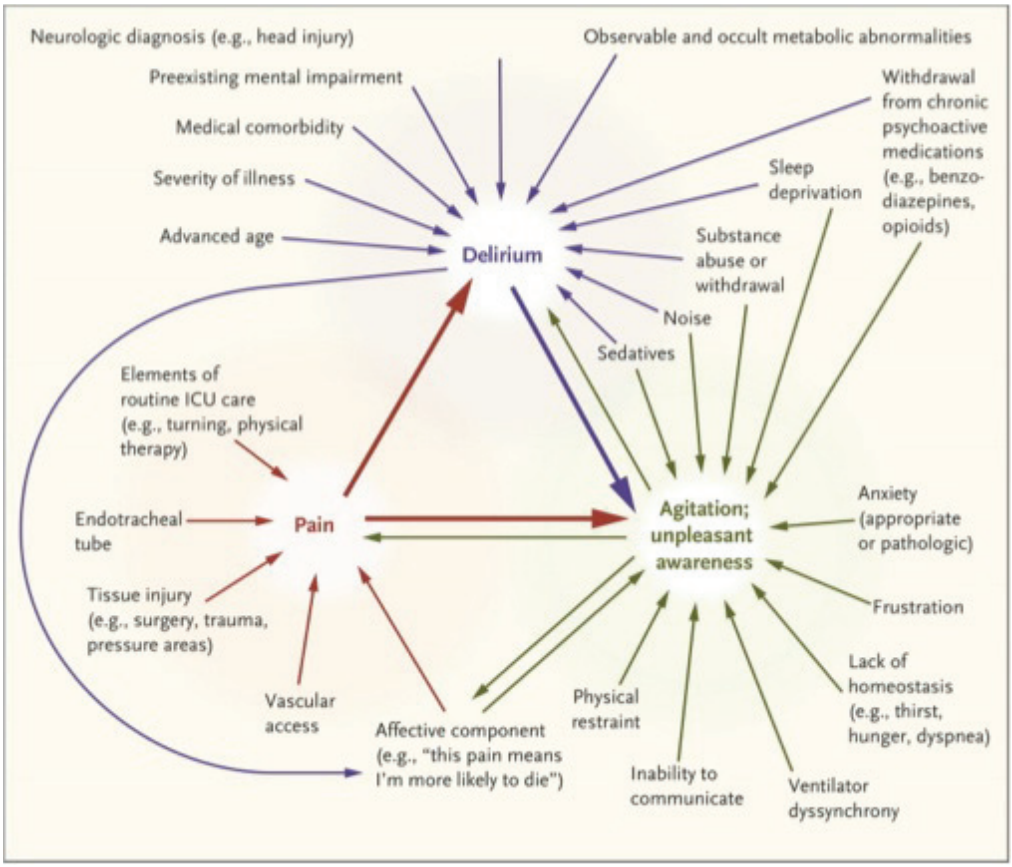

Figure 2-1: Causes and Interactions of Pain, Agitation, and Delirium.
From: Reade MC, Finfer S. N Engl J Med 2014; 370:444-454. Publication Title: Sedation and Delirium in the Intensive Care Unit. Copyright © (2014) Massachusetts Medical Society. Reprinted with permission from Massachusetts Medical Society.

Delirium is most prevalent in the hospitalized elderly, and its diagnosis varies based on the patient's medical status, the type of care, and the detection tools used. The general prevalence of delirium across the entire community is low (l-2%) but increases with age, rising to 14% among individuals older than 85 years. The prevalence is 10-30% in elderly (age >65 years) who present to emergency departments, in which case the delirium is often secondary to another medical problem. The incidence of delirium during hospitalization is between 6 and 56%. Delirium occurs in 15-53% of older individuals

post-operatively and in 70-87% of those in intensive care. It occurs in up to 60% of individuals in nursing homes and in up to 83% of all individuals at the end of life[21].

Before diagnosing a patient with delirium, the health care team needs to rule out reversible causes. These include hypoxia, hypoglycemia, complex partial seizure, encephalopathy (viral, bacterial, metabolic, or hypertensive), renal failure, heart failure, infection, subdural hematoma, electrolyte disturbance, and as a side effect of several medications. The team needs to consider the possibility that delirium may be a secondary manifestation of an ongoing pathological process. If the underlying insults are discovered early and acted on in a timely fashion, delirium can be reversed.

The diagnosis of delirium is mainly based on clinical suspicion. The National Institute for Health and Clinical Excellence in the United Kingdom (UK) produced a comprehensive 447 page set of guidelines, entitled "DELIRIUM: diagnosis, prevention and management," in an effort to educate the medical community and standardize the process of managing delirium"[24]. They stated, "Delirium is common but is frequently unrecognized by doctors and nurses despite the fact that it can be life-threatening and lead to serious preventable complications." In this document, they identified several diagnostic tools. In the general floor and long-term facilities setting, these tools include the Abbreviated Mental test (AMT), the clock-drawing test, the MMSE, the Confusion Assessment Method (CAM) and the Delirium Index (DI). The AMT, a 10-item questionnaire that was first used by Ni Chonchubhair in 1995 to diagnose delirium in the elderly, is administered the day before surgery and then again on the third day after surgery[77].

Formally described in 1990 by Inouye et.al[78], the CAM and has undergone several modifications, exists in both a long and a short version. The short version evaluates the criteria: acute onset, fluctuating course, inattention, and disorganized thinking or altered level of consciousness. The long version has 6 additional criteria (disorientation, impaired memory, perceptual disturbance, psychomotor agitation, psychomotor retardation, and altered sleep cycle).

In the ICU setting, the available tools include the confusion assessment method in the intensive care unit (CAM-ICU) and the Richmond agitation sedation scale (RASS). The CAM-ICU, described in 2001 by Ely et al., assesses the following features: acute onset or

fluctuation course and inattention, with either disorganized thinking or altered level of consciousness[79]. The inattention aspect of the assessment is based on Attention Screening Examination (ASE) scores.

In the UK guideline, the CAM short version has the highest specificity and positive predictive value (PPV) when administered by a general physician, psychologist, geriatrician, or resident during their geriatrics rotation or a psychiatrist in a non-ICU setting. Likewise, when the CAM-ICU is administered by an ICU nurse, it has a high specificity and PPV in all studies[24].

The guidelines have established that the diagnosis of delirium should be a 2-stage process. The first stage is intended to alert the primary health care team to the possibility that a patient may be developing delirium. Following from that, the second stage involves a comprehensive clinical assessment by an appropriately trained health care professional.

The diagnosis of delirium has been associated with a variety of additional complications. Mortality and readmission appear are significantly linked to the occurrence of delirium, in a time-independent manner. There is also strong evidence that delirium decreases the likelihood of discharge, resulting in a longer stay in the hospital; this effect is especially strong when the delirium has developed in the ICU[24].

2.2.2. Delirium and Cardiac Surgery

Patients who are undergoing cardiac surgery are considered to be at higher risk of developing delirium due to several factors, including: surgical complexity, presence of other co-morbidities, and age. Several studies have indicated that there has been a dramatic shift of the demographics of the cardiac surgical population, with fewer smokers, more diabetics, and older patients[33, 34, 80]. In one study by Buth et al. that looked at the demographics and characteristics of a cohort of cardiac surgery patients from 2001-2010, frailty increased dramatically over time[33]. In another study by Pierri et.al that examined a cohort from 1999-2007, they found that over the 9-year study period, patients were getting older, and they were operating on sicker patients (shock, higher New York Heart Association (NYHA) classification, and unstable angina), with more co-morbidities (hypertension and morbid obesity), and a lower ejection fraction (<30%)[34].

Older age, critical illness, and undergoing cardiac surgery are some of the well-established risk factors for developing post-operative delirium[21]. Some of the factors that are particular to cardiac surgery include: pre-operative EURO-Score[8], length of stay in the ICU[19, 23], prolonged mechanical ventilation [19, 23], prolonged aortic cross clamp time[20], undergoing valve surgery[22], history of cerebrovascular disease[18], left ventricular dysfunction[18], and diabetes mellitus[18]. The biggest issue with these risk factors is that most of them are based on an observational small cohort of patients (44-142 patients). One literature review on medication that can cause delirium after cardiac surgery concluded that: intraoperative fentanyl, intraoperative ketamine, preoperative antipsychotics, and post-operative inotropes were associated with post cardiac surgery delirium[29].

In several studies, delirium after cardiac surgery was associated with early, intermediate, and late morbidity and mortality. One study identified delirium as a predictor of post-CABG sepsis[9]. In this study, delirium was the second most important predictor of post-operative sepsis after emergent operation, with an odds ratio (OR)=2.32 and a 95% confidence interval (95% CI)=1.59-3.39[9]. A key conclusion of this paper is that delirium is a malignant process, and not the benign and self-limiting process that it is often regarded as in the medical community. This group of researchers has also demonstrated that patients who develop delirium have a median post-operative hospital stay of 12 days, compared to 6 days for those who did not develop delirium. Delirium was identified as an independent predictor of all-cause mortality, with a Hazard Ratio (HR)=1.52 and a 95%CI, 1.29-1.78; and hospitalization for stroke HR=1.54, 95%CI=1.10 - 2.17[14]. Another study of 5,034 consecutive patients undergoing CABG surgery at a single institution from 1997-2007 identified delirium after cardiac surgery as a strong independent predictor of mortality up to 10 years post-operatively (adjusted HR=1.65, 95% CI=1.38–1.97), even when data was adjusted for perioperative risk factors. This was more prominent in patients younger than 65 years (HR=2.42) and in those without prior stroke (HR=1.83)[10].

In recognition of the importance of delirium within the cardiac surgical population, some have attempted to develop a predictive model. For example, Afonso et al. conducted a prospective observational study on 112 consecutive adult cardiac surgical patients.

Patients were evaluated twice daily for delirium using RASS and CAM-ICU, and the overall incidence of delirium was 34%. Increased age (OR=2.5, 95% CI=1.6-3.9, for every 10 years) and surgical procedure duration (OR=1.3, 95% CI=1.1-1.5, for every 30 minutes) were found to be independently associated with post-operative delirium[17]. Similarly, Bakker et al. prospectively enrolled 201 cardiac surgery patients aged 70 and above. They found that a low MMSE score (27 vs. 28, p-value=0.026), a higher pre-operative creatinine level (98 vs. 88 μmol/l, p-value=0.003) and an increased ECC time (145 vs. 113 min, p-value < 0.001) were independent predictors of post-operative delirium[15].

Unfortunately, all of the previously published models focus on selecting a subset of well-known attributes. These attributes have already been demonstrated to influence the development of post-operative delirium. Also, all of these models used the same technique, LR.

Logistic regression takes a binary outcome expressed as a probability that must fall between 0 and 1 and transform a non-linear outcome to a linear probability using the natural logarithmic scale. This generally includes only a subset of attributes that are considered important based on their statistical contribution. Selecting a subset limits the number of allowed attributes and possible interactions in the model. However, there are situations where the use of these models is unfitting, as it will require many assumptions that in reality may not be true. Violating these assumptions may produce an error in prediction and hypothesis testing. In addition, logistic models use a linear combination of variables and, therefore, are not suitable for modeling multifaceted relationships. Traditional methods suffer from their inability to capture pattern complexity and uncover hidden process dynamics. To overcome the limitations of traditional methods, machine learning techniques were developed. Machine learning techniques are based on the rigorous mathematical principles of traditional methods but augment their performance by allowing complex interactions and discovering the process dynamics as they unfold in reality.

2.3. Machine Learning and Medicine

2.3.1. Machine Learning, Knowledge Discovery, and Data Mining

For as long as humans have existed, people have tried to analyze information in the hopes of finding patterns and making accurate predictions. A classic example is weather forecast: in earlier times, knowing what kind of weather the coming days and weeks would bring was crucial because it affected whether or not people could plant, harvest, travel, and survive. Then, people had only limited information upon which to predict the weather – phenomena such as the behavior of animals and the positions of the sun, moon, and the stars. Today; however, meteorologists use complex equations and state-of-the-art computer systems linked to networks of satellites and weather stations around the globe, allowing for a much better understanding of weather and, therefore, more accurate predictions. These sensors gather massive amounts of information, which needs to be captured, stored, normalized, and analyzed in a very short time span. However, even with the modern technology of today, there are so many attributes that forecasting the weather is still not perfectly accurate.

Russell Ackoff, who in 1989 described a hierarchy of knowledge, ranging from data to wisdom (Figure 2-3 [81]). In Ackoff's hierarchy, data – mere symbols with no inherent significance beyond their existence – is transferred via analysis into information, which does have applicable meaning. At the next level, knowledge results from the addition of context, such that the information can be better understood. Next, insight into how things work results from the synthesis of new knowledge and information. Lastly, wisdom reflects the broad understanding of a phenomenon, including the acknowledgement that much is and will remain unknown.

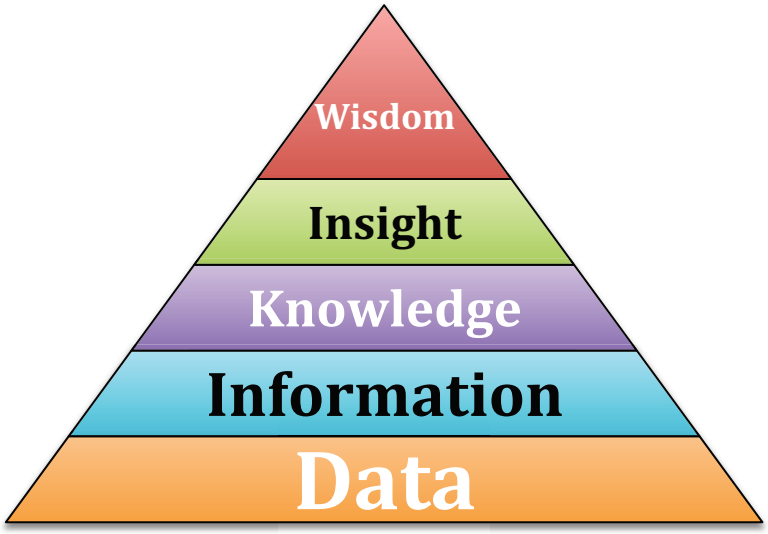

Figure 2-2: Data to Wisdom Pyramid
(Based on the hierarchy of profound knowledge from Russell Ackoff)

Starting in the early-1980s, the amount of data available for analysis began to exceed current capabilities to efficiently analyze and extract meaningful information from that data. The amount of generated data has continued to increase exponentially, primarily because of the computerization of our society and the fast development of powerful collection, storage, and analytical tools. Techniques such as data mining (DM) rose to prominence as a way of dealing with several aspects of this data explosion. One issue was how to efficiently store and collect such large amounts of data. Computers becoming cheaper and more powerful, ultimately giving most consumers access to these masses of information has addressed this, partly. The extraction of meaningful information from this data; however, requires costly physical resources, including teams of human analysts and many man-hours. In general, without proper tools, some of any data set's meaning will go undiscovered, with much of the data never analyzed at all.

Data mining is the interdisciplinary process of discovering interesting patterns and knowledge from large amounts of data, or big data, and integrates techniques from several disciples, including statistics, machine learning, pattern recognition, data visualization, and databases[64]. The concept of 'big data' does not have an exact size – it is more of a moving target that grows in parallel with the size of data sets, often defined

in terms of the 3 Vs: volume, variety, and velocity. Essentially, big data refers to any data sets with sizes that exceed the abilities of commonly used software tools to capture, curate, manage, and process the data within a reasonable amount of time[82].

To follow the application of data mining described in this thesis; it is important to understand some related concepts. Statistics is defined as the science that deals with the mathematics of the collection, organization, and interpretation of numerical data, especially the analysis of population characteristics by inference from sampling. Machine learning is the study of providing computers with the ability to learn without being explicitly programmed. Database science is the science and technology of collecting, storing, and managing data so that users can retrieve, add, update, or remove that data. The major distinction between DM and a database query is that in a query you simply ask "How many" whereas in DM you take this question a step further and ask, "What are the important/associated features that resulted in that pattern." Predictive analytics is an approach in which data are analyzed for meaningful patterns that can provide actionable insights, which can in turn be used to enhance preventive actions and provide cost-effective management that improves the domain products (examples in health care include patient outcomes, patient satisfaction, resources allocation, system re-structuring, and others)[40]. Sometimes used as a synonym for DM, knowledge discovery from databases (KDD) is the process of analyzing data from different perspectives and summarizing it into useful information through the extraction of interesting (non-trivial, implicit, previously unknown, and potentially useful) patterns or knowledge from huge amounts of data. DM can be viewed as a temporal phase in the development of machine learning and statistical learning, as a new application area has emerged, and ad hoc techniques are gradually being usurped by established techniques that have long pedigrees in learning theory or statistics.

Some experts consider DM to be one step in the process of knowledge discovery (Figure 2-3), a process that starts with the posing of a question based on a problem and the existing knowledge regarding that problem. Identification of the problem and the question is followed by determination of the appropriate data for answering the question and collection of the data form available sources. Target data is then selected, pre-processed and subjected to data mining to produce a pattern or a model that can be analyzed. Such

analysis will generate some insight and knowledge that can be used to facilitate and inform future decisions.

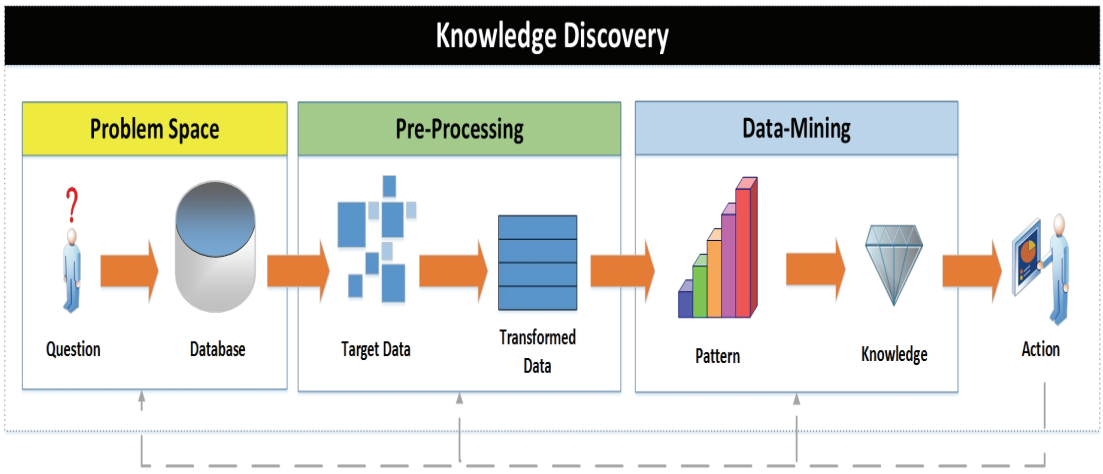

Figure 2-3: From Data to Action (Knowledge Discovery Process)

The most troubling and critical aspect of the growing gap between the volume of data and our abilities to timely process and analyze it; is the possibility that patterns of significant meaning – that may, in the context of medicine, save lives – remain hidden owed to our limitations[83]. Nonetheless, in marketing, banking, insurance, aviation, automotive industries, and many other disciplines, the massive amounts of generated data are successfully analyzed in a matter of seconds (whether by fully or partly automated processes) in order uncover useful patterns. In banking, data mining outlier identification techniques are used for the detection of fraud transactions. Data mining clustering techniques can be used to identify different customer clusters that have similar shopping behaviors, in which targeted advertising can be applied on to improve future revenue and product sales.

The information acquired via the data mining process acts as the basis for proper actions, applicable adjustments, and intelligent decisions. These sectors have caught up with the explosion of data, apparently understanding that being able to acquire meaningful information from large data sets ultimately leads to customers coming back and waste of time and resources being reduced[83].

2.3.2. Machine Learning in Medicine

Medical science generates incredible amounts of data. However, one might argue that medical science fails to make the best possible use of this information (primarily in the form of study results in medical literature and data points in health records) to serve its goal, which is the maintenance of health. MEDLINE, the U.S. National Library of Medicine's (NLM) premier bibliographic database, is the most comprehensive medical literature database, with over 19 million references to journal articles in the life sciences. It includes citations from nearly 5,600 worldwide journals, and 2,000-4,000 new references are added each day[84]. In a study published in the Journal of the Medical Library Association (JMLA) in 2004, authors found that 7,287 articles are published each month within the subset of 341 active primary care journals. A well-trained physician would need to devote about 627 hours – or 26 days! – every month in order to critically appraise these articles, extract useful information, and integrate this information into his or her practice[85].

In reality, though, 81% of physicians report spending less than 5 hours per month reading medical journals. Clearly, there is no lack of data available. The critical task is to extract useful information and to detect and understand recurrent patterns in an efficient manner, promoting the delivery of evidence-based, data-supported, value-driven and patient-centered care.

There are numerous examples of authors using this non-traditional, data-driven approach to tackle medical problems, with several articles using such methodologies to improve early and precise detection of diseases (e.g., breast cancer, atrial fibrillation, etc.)[86, 87]. One such study employs a voice signal analysis for early detection of Parkinson disease[88]. Other studies use additional data-driven approaches to better understand specific grouping patterns (like genetic DNA clusters, disease groups, complications patterns, infectious disease epidemics, and others)[89]. Optimization of health care resources has also been tackled using such an approach (e.g.: outbreak detection and cost forecasting)[90-92]. The prevalence of medical research articles employing data mining procedures has increased dramatically in recent years. The first such medical publication appeared in 1978, according to PubMed[93]. The appearance of these types of papers

increased slowly in the years that followed and then more rapidly during the last 20 years, with 265 such articles per year published between 2010 and 2013 (Figure 2-4).

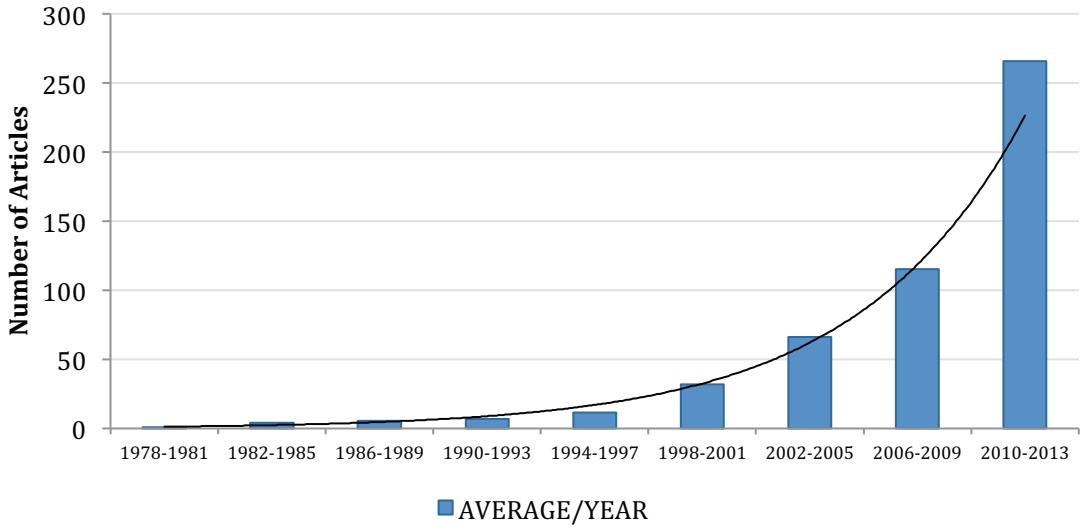

Figure 2-4: Medical Data Mining Publications on PubMed
(Search conducted at the 31-January-2014)[93]

In one example of a medical research study that applies data mining methods, Yan et al. used a clustering method called self-organizing maps of optic nerve images obtained via confocal scanning laser tomography (CSLT) in normal subjects and patients with glaucoma, with the goal of devising a new 5 categories sub-classification of glaucoma based on morphological features[94]. In another paper by Qiang et al., the researchers report a technique for recognizing vascular patterns associated with cervical lesions in colposcopic images. The lesion patterns are usually confusing and complex, such that even trained physicians can have difficulty recognizing typical patterns. Using only 12 out of 24 features, the authors were able to generate a model that can differentiate each pattern with 80% accuracy[95].

Mao et al. used a comparable approach to address an important problem: early detection of deteriorating health conditions (mainly pre-shock status) in patients who are in the general hospital wards (prior to transfer to intensive care). They applied a data mining framework to solve this problem, developing an early warning system designed to identify the signs of clinical deterioration and provide early warning of serious clinical

events. They developed this system using a very large amount of data – 41,503 patient visits. The proposed system achieved a high specificity (95%) and was able to detect deterioration in a patient's health condition at least 4 hours before transfer to ICU, and all the patients who were transferred to the ICU had an alert less than 24 hours before they were transferred[96]. In a paper by Pandey et al., they tested several clustering algorithms to assist in the prediction of heart disease[97]. Their findings indicate that, with the right algorithm, patients that are at higher risk of developing heart disease can be identified early, enabling more effective interventions.

2.4. Chapter Summary

Unlike some disciplines, medical science has mostly failed to keep up with the data explosion during recent decades, so that far more information is produced than professionals in the field could possibly assess and apply appropriately. In this era of open access and large data sets, physicians must harness these options in order to be up-to-date and acquire the best, most useful information. Data mining shows great potential for addressing this problem and enabling people working in the medical field to make the most of the massive amount of data available.

Although data mining can explain the past, the real power lies in its ability of predicting the future with great confidence. Predictive analytics will help the health care community uncover patterns and hidden relationships between data points previously buried or thought to be unrelated. That, in turn, will fill in gaps in knowledge; optimize the flow of care, and trigger a significant paradigm shift away from a volume-based health care system and towards a value-based health care organization.

This work will illustrate that the use of a knowledge discovery techniques to address post-operative delirium in cardiac surgery by potentially enhancing our understanding of the underlying process. It will not only provide further support for many already-recognized attributes, but it will reveal new and unexpected links between attributes that might influence the development of post-operative delirium. This research will also demonstrate the predictive power of DM techniques compared to conventional statistical methods, as

they are able to produce a more accurate representation and solution. The use of these solutions will improve the identification of high-risk patients, alerting the health care team and triggering appropriate interventions in a timely fashion. Preventing further deterioration, and negative effects.

CHAPTER 3: THE DATASET AND PREPROCESSING STEP

This chapter will discuss the dataset, and the data preparation steps (Figure 3-1: Data Pre-Processing methodology, page: 26). The dataset is a cohort of actual patients who underwent cardiac surgery in Halifax, Nova Scotia, Canada. The data is a portion of the Maritime Heart Center Cardiac Surgery Registry. The registry is from a comprehensive, ongoing detailed clinical database that captures all cardiac surgery patients in Atlantic Canada since June 1995. It has more than 20,000 patients. The acquired data represent a cohort of patients who underwent cardiac surgery in the center between 2006 and 2012. The original dataset contained 5,798 patients, and each patient had 220 attributes.

After acquiring the data, the first task was to construct attributes that are not explicitly captured in the registry, but their components are (e.g.: length of stay in days, creatinine clearance and etc.). Next, redundant attributes were removed (e.g.: date of admission, creatinine and etc.). Afterward, univariate and bivariate analysis were applied to identify key attributes that influence delirium.

Several preprocessing steps were applied to prepare the data for model construction, reduce the attribute vector space (262 attributes) down to a more practical size, and isolate key attributes. These steps include: dimension reduction using clustering, discretization, and feature selection. Ultimately, we were able to identify 22 pre-, intra-, and post-operative attributes that significantly influence the development of post-operative delirium in our cohort. The pre-processing steps are summarized in Figure 3-1.

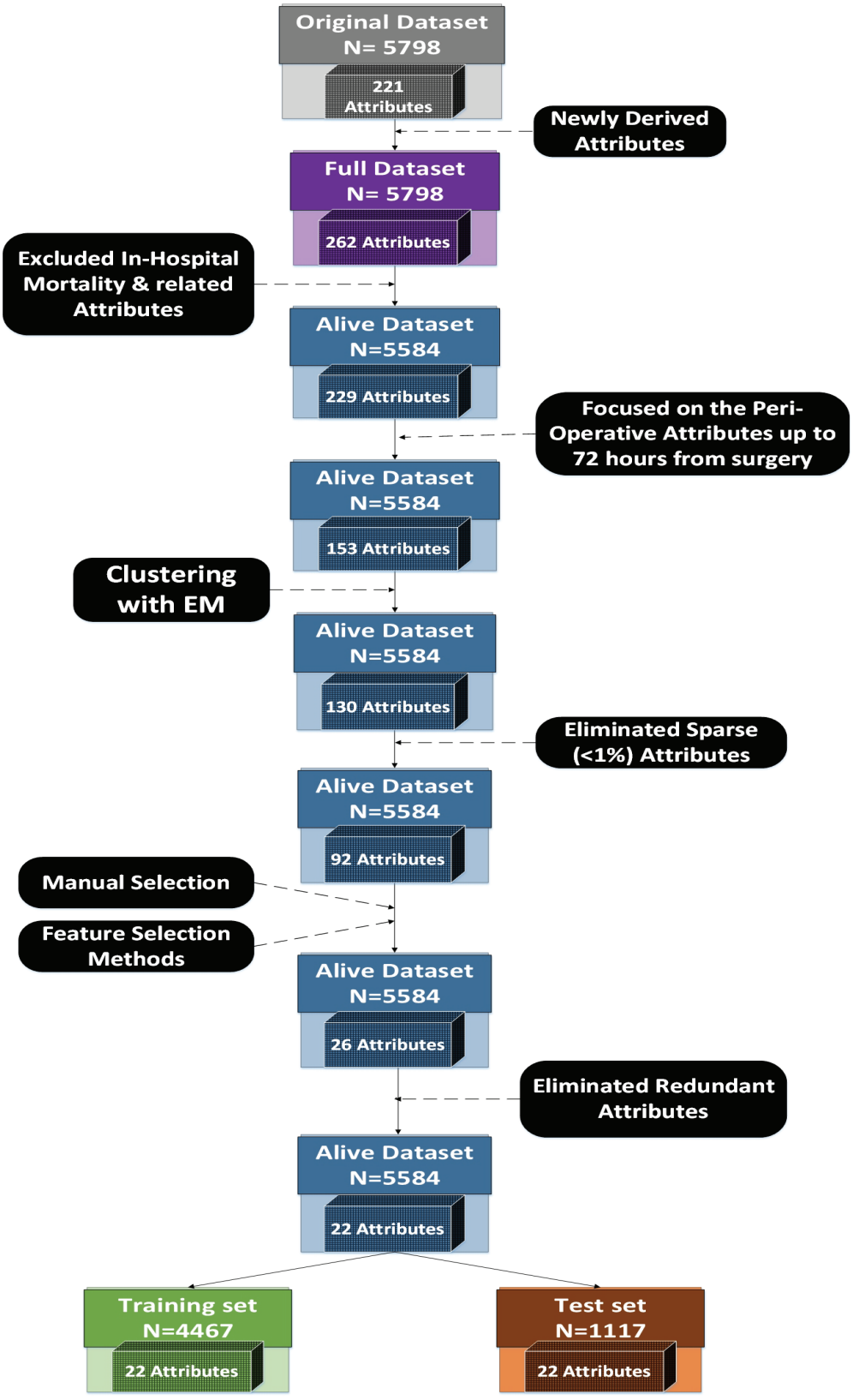

Figure 3-1: Data Pre-Processing methodology

3.1. The Dataset

3.1.1. Maritime Heart Center Cardiac Surgery Registry

The Maritime Heart Center Cardiac Surgery Registry is an ongoing detailed clinical database that collects pre-, intra-, and post-operative information on cardiac surgery patients. The database captures information related to all cardiac surgical procedures that take place in the Queen Elizabeth II Health Sciences Center (QEII-HSC), which is the sole cardiac surgical center in the province of Nova Scotia, Canada as well as for parts of the surrounding Atlantic provinces (New Brunswick, Newfoundland, and Prince Edward Island), constituting a region of 2 million people. The QEII-HSC is the tertiary referral center in Atlantic Canada that has the capability to perform complex cardiac procedures. An average of 1,000 open heart surgeries are being performed every year. It is also the only cardiac transplant and advanced heart failure center in Atlantic Canada. The registry started collecting patient information in June 1995, and has since undergone numerous iterations to improve it and make it more comprehensive. The latest version of the registry is based on the framework and data definitions used by The Society of Thoracic Surgeons (STS), and has more than 20,000 patients and more than 500 different variables. Many of the definitions are altered and customized to the fit the MHC registry primary goals (e.g.: Quality Improvement).

Following the STS definition, this data set defines delirium as mental disturbance marked by illness, confusion, and cerebral excitement, with a comparatively short course[98]. This definition will only include patients with agitated delirium, which represents the smaller portion of patients who develop post-operative delirium compared to the mixed and hypoactive types[27, 28, 52, 99]. In the MHC, Delirium is defined as short-lived mental disturbance marked by illusions, confusion, or cerebral excitement, requiring temporary medical/physical intervention, a consult, or extends the patient's hospital stay. Delirium is captured via manual chart abstraction, which is done by trained chart abstracters. The chart abstracters look for any subjective documentation of delirium by the medical team in the chart.

3.1.2. Research Ethics

Full ethics approval was obtained from the institutional research ethics board, in keeping with the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans[100]. Our request was reviewed by the Capital District Health District Health Authority (CDHA) Research Ethics Board (REB) and was approved on the 18th of September 2013 (File Number: CDHA-RS/2014-087). Under Section 2.1 c) of the Tri-Council Policy Statement[100], the ethics board granted a waiver of informed consent. An additional REB application to the MHC data access committee was submitted and approved at the 22nd of October 2013 File Number: MHC proposal 2013/032). Access to the full data set was granted at the 12th of November 2013.

3.1.3. Study Population

The study included all patients undergoing isolated CABG, valve, and CABG with valve cardiac procedures, including redoes (patients who had undergone a previous cardiac surgery), at the QEII-HSC in Halifax, Nova Scotia, between January 2006 and December 2012. We focused on that segment of the database because it was comprehensive and stable, although throughout the observation period the definition of delirium did not change. The valve procedures that we included are: aortic valve replacement (AVR) with mechanical or prosthetic valve, mitral valve replacement (MVR) with mechanical or prosthetic valve and mitral valve repair (MV-Rep), irrespective of the repair technique. Given the abovementioned criterion, our initial dataset included 5,978 patients.

3.1.4. Requested attributes

To determine the appropriate variables to include in our analysis, we conducted an extensive literature review of all articles that were published on delirium after surgery, with a focus on cardiac surgery. Based on this literature review, in combination with a review of the variables available in the MHC registry, 220 attributes were requested (all available in the MHC registry) and they were divided into 10 main categories: a) demographics, b) frailty, c) comorbidities, d) previous cardiac intervention, e) cardiac

specific, f) pre-operative, g) intra-operative, h) post-operative, i) discharge, and j) newly created. For more details on how did we create these attributes and the equations, please refer to APPENDIX A: ATTRIBUTES, page 120. After that, redundant attributes were removed (e.g.: date of admission, creatinine and etc.). Univariate and bivariate analysis were applied to isolate key attributes with significant influence on delirium. Delirium was coded as a binary outcome (present=yes and absent=no).

Several preprocessing steps were used in an attempt to reduce the attribute vector space (262 attributes). These steps include: dimension reduction using clustering, discretization, and feature selection. Non-normally distributed continuous attributes were discretized in an attempt to overcome. We were able to identify 22 pre-, intra-, and post-operative attributes that significantly influence the development of post-operative delirium in our cohort.

3.1.5. Clustering Correlated Attributes With Expectation Maximization

The expectation-maximization (EM) algorithm is a probabilistic algorithm that belongs to the partitioning clustering methods. It attempts to construct a "latent" attribute that can be used to maximize the likelihood estimate of the model[64, 115]. The expectation maximization (EM) algorithm was used to cluster correlated attributes in an attempt to reduce the attribute vector space, while maximizing retained information.

3.1.6. Original Data Format and Analytical Software

The dataset contained 5,798 patients/rows and 221 attributes/columns (requested attributes plus a study ID). All binary attributes were originally coded as 0/1 (0=no and 1=yes). Missing attributes were coded in different ways based on the type of the data; categorical were coded as "9," numeric were coded as "-9," dates were coded as "01/01/1900". Also, some attributes had "unknown" or "unk" to denote missing or un-coded values for character attributes with text.

In this study we used multiple platforms based on the software capabilities and the analyst's comfort with the available functionalities in each package. The following software was used in this thesis: Microsoft Excel 2013[101], Statistical Analysis Software

(SAS) V9.3[102], Waikato Environment for Knowledge Acquisition (WEKA) V3.7.10[103], and R V3.0.1[104].

3.1.7. Data Exploration and Statistical Analysis

After recoding, 221 attributes were examined (the first attribute was excluded, as it was a study ID but was kept to cross reference the rows). All date attributes were used to generate length of stay continuous attributes. Statistical measures of central tendency (mean, median, mode, inter-quartile ranges, standard deviation and others), shape of distribution and outliers were examined. Continuous attributes were examined and when the shape of the distribution was of close to a normal (Gaussian) distribution; mean and standard deviations were used. When the shape was extremely skewed (non-normal distribution); median, 25% inter-quartile range (IQR) and 75% IQR were used. Categorical attributes are reported with frequency in percent. Imputation of missing values was not used in an attempt to prevent the introduction of bias, through artificial speculation of a single result.

Patient characteristics of the full data set are displayed in Table 3-1. Delirium was documented only in 661 patients (11.4%). The mean age was 67 years. The majority of patients were male (74%). Ten percent of patients were older than 80 years. Only 13% had history of cerebrovascular disease (CVD). CABG was the most commonly performed procedure (67%). Out of the 3,886 patients that underwent CABG, 67 % had at least 3 distal grafts and 91% of them had at least one internal mammary artery distal anastomosis. 19% of patients received a blood product intra-operatively. Eighteen percent of patients required mechanical ventilation for >24 hours. Almost 56% stayed in the ICU for 24 hours or less. Only 7% had pneumonia and only 2.7% developed sepsis. Seventeen percent of patients developed a neurological complication, but only 2% suffered a permanent stroke.

In-hospital mortality was 3.7% and 83% of patients were discharged home. Fifty-two percent of patients spent less than a week in the hospital from the day of surgery. Figure 3-2 illustrates some of the key attributes distributions in the full data set. It clearly displays the imbalanced representation of delirium in our dataset.

Table 3-1: Full Dataset of Patient's Characteristics

	TOTAL (N=5798)
Preoperative	
Age, y, mean (± SD)	67 (11)
Age < 60, %	25.4
Age ≥ 80, %	11.9
Male, %	74
CVD, %	13
EUROII Score, median	2
EF >50, %	72
Frail, %	7
HTN, %	76
DM, %	37
DLP, %	81
COPD, %	14.5
NYHA Class ≥ 3, %	43.5
A-Fib, %	12
Intraoperative	
CABG, %	67
Valve procedure, %	21.5
Cross clamp time, min, mean (± SD)	84 (39)
CPB time, min, mean (± SD)	125 (54)
On-Pump, %	99
≥3 Distal anastomosis, %	48
Intra-operative TEE, %	65
Intra-operative blood products, %	19
Post-operative	
Mechanical ventilation >24hrs, %	18
ICU stay ≤24hrs, %	52
Readmission to the ICU, %	4.5
Cardiac tamponade, %	2.7
Post-operative A-Fib, %	33
Post-operative pneumonia, %	7
Post-operative new renal failure, %	7
Post-operative permanent stroke, %	2
Discharge home, %	82.5
Length of stay from surgery < 1 week, %	52.3
In-hospital mortality, %	3.7
Delirium, %	11.4

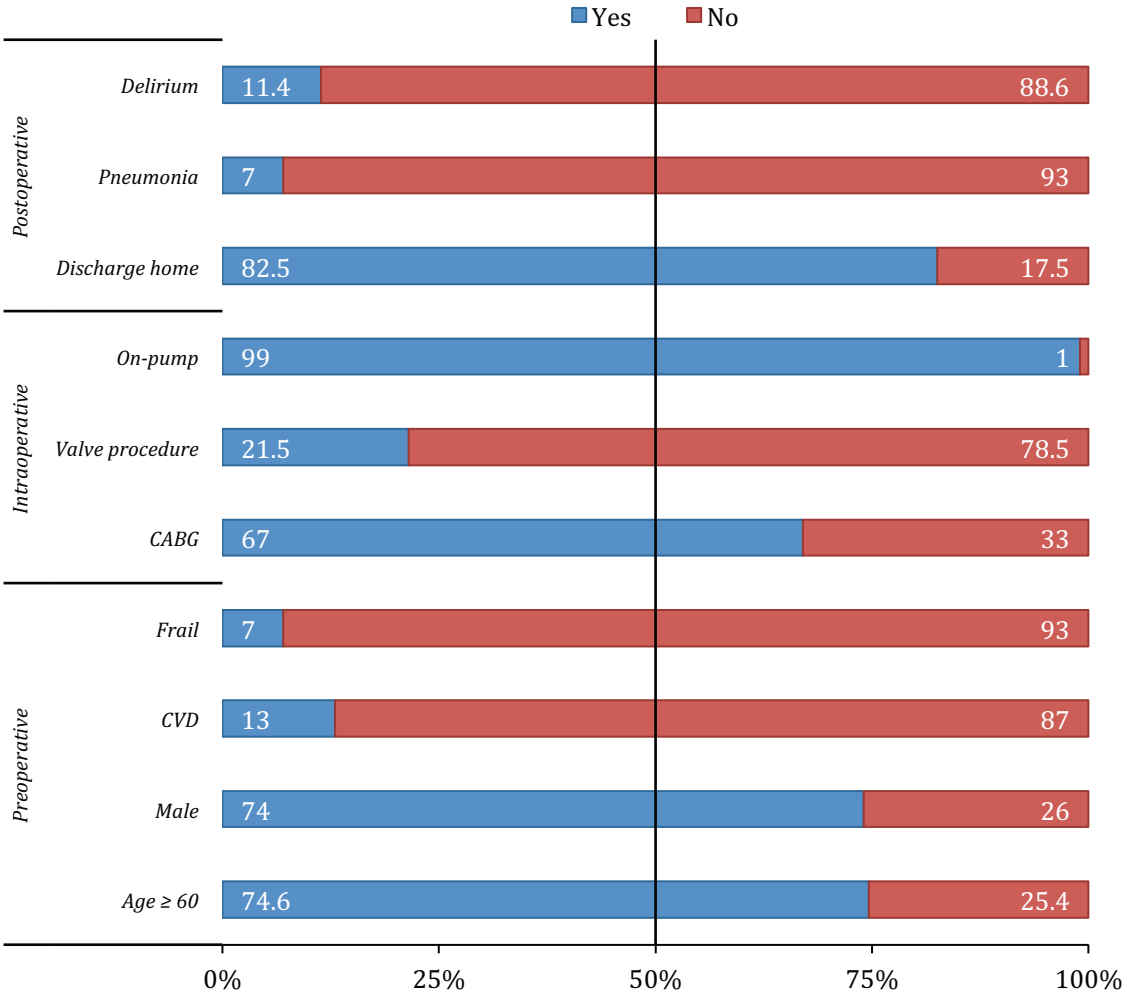

Figure 3-2: Full Dataset of Patient's Characteristics

3.1.8. Delirium in the Full Dataset

After examining the general characteristics of the full original dataset; the influence of the available attributes on delirium was evaluated. As indicated earlier, only 11.4% of the study patients developed post-operative delirium (Table 3-1 and Figure 3-2).

The attributes were divided into continuous and categorical. Secondary to the large sample size, a large number of the independent attributes had a significant p-value (<0.05). Summary tables of all statistical analysis results are provided in APPENDIX B Table B-1 to Table B-12.

CABG surgery was the most frequently performed surgery (67%). Delirium had a higher prevalence in patients who underwent a valve procedure, a combined procedure, required intra-operative trans-esophageal echocardiography (TEE), and required the use of intra-operative inotropes. The length of stay in the ICU was exhibiting a non-Gaussian distribution. The median length of ICU stay for patients who developed delirium was 95 hours (first, third IQR=26, 216) compared to a median of 23 hours (first, third IQR=20, 46) in the patients who did not (See Figure B-2 and Figure B-3). Several approaches were exploited to maintain CVICUhrs as continuous and fitting it to a normal distribution, but unfortunately all our attempts did not improve its implementation. Several authors identified the length of ICU stay as an important predictor of post-operative delirium and because of its skewed behavior; CVICUhrs was discretized into 3 categories based on the ICU literature[12, 19, 23, 24, 52, 76].

Post-operatively, only 7.2% of patients developed pneumonia. However, patients who had delirium were 3.5 times at higher risk of having pneumonia. Fourteen percent of the patients were discharged to another health care institution (e.g.: rehabilitation center, another hospital, or a nursing home), defined as a health care institution transfer to recover as they were not yet fit to go home, but patients who had delirium were 3.8 times more likely to be discharged to an institution. Median in-hospital stay from surgery for patients who developed delirium was 16 days (first and third Quartiles: 10 - 27) compared to 7 days (first and third Quartiles: 5 -10) for patients who did not develop delirium.

3.1.9. Alive Dataset

After excluding patients who had in-hospital mortality (APPENDIX B.I In-Hospital Mortality, Post-Operative Stroke and Delirium page 125), we attempted to guarantee that the new "Alive" data set had a similar representation as the full one. During the study period, the center performed 7,209 open-heart surgeries. The "Full" dataset represented 80% of the surgical procedures performed during the study period. After excluding in-hospital mortality patients, the "Alive" dataset comprised 77.5% during the study period (Figure 3-3).

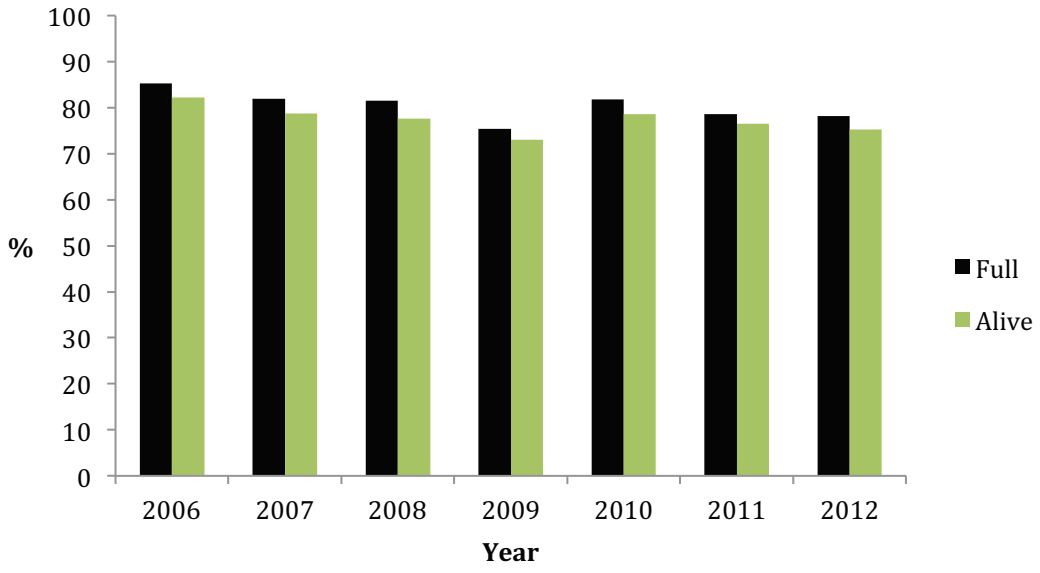

Figure 3-3: MHC Surgical Case Load During 2006-2012
(Datasets representation compared to the total number of cases for each year)

Similarly, the "Alive" dataset's representation was similar to the full dataset in terms of patients' pre-, intra-, and post-operative characteristics (Table 3-2 and Figure 3-5). Delirium was documented in 634 patients (11.4%), which means that we only lost 3 patients due to in-hospital mortality. Mean age was 67 years. The majority of patients were still males (75%). Only 13% had history of CVD. CABG was the most commonly performed procedure (68%). Fifteen and six tenths percent of patients required prolonged post-operative mechanical ventilation. Thirty-three percent of patients developed post-operative atrial fibrillation (A-Fib). Only 6.4% had pneumonia, 1.7% developed sepsis, 1.5% suffered a permanent stroke, 14% of patients were discharged to an institution, and 52% of patients spent less than a week in the hospital from the day of surgery.

After excluding patients with in-hospital mortality, the "Alive" dataset had an almost similar patient's characteristics (Figure 3-4). Discharge home showed a statistically significant difference between the 2 datasets, but this can be justified. As most patents that survive after 30 days from their surgery will eventually be able to be discharged home.

Table 3-2: Alive Dataset Patient's Characteristics

	TOTAL (N=5584)
Preoperative	
Age, y, mean (± SD)	67 (11)
Age < 60, %	26.1
Age ≥ 80, %	9.31
Male, %	75
CVD, %	13
EUROII score, median	1.8
EF >50, %	72.6
Frail, %	6.6
HTN, %	76
DM, %	37
DLP, %	82
COPD, %	14
NYHA class ≥ 3, %	42.5
A-Fib, %	12
Intraoperative	
CABG, %	68
Valve procedure, %	21.5
Cross clamp time, min, mean (± SD)	83 (38)
CPB time, min, mean (± SD)	122(50)
On-pump, %	99
≥3 distal anastomosis, %	48
Intra-operative TEE, %	64
Intra-operative blood products, %	17
Post-operative	
Mechanical ventilation >24hrs, %	15.6
ICU stay ≤24hrs, %	44
Readmission to the ICU, %	4
Cardiac tamponade, %	2.3
New A-Fib, %	33
Pneumonia, %	6.4
New onset renal failure, %	5.8
Permanent Stroke, %	1.5
Discharge home, %	86
Length of stay from surgery < 1 week, %	52.2
Delirium, %	11.4

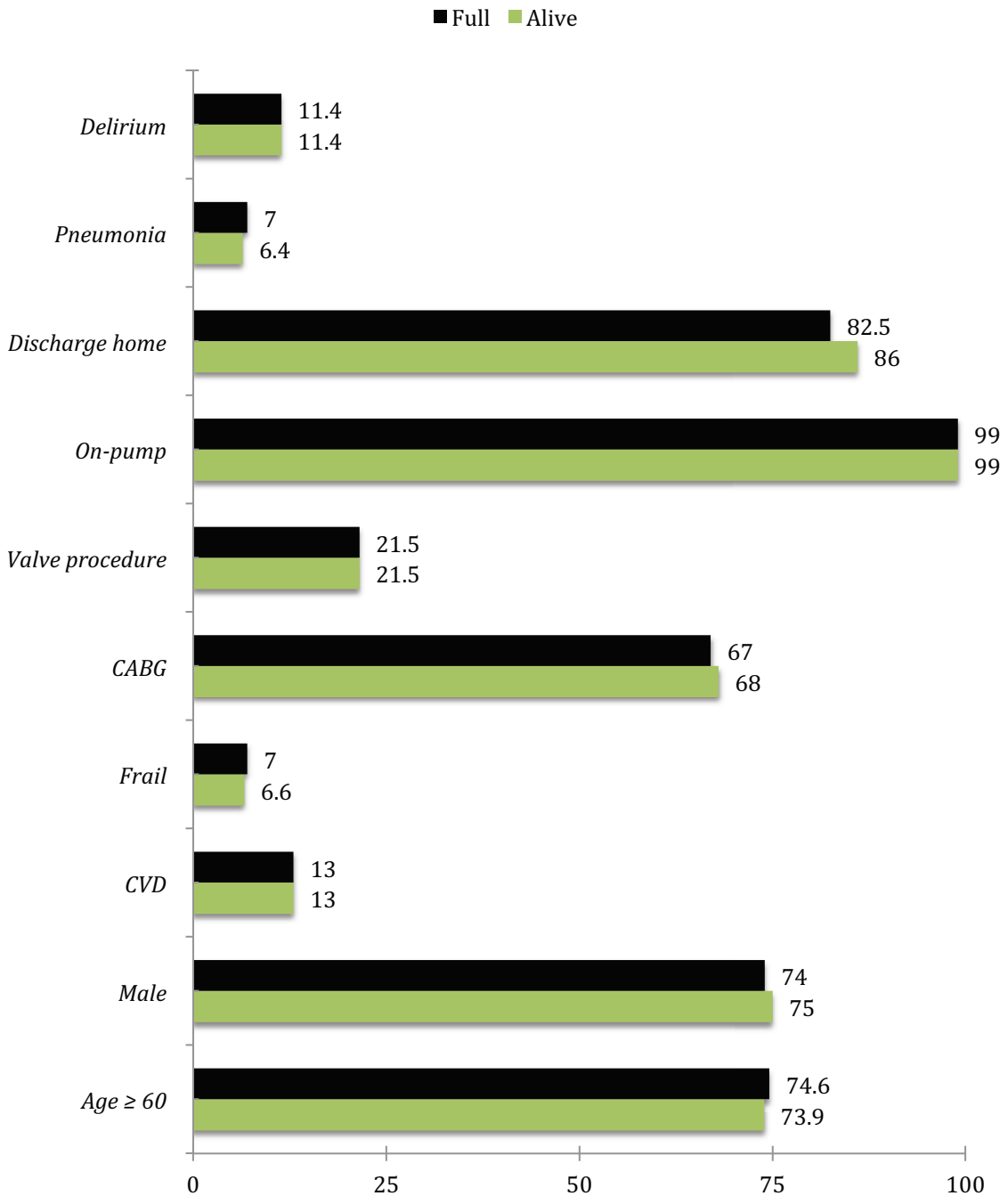

Figure 3-4: Comparison Between the Full and Alive Datasets Characteristics

3.2. Dimension Reduction and Features Selection

In many learning problems there are hundreds or thousands of potential features. The majority of learning methods do not behave well in this circumstance because, from a statistical point of view, examples with many irrelevant, but noisy, features provide very little information. A feature subset selection is a task of choosing a small subset of features that ideally are necessary and sufficient to describe the target concept[105].

The primary objective of dimension reduction is to reduce the number of predictors and try to confirm their independence[106]. In classification, the primary goal is to find a low-dimensional transformation of the feature vectors that retains information needed to predict the class labels[107]. By doing so; we can avoid overfitting, generate a more efficient faster model, improve the understanding of the underlying process, and create model that can be comprehended by the domain experts[108]. A lower dimensional feature space can also improve data visualization and improve exploratory data analysis.

The attributes vector space was reduced down to 92 attributes. After that, several steps were taken in an effort to decrease the attribute vector space to a reasonable one.

3.2.1. Clustering in Medical Data Mining

Clustering is categorizing observations based on their similarity so that 2 observations that belong to a cluster are more similar than 2 observations from a different cluster. In medicine, clustering can be used to group the patients into different categories such as: normal and abnormal; low, medium, or high-risk; class 1, 2, 3, or 4. Several authors have used different clustering as a dimension reduction technique[109]. Similarity measures are usually used for clustering the attributes[64, 65, 109]. In contrast to classification, clustering is considered to be an unsupervised learning technique, in which the class label will not be provided. This means that there is no pre-determined assignment of clusters and the clusters are created based on the available information that is presented to the algorithm.

Different clustering techniques were successfully applied in the medical domain. Belciug et al. used 3 different clustering techniques (k-means, self-organizing map/Kohonen

network, and cluster networks) to detect the recurrence of breast cancer[110]. Escudero et al., was able to divide a dataset of a patient with Alzheimer disease into pathologic and non-pathologic utilizing the patients bio-profile (patient's medical history that is temporally correlated with their biochemical markers) using k-means[111]. Hierarchical clustering methods are heavily used in biochemical medicine and genetics[112, 113]. Other methods have also been used in the medical domain[69, 114].

3.2.2. EM for Dimension Reduction

The expectation-maximization (EM) algorithm is a probabilistic algorithm that belongs to the partitioning clustering methods. It attempts to construct a "latent" attribute that can be used to maximize the likelihood estimate of the model[64, 115]. Several authors have successfully implemented EM as a technique for dimension reduction[106, 107, 116, 117].

Based on our clinical domain knowledge and the patient population, we knew that some of the attributes are correlated. Some examples of these attributes are: diabetes mellitus and the use of insulin, the use of angiotensin converting enzyme inhibitors (ACEI) and angiotensin receptor blockers (ARB), aspirin (ASA) and lipid lowering agents, several pre-operative arrhythmias other than A-Fib, and others. In an attempt to reduce the attribute vector space but simultaneously retain as much information as possible, the EM algorithm was used as a dimension reduction method. Clustering was done with the EM algorithm that is provided in WEKA using the default setting. Seven latent attributes were produced, but only 2 were clinically interesting (DM clustering and preoperative arrhythmia clustering).

3.2.3. Excluded Attributes

Several attributes were excluded from the analysis for the following reasons:

- More than 33% of the observations are missing: these attributes were removed as replacing these observations with their mean or median would lead to questionable results that might be not accurate. The use of classifier-based imputation techniques on the training set was another option[118-120], but it was not attempted as the implemented classifiers adapt to missing values, and several

authors discourage the use of these techniques to impute a large portion of the data (>25%)[118, 121] because it might lead to false conclusions. Some examples of these attributes include: CVA type and Pre-operative coronary catheterization.

- Too infrequent (less than 1% occurrence): these are attributes that are sparse in our dataset and no meaningful conclusions can be drawn from them. Several techniques exist to make use of these sparse attributes, such as generating new artificial observation, resampling, and others. The use of these techniques in the presence of a very small representation will raise some red flags on the validity of these attributes. Some examples of these attributes include: conversion from off-pump to on-pump (9/5584=0.16%) and active endocarditis (37/5584=0.662%).
- Unclear temporal relationship: the main interest in this work is to identify patients who are at risk of developing delirium, which usually occurs within the first 48-72 hours after surgery. Attributes that did not have a clear temporal relationship with delirium were excluded. Some examples of these attributes include: low cardiac output syndrome (in the database definition this attribute can occur at any time within the patient hospital stay) and subsequent or (re-operation after the index operation within the same admission, this can be less or more than 72 hours).

3.2.4. Attributes Selection

After including the new cluster-based attributes (7 new attributes), feature space reduction was applied utilizing:

a. A conventional approach of manual selection based on statistical significance and domain knowledge
b. A data mining approach was used to support the selected attributes by the conventional approach and to uncover overlooked, but important, attributes using feature selection methods

3.2.4.A. Manual Selection Based on Statistical Analysis

All "Date" attributes were used to generate length of stay continuous attributes. Statistical measures of central tendency (mean, median, mode, inter-quartile ranges, standard deviation, and others), shape of distribution, and outliers were examined. Continuous attributes were examined and when the shape of the distribution was close to a normal (Gaussian) distribution, mean and standard deviations were used. When the shape was extremely skewed (non-normal distribution), median, 25% inter-quartile range (IQR) and 75% IQR were used. Categorical attributes are reported with frequency in percent.

Continuous attributes with normal distribution were tested using the student t-test and analysis of variance (ANOVA). Continuous attributes with non-normal distribution were tested using the Wilcoxon-Mann Whitney test. The Kruskal-Wallis test was used for ordinal attributes. Categorical attributes were examined using the chi-square test. Univariate and bivariate analysis of the 92 attributes was conducted, and a candidate list based on a statistical significance of a p-value <0.05 and the appropriate measure of association was composed of 26 candidate attributes (Table 3-3).

Table 3-3: List of Candidate Attributes

Attribute	Attribute
Length of stay in the ICU	CHF
Prolonged ventilation	Pre-op A-Fib
EUROII score	AS
Age	CVD
Procedure difficulty	Clustered DM
Blood product within 48 hrs	Frail
Intra-operative TEE	History of turn down
Timing of IABP	Hemo-dynamic instability
Intra-op inotropes	MR
Pre-op creatinine clearance	Clustered arrhythmia
EF categories	COPD
Pre-op hemoglobin	Pre-op intubation
Pre-op inotropes	Gender

3.2.4.B. Feature Selection Methods

Several feature selection methods were used, aiming to corroborate the statistical analysis and isolate features that were not detected by the conventional approach. For more details on the feature selection methods, please referee to APPENDIX B.VII Feature Selection Methods, page 146. All of the feature selection experiments were conducted in WEKA 3.7[103].

Five different attribute evaluators were independently applied. The filter-based attribute selection methods were applied on the "Alive" Training dataset. The top 30 attributes nominated by each method were matched to the 26 selected attributes based on conventional statistical approach (see Table 3-3). Length of stay in the ICU and prolonged ventilation were really important as they appeared in all 5 methods. EUROII score and blood product transfusion within 48 hours was considered important in 4 out the 5 methods. Nearly all the attributes that appeared 3 times (e.g.: Age, CVD, etc.) were picked up by the same methods (GainRatioAttributeEval, SignificanceAttributeEval, and SymmetricalUncertAttributeEval), this might be due to the fact that these methods rely on the attribute influence probability on the target class (Figure 3-5). Surprisingly, some of the attributes that were picked up by multiple feature selection methods were not deemed important by the classical statistical approach although they were clinically relevant.

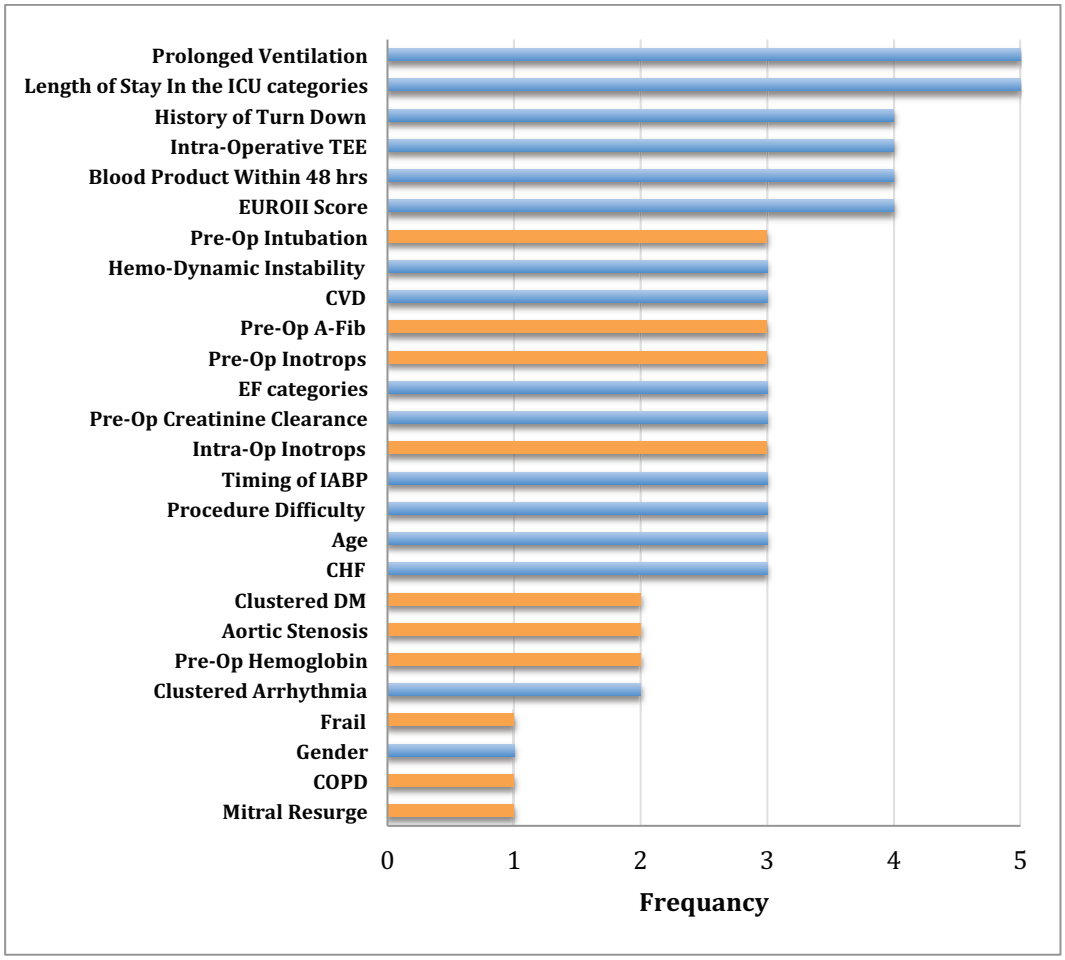

Figure 3-5: Co-occurrence in the top 30 Attributes in Filter Selection Methods *(Blue bar:* the attribute *was* considered to have an influence on the development of post-operative delirium by the conventional statistical approach and some of feature selection methods, *Orange bar:* the attribute *was not* considered to have an influence on the development of post-operative delirium by the conventional statistical approach, but was identified as an important attribute by some of the feature selection methods)

3.3. Description of the Final Dataset

Based on the conventional feature selection approach (statistical and domain knowledge), 26 attributes had a significant influence on the development of post-operative delirium in the "Alive" dataset. At the same time, feature selection methods were independently applied on the "Alive" dataset. Nearly all of the 26 candidate attributes were within the top attributes chosen by several feature selection methods. Attributes with very low frequency or redundant ones were removed (e.g.: hemo-dynamic instability is correlated with European System for Cardiac Operative Risk Evaluation II (EUROII) score and was documented in 70 out 5,584 patients). In the case of highly correlated attributes, the objective one was retained (e.g.: congestive heart failure [29] and EF categories, EF was retained as it is more objective). While "history of turn down" had a low frequency (75 out 5,584 patients), it was preserved in the candidate list as it was highly ranked by all of the feature selection methods. This final dataset was included 22 candidate attributes (Table 3-4 and Figure 3-6) and 1 binary outcome class (delirium: yes/no).

Table 3-4: Final Alive Dataset General Characteristics

Attribute	Type	Possible Values	Missing (N=5584)
Length of stay in the ICU	Ordinal	<24, 24-72, >72 hrs	0
Prolonged Ventilation	Categorical	Yes/No	0
EUROII score	Continuous	0-100 %	0
Age	Continuous	19-95 years	0
Procedure difficulty	Categorical	Single/Combined	0
Blood product within 48 hrs	Categorical	Yes/No	0
Intra-operative TEE	Categorical	Yes/No	16
Timing of IABP	Ordinal	None, pre-, intra-, post-operative	0
Intra-op inotropes	Categorical	Yes/No	2
Pre-op creatinine clearance	Continuous	4-203 ml/min	14
EF categories	Ordinal	<30 30-50, >50 %	25
Pre-op hemoglobin	Continuous	10-196 mg/dl	13
Pre-op A-Fib	Categorical	Yes/No	0
AS	Ordinal	None, trivial, mild, moderate, critical	141
CVD	Categorical	Yes/No	0
Clustered DM	Categorical	Cluster 1 or cluster 2	0
Frail	Categorical	Yes/No	0
History of turn down	Categorical	Yes/No	0
MR	Ordinal	None, trivial, mild, moderate, severe	143
Clustered arrhythmia	Categorical	Cluster 1 or cluster 2	0
COPD	Categorical	Yes/No	0
Gender	Categorical	Male/Female	0
Delirium	Categorical	Yes/No	0

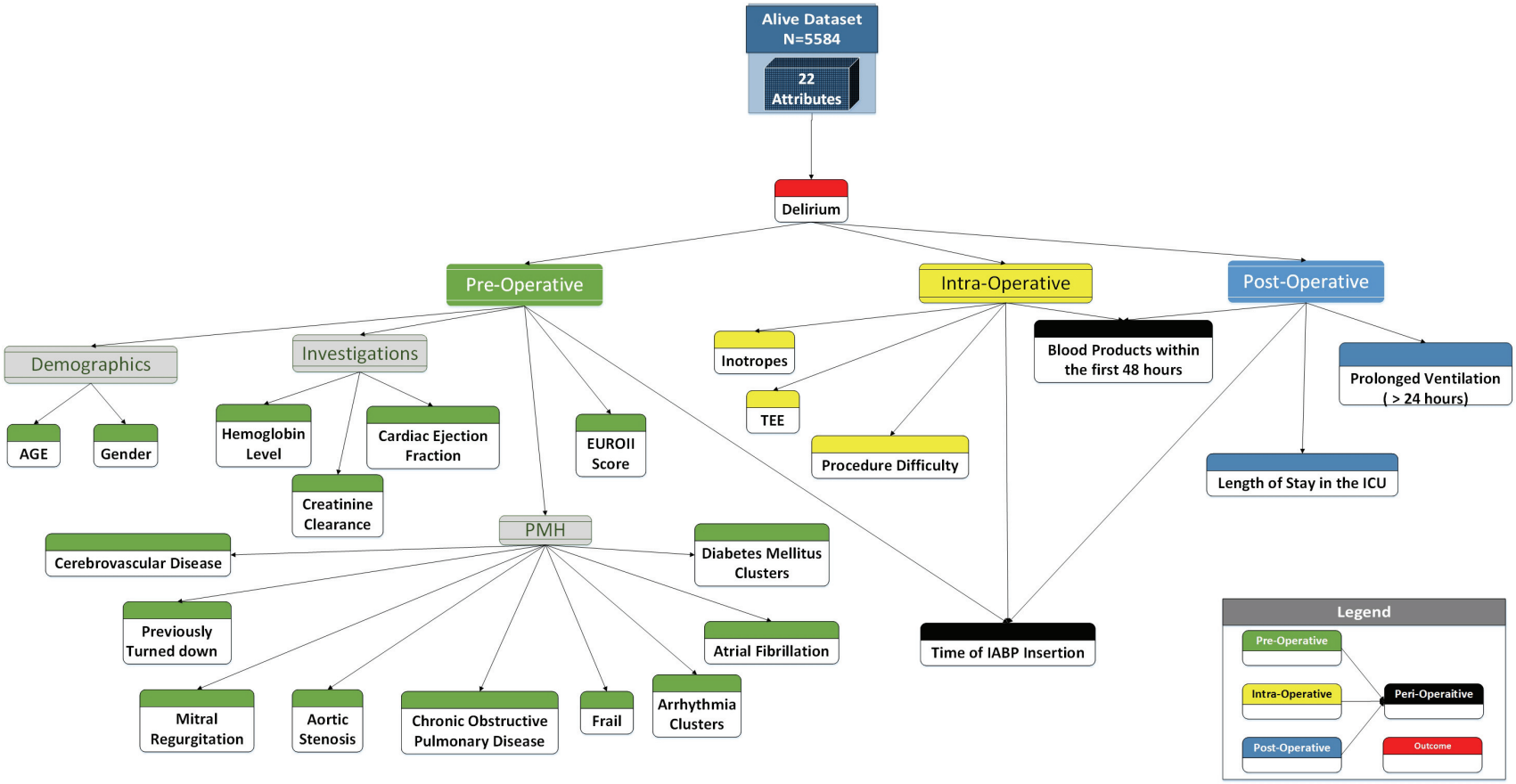

Figure 3-8: Attributes Chart in the Alive Dataset

3.3.1. Selected Attributes Definitions

Delirium: defined as a mental disturbance marked by illusions, confusion, or cerebral excitement requiring medical/physical intervention, a consult, or extends the patient's hospital stay. This definition will only include patients with hyperactive delirium, which usually represent the smaller portion of delirium. In reality, hypoactive delirium is more common in elderly (>65 years) and has been linked to worse outcomes including prolonged ventilation, prolonged hospital stay, and even death [12, 19, 20, 24, 28, 76, 99].

Delirium was detected in 11.4% of the study cohort. The MHC registry is mainly based on chart abstraction, the MHC does not apply a standard delirium detection tool, and the definition dose only include hyperactive delirium; all of these reasons attribute to the low incidence of delirium.

Age: defined as the age of the patient at the time of surgery in years. The mean age for patients who developed delirium was 71 compared to 66 for those who did not.

Frail (Frialty): defined as any impairment in activities of daily living using the Katz index of activities of daily living (ADL), ambulation, or a documented history of dementia. Only 7% of the study cohort was considered frail. The Katz index of ADL, an internationally validated measure of dependency in elderly patients, ranks adequacy of performance in 6 functions (feeding, bathing, dressing, transferring, toileting, and urinary continence). Each function receives a score of 0 or 1, and the total score can range from 0 (complete dependence) to 6 (complete independence)[26, 122].

Length of stay in the ICU: patients were divided into 3 categories based on length of stay in the ICU in hours. 52% of patients spent less than 24 hours in the ICU and 19% stayed in the ICU for more than 72 hours. Of note, the definition of hours in the ICU includes the total hours a patient spent in the ICU that will include readmission and is not specific for the single longest admission. We realize that this might be a weakness in the definition and might bias the data.

Prolonged ventilation (>24 hours): patient required mechanical ventilation for more than a total of 24 hours during the index admission. 16% of patients required prolonged mechanical ventilation. The definition of prolonged ventilation includes the total hours a

patient was on mechanical ventilation that will include re-intubation time and is not specific for the single longest one. We realize that this might be a weakness in the definition and might bias the data.

Procedure Difficulty: defined as a single versus a combined index procedure, regardless of the combination (procedures: CABG, AVR, MVR, or MV-Rep). 12% of the study cohort underwent a combined procedure.

Blood product within 48 hours: defined as a patient receiving a blood product (packed red blood cells transfusion, Platelets, Cryo-participate, or fresh frozen plasma) in surgery and within the first 48 hours from surgery. Twenty-nine percent of the study cohort received a blood product within 48 hours.

Intra-operative TEE: defined as a patient having an intra-operative TEE during surgery. 64% of the study cohort had an intra-operative TEE.

EUROII score: patient calculated European System for Cardiac Operative Risk Evaluation II score. The median EUROII score for patients who developed delirium was 4.2% c**ompared to 1.6% for those who did not.**

Timing of Intra-aortic Balloon Pump (IABP): defined as the time of insertion of an IABP. 5% required an IABP pre-operatively and only 1% required it post operatively.

Intra-operative inotropes: defined as the start of inotropes (drugs to support the circulation) during the operation. Thirty percent of patients required intra-operative inotropic support.

Pre-operative creatinine clearance: pre-operative calculated patient creatinine clearance based on the Jelliffe formula. The mean creatinine clearance for patients who developed delirium was 54.5 ml/min compared to 68.3 ml/min for those who did not.

EF categories: patient's ejection fraction divided into 3 categories. Seventy-three percent had an EF >50% and only 6 % had an EF <30%.

Pre-operative hemoglobin: patient's pre-operative hemoglobin level. The mean hemoglobin level for patients who developed delirium was 125 mg/dl compared to 133 mg/dl for those who did not.

Pre-operative atrial fibrillation: patient's documented history of pre-operative A-Fib. Twelve percent of the study cohort had a history of pre-operative A-Fib.

Aortic valve stenosis (AS): patient's documented history of aortic valve stenosis and its severity. Twenty-one percent of the study cohort had critical AS and 74 % did not have a documented history of AS.

Mitral valve regurgitation (MR): patient's documented history of mitral valve regurgitation and its severity. Five percent of the study cohort had severe MR and 60 % did not have a documented history of MR.

Cerebrovascular disease (CVD): defined as any pre-operative history of transient ischemic attack, reversible ischemic neurologic deficient, cerebrovascular accident, cerebrovascular surgery, or any carotid disease. Twelve and five tenths percent of the study cohort had CVD.

Chronic obstructive pulmonary disease (COPD): defined as a patient requiring pharmacological therapy for COPD, or has a documented FEV1 (forced expiratory volume at the first second) <75% of predicted value for age. Forteen percent of the study cohort had COPD.

History of turn down: defined as a patient who was previously referred for surgery and was labeled non-operable by another surgeon. One and four tenths percent of the study cohort were turned down before.

Gender: seventy four percent of patients were male. Out of the 634 patients who developed delirium, 24% of them were female. The distribution of males and females was almost the same across patients who did and did not develop delirium.

Clustered DM: clustering of DM and Diabetic Control. Sixty-four percent of the patients were in cluster "0."

Clustered Arrhythmia: clustering of pre-operative arrhythmia other than A-Fib (ventricular, AV-Block, complete heart block and others). Twenty-three percent of the patients were in cluster "0."

3.4. Chapter Summary

In this chapter, we discussed an essential data mining task; pre-processing. The data source and the dataset were described. Then, the attributes impact on the outcome of interest (delirium) was evaluated. Candidate attributes were nominated (Table 3-3). In an attempt to reduce the attribute's vector space, clustering of correlated attributes was done using the EM algorithm. Subsequently, the most significant attributes were chosen. The candidate attributes selection was based on the conventional approach (statistical significance, domain knowledge, and problem understanding). The candidate attributes list was then corroborated with the use of several feature selection methods to validate the selection of these attributes and uncover potentially important ones that were overlooked. Ultimately, 22 attributes were considered to have a strong influence of the development of delirium after cardiac surgery. Several pre-operative modifiable features were identified, some of which have not yet been described in the literature (e.g.: frailty, diabetic control, pre-operative hemoglobin, and history of turning down). These features were used as an input vector for building the LR, ANN, and BBN models.

CHAPTER 4: PREDICTIVE MODELS FOR DELIRIUM

"Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful."

George E. P. Box (1919-2013)

Outcome prediction is considered a classification problem in the data-mining world. According to the Merriam-Webster online dictionary, classification is defined as "the systematic arrangement of things in groups or categories according to established criteria and observed similarities"[123] .

In predictive modeling, it is vital to appreciate the nature of the predicted class and the impact of class imbalance on the model performance. The most commonly used model performance measure is predictive accuracy[58, 64, 65, 68]. The use of predictive accuracy in the presence of class imbalance is inappropriate and can lead to false conclusions[58, 59, 64, 65]. In the presence of class imbalance, others measures of performance should be used to compare different models[56-59, 64, 65, 67, 68, 124, 125].

In this chapter we start by explaining the steps that were taken to prepare the final dataset for model building (Figure 4-1). Then, model performance evaluation measures will be explained in the case of balanced and unbalanced class representation. We will address the issue of class imbalance, its effect on model building, and proposed solutions to overcome its negative effect. After that, we will discuss the currently available models for detecting delirium after cardiac surgery. Then we will describe the different molding techniques that were used in this work. After that, we will describe the application of these modeling techniques on the "Alive" dataset in an effort to generate a model that is capable of predicting delirium. We start by developing a Logistic regression model that will be used as a reference model that other models will be compared too. Several models will be developed and compared to the reference model to discern the best model. Then several experiments will be conducted in an effort to mitigate the effect of the outcome class imbalance on model performance.

4.1. Classifier Building and Performance Evaluation

This section will explain the steps that were taken to prepare the "Alive" dataset, described in chapter 3, for model/classifier building and evaluation. Then, model/classifier performance evaluation measures will be explained. The choice of these measures is based on an extensive literature review that was conducted, in addition to recommendations from authorities in data mining and predictive analytics domains[56-59, 64, 65, 67-70, 124, 126-128]. Figure 4-1 summarizes the classifier building process and evaluation methodology for the classification experiments.

The low prevalence of delirium (11.4%) results in a target class imbalance. Here, only 1 of every 9 patients in the "Alive" dataset exhibits post-operative delirium. We will discuss the negative effects of target class imbalance on the model performance and how to deal with it[57-59, 64, 65, 69, 70, 125].

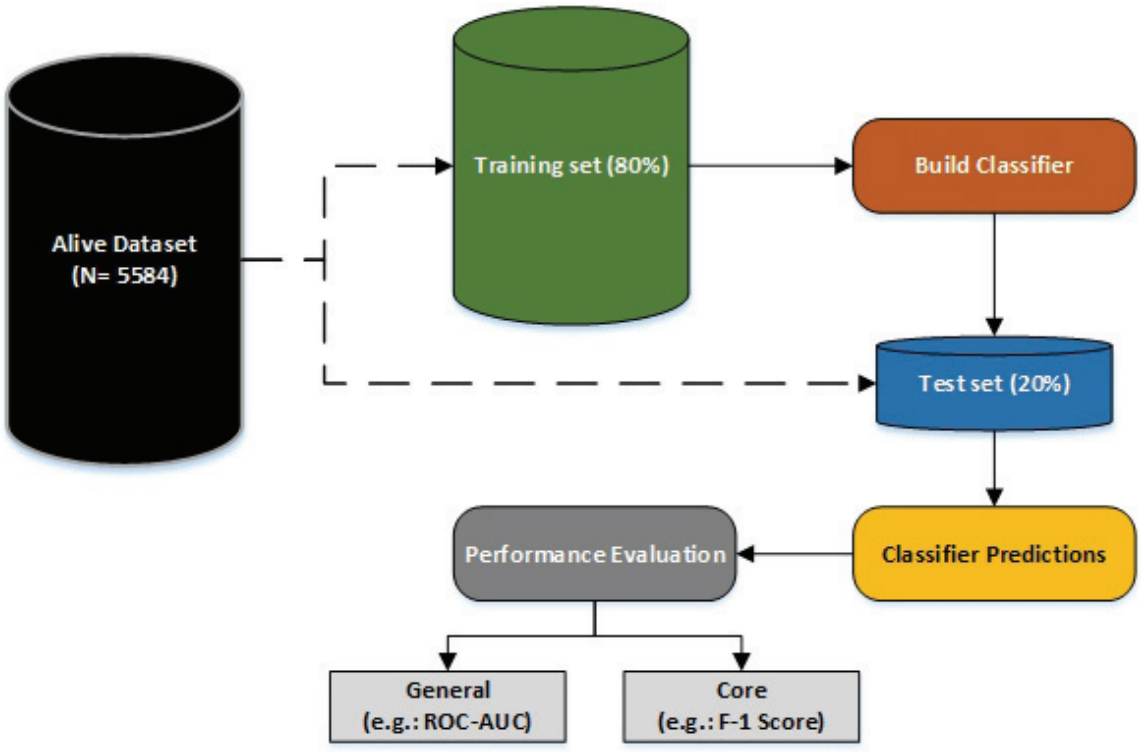

Figure 4-1: Overview of classifier building and evaluation methodology

4.1.1. Hold-Out or Cross Validation

Several methods are available for testing a model's performance. The effectiveness of these methods – and, ultimately, the choice of method – depends primarily on the amount of data available. The goal is to maximize the amount of data available for training, in order to support the algorithm learning process. Cross-validation starts by dividing the data set in "k" mutually exclusive partitions, which will represent the test sets for each of the "k" models that will be built using the remaining data. The hold-out method splits the original full data set into 2 partitions, one for training and another for testing.

The cost of the hold-out method relates to the amount of data kept out for testing. If the model is trained on a small dataset, its error will be exaggerated. Such pessimistic predictions are always better in real life implementations in comparison to overconfident ones[128]. The hold-out method is usually preferred when the dataset is large: while there is no exact threshold for choosing this method, most authors prefer it when there are more than 1,000 observations in the dataset[56, 64]. Experts encourage stratification (stratified hold-out sample), which ensures that most attributes – particularly the outcome/class – have a comparable distribution across datasets[65]. Some of the pros and cons of the Hold-out method are displayed in APPENDIX C.

The stratified Hold-out method was used here in order to ensure equal representation of essential attributes across the datasets (Table 4-1), to reassure the generalizability and applicability of the results. An 80/20 split was used, with 80% for training and 20% for testing. This split, chosen in order to maximize the amount of available data for the learning step, resulted in a training dataset of 4,467 patients and a testing set of 1,117 patients. This final datasets (training and test) had a list of 22 candidate attributes and 1 binary outcome class (delirium: yes/no). The characteristics and missing values frequency counts for each attribute in the training and test subsets are summarized in Table C-14, page 151.

Table 4-1: Training and Test Subsets General Characteristics

Attribute	"Alive" Dataset (N=5584)	
	Training (N=4467)	Test (N=1117)
Delirium, n (%)	507 (11.4)	127 (11.4)
Male gender, n (%)	3327 (74.5)	832 (74.5)
EF, n (%)		
• <30%	261 (5.8)	51 (5)
• >50%	3228 (72.2)	809 (72.5)
Frail, n (%)	289 (6.5)	78 (7)
EUROII score, median	1.7	1.8
Pre-operative creatinine clearance based on Jelliffe formula, mean (SD)	67 (25)	68 (25)
Pre-operative A-Fib, n (%)	527 (11.8)	131 (11.7)
CABG procedure, n (%)	3034 (67.9)	742 (66.5)
Blood products transfusion within 48 hours, n (%)	1267 (28.4)	327 (29.3)
Mechanical ventilation >24 hours, n (%)	695 (15.6)	179 (16)
CVICU length of stay in hours, n (%)		
• <24	2523 (56.5)	623 (56)
• >72	824 (18.5)	215 (19)

4.1.2. Evaluating Classifier Performance

In the presence of class imbalance, predictive accuracy may not be the most appropriate measure of model performance[57-59, 64, 65, 125]. As a result, more objective evaluation measures, such as F1-measuere, Kappa statistic, ROC-AUC and others may be used[57-59, 64, 65, 125]. These measures are derived from the contingency table produced by the classifier.

A binary classifier will generate a 2x2 confusion matrix (contingency table) that describes the performance of a model/classifier with its predictions (predicted) against the actual target classes (actual). The confusion matrix of a binary class problem and the resulting definitions of measures used in this work are defined in Table 4-2.

Table 4-2: Confusion Matrix Example

		Predicted Class		
		+	-	
Actual Class	+	True Positive	False Negative	Sensitivity (Recall)
	-	False Positive	True Negative	Specificity
		PPV (Precision)	NPV	Accuracy

Sensitivity (recall) relates to how inclusive the test is; that is, the degree to which it can correctly detect observations with the condition. In a medical context, it is the probability that a test will be positive in an individual that actually have or will develop the disease. Specificity relates to the test's ability to correctly identify observations without the condition. In a medical context, it is the probability that a test will be negative for individuals that actually do not have and will not develop the disease.

Accuracy refers to the test performance using the entire data set, irrespective of the classes or their distribution. It relates to a test's ability to correctly identify observation assignments. In the presence of class imbalance, this measure will be misleading because the minority class (positive cases) has a small contribution to the error rate, such that the algorithm will tend to favor the majority class.

In the case of the "Alive" dataset, the algorithm would wrongly label all positive observations in the test dataset as negative (127 positive for delirium); the predictive accuracy of such a model would be 89%. Given the importance of early detection and prevention of delirium[12, 21, 24, 25, 36, 76], missing high-risk patients has serious consequences in the case of delirium.

Because predictive accuracy is not suitable measure in this case, the ROC-AUC was used. The ROC-AUC is a general model performance measure that is common in several disciplines (e.g., medicine, engineering, computer science, biology, and others), and is the advocated measure in the presence of class imbalance[56-59, 64, 65, 68, 124, 125]. the receiver operator characteristics (ROC) illustrates the performance of a classifier in a 2-dimensional graph by plotting the false positive rate (1-specificity) on the x-axis against the true positive rate (sensitivity) on the y-axis for all potential points, thus demonstrating the trade-off between true and false alarm rates.

The area under the curve (AUC) refers to the area covered by the model in the square formed by the x-axis and y-axis. It represents the overall accuracy of a test, ranging between 0 and 1. It is equivalent to the probability that a model will appropriately rank a positive instance higher than -1.0[126, 127, 129, 130], with values approaching 1.0 indicating higher sensitivity and specificity. Values close to 0.5 indicate accuracy similar to that of random chance[126, 127, 129, 130]. The ROC-AUC allows visualization of performance over a spectrum of different conditions, rather than relying only on a point estimate (e.g., accuracy)[131].

When comparing several models developed from the same dataset, Hanley and McNeil[132] developed a method to that takes into account the correlation between the AUC induced by the paired nature (same cases) of the data. This test was used to compare the ROC-AUC of the different models.

If several models had a comparable general performance, the McNemar's test should used to verify the ROC-AUC test results. Mainly used to analyze matched pairs of data, McNemar's test has been described and used by several authors in machine learning literature, and some authors advocate its use as an adjunct to the ROC-AUC for the development and improvement of algorithms[62, 133].

Precision (positive predictive value (PPV)) refers to the degree of correctness. It is the percentage of actually positive observations that were considered by the test as positive. In a medical context, it is the probability that a patient has the disease if the test is positive.

The Kappa statistic is a measure that compares observed accuracy with expected accuracy (that of random chance)[134]. It is used to evaluate a single model or evaluate different models that were evaluated on the same data, and is often used as a measure of reliability when 2 or more independent observers are evaluating the same thing[135]. In machine learning, one rater is the actual labeled values and the other rater is the algorithm used to perform the classification[64-66, 134]. The Kappa statistic is standardized to fall on a scale of (-1, 1). A value of 1 indicates perfect agreement, 0 indicates the degree of agreement expected from random chance, and negative values indicate potential systematic disagreement[135].

The F1-score is defined as the harmonic mean of precision and recall[64, 65]. When comparing ratios, the harmonic mean gives a more realistic picture of the true mean compared to the arithmetic mean[61, 66]. Since the harmonic mean of a list of numbers tends more strongly towards the least elements of the list, compared to the arithmetic mean, it tends to diminish the influence of outliers and augment the impact of small ones[136]. Thus, it expresses a realistic picture of system performance, essentially demonstrating the worst-case scenario.

For further details on the applicability of a screening test in relation to its performance measures, please refer to APPENDIX C: MODELING, C.III Screening Test and Performance Measures in the Context, page 159.

4.1.3. Dealing with Imbalance

Data sets are unbalanced when one or more of the classes represent a small proportion of the set (called the minority class) while other classes make up the majority. In this situation, classification algorithms tends to predict the majority class very well but perform poorly on the minority class, an effect that has been attributed to 3 main reasons[57-59]: 1) the goal of minimizing the overall error (maximize accuracy), to which the minority class contributes very little; 2) the algorithm's assumption that classes are balanced; and 3) the assumption that impact of making an error is equal. In reality, these assumptions usually do not hold.

Class imbalance is a popular problem in the data science community[56-59, 137]. Two main strategies for dealing with the issue depend on the level of the intervention (data level manipulation or algorithm parameters manipulation)[56-59]. For more details, please refer to APPENDIX C: MODELING, C.IV Addressing Class Imbalance, page 160.

In an attempt to fairly evaluate the effect of data manipulation on the developed model's performance, both approaches were independently applied. Two data level techniques (spread sub-sampling and Synthetic Minority Over-sampling Technique (SMOTE)[58]) and 1 algorithm parameter manipulation technique (cost sensitive classification[59]) were used in this work.

4.1.4. Classifier Performance Evaluation

The "Alive" dataset is reasonably large (5,584 observations) and is composed of a homogeneous population of cardiac surgery patients, with a prevalence of delirium of only 11.4% (Figure 4-1 & Table 3-4). During preparation of the training and testing sets, the distribution of several attributes was intentionally preserved across the 2 datasets in order to ensure the reproducibility and validity of the tested algorithms, with the specific aim of minimizing the impact of prevalence upon precision and therefore decreasing the need for re-calibration.

Due to the large imbalance within the dataset, with far more negative cases (1:9 ratio of delirium: no delirium), accuracy was not the appropriate measure to use[56-59, 131]. The ROC-AUC was used to evaluate the model's general performance. Compared to accuracy, the ROC-AUC, takes into account the class distribution and gives more weight to correct classification of the minority class, thus generating genuine results[56-59, 70, 126, 127, 129, 130].

4.2. Current Predictive Models for Delirium after Cardiac Surgery

Undergoing cardiac surgery is an independent predictor for developing delirium[17, 24, 25, 78]. The diagnosis of delirium is mainly based on clinical suspicion. Currently, multiple assessment tools are available, including the Abbreviated Mental test, clock-drawing test, MMSE, CAM, Delirium Index, CAM-ICU, and the Richmond Agitation Sedation Scale[12, 21, 24, 25]. While all these detection tools aim to detect delirium in the post-operative setting, none of them are specific to cardiac surgery patients.

Several authors have tried to create models that can detect cardiac surgery patients who are at higher risk of developing post-operative delirium. Existing research relies mainly on LR as the main statistical modeling tool. Afonso et al. developed a predictive model for the detection of post-operative delirium in cardiac surgery patients by conducting a prospective observational study with a series of 112 patients[17]. Using LR, they were able to show that increased age and procedure time were independent predictors of post-operative delirium. In a retrospective review with 2,160 patients – of which only 90 developed delirium – LR was used on a small number of patients (n=16) in order to develop a model to predict severe delirium[20].

Stransky et al. conducted a prospective study on 506 patients who underwent CABG and/or valvular surgery, primarily focusing on hypoactive delirium after cardiac surgery[28]. Patients who developed hyperactive and mixed delirium were excluded. Forty-two patients (9%) had documented hypoactive delirium within the first 3 days after surgery. A LR model was developed; several factors (age, preoperative depression, preoperative diuretics, aortic cross clamp time, and number of pRBCs transfused) were associated with hypoactive delirium. The model was not tested on an independent validation cohort. Smulter et al., focusing on patients 70 years and older who were undergoing cardiac surgery (n=142), found that combining predisposing and precipitating factors resulted in a better model performance (ROC: 0.802 compared to 0.729 for precipitating and 0.695 for predisposing)[19]. Katznelson et al.[37] had the largest cohort (n=1,059 patients). Using stepwise LR, they identified several independent predictors of post-operative delirium, including older age, preoperative depression, preoperative renal

dysfunction, complex cardiac surgery, perioperative intra-aortic balloon pump support, and massive blood transfusion with an ROC of 0.77. In this study, bootstrap sampling was used to validate the model.

4.3. Contemporary Classification Methods for Forecasting Delirium

A predictive model is considered to be useful if it is simple and easy to calculate, has a clear structure, and is validated in independent data sets with good generalization[138].. Several machine-learning algorithms have been utilized to solve real life medical problems[69, 70, 87-89, 93-97, 109-111, 113, 114, 139-145]. Some of the commonly applied algorithms to solve medical classification problems are: LR, ANN, Bayesian techniques, decision trees and support vector machines, K-nearest neighbor and ensemble methods[53, 69, 86-89, 93, 94, 96, 97, 109, 111, 113, 141-147].

Studies in the medical literature primarily focus on the classical statistical approach, LR. Yet, this technique has not been able to produce consistent results [8, 10, 11, 13, 15, 17-20, 22-24, 29, 51, 52, 75, 79, 180]. No other approaches have been explored in the medical community.

The choice of algorithms in this study was based on the commonly used methods that are usually compared to LR in the medical data mining literature. In medical data mining classification tasks, LR is usually compared to ANN[144, 148-150]. Bayesian approaches have been used in the medical domain but are less common because of their computational complexity, and are rarely compared to LR[66, 67, 124, 125, 147, 151]. Compared to LR, ANN and BBN allow complex nonlinear interactions, can handle missing data, and are capable of self-learning. All of these features allow them to handle the complexity that characterizes medical data and its nature.

The following section will briefly introduce the 3 modeling approaches that were used in this work: LR, ANN, and BBNs. For more details, please see APPENDIX C: MODELING.

4.4. Predictive Models

4.4.1. Logistic Regression

Logistic regression, the preferred and most widely used binary classification method in medical literature [8, 10, 11, 13, 15, 17-20, 22, 23, 28, 29, 35, 37, 51, 52, 75, 79], examines the relationship between a binary outcome (dependent) attribute, such as presence or absence of disease, and predictor (independent) attribute(s), which may be continuous (e.g., age), categorical (e.g., gender), or even ordinal (e.g., level of education). The presence or absence of a disease within a specified time period may be predicted based on many attributes including, the patient's age, past medical history, and family history [53, 54]. LR assigns a regression coefficient to each attribute, indicating the amount of influence it has on the outcome. This effect is usually expressed as an odds ratio (OR), which represents the effect by which the odds of an outcome change for a 1-unit change in the independent attribute[53, 54, 144].

One of the specific goals of LR, and of predictive modeling in general, is to generate a concise model that explains the outcome while neither over-fitting, nor losing important information. Another important consideration when building the model is the candidate attributes selection method. Candidate attributes selection methods, which include direct, sequential, and stepwise; are all based on assessing the statistical contribution of the attribute to the improvement of the model fit. Stepwise selection chooses attributes based on predefined statistical criteria, as influenced by the data itself [54].

A major advantage of LR is that it creates simplified picture of the relationship between the outcome and predictors that can be easily explained, since odds ratios are interpreted directly. One of its major drawbacks is that, by creating a linear combination of attributes, it cannot handle nonlinear complex interactions, which are characteristic of complex biological, chemical, and eco-systems. For more details, please refer to APPENDIX C: MODELING Logistic Regression, page 164.

4.4.2. Artificial Neural Networks

Artificial neural networks (ANNs) are computational models inspired by the biological function of the nervous system, specifically the brain. It consists of highly interconnected neurons (nodes), and the overall ability to predict outcomes is determined by the connections between these neurons [64, 66]. In comparison to LR, ANNs apply nonlinear mathematical models that simulate the brain's own problem-solving processes[150]. They are considered to be complex non-linear systems that can deal with noisy or incomplete data, allow multiple and simultaneous multilevel interactions, and have a high ability to generalize based on the input [57, 65, 144]. Like a network of neurons in the brain, they learn by adjusting the connection weights between present neurons. ANNs are dynamic models that learn from reality (adaptive learning), exhibit self-organization and parallel processing, deal efficiently with non-linear relationships, and feature high tolerance to redundant information[152].

Usually, a neural network is organized in layers, each of which is composed of several nodes (neurons). The layers are usually divided into 3 main categories: input, hidden, and output (Figure 4-2). In a classification task, the input layer represents the attributes and the output layer represents the outcome; it is the hidden layer, representing the actual processing of information, where most of the work happens. There is usually a single level in the input and output layers. In contrast, the hidden layer can be made of multiple levels. Increasing the levels in the hidden layer will increase the network complexity, although doing so may lead to overfitting and does not necessarily improve performance [64-66].

In back-propagation neural networks (BPNN), the most commonly used architecture in classification tasks, learning occurs with each cycle through a forward activation flow of outputs combined with backwards error propagation of weight adjustments. In a binary classification, the sigmoid, or logistic, function is usually used as an activation function, allowing the BPNN to model classification problems that are linearly inseparable[64]. BPNNs suffer from some drawbacks: it is difficult to directly interpret the symbolic meaning of the denoted connection weights, it is unpredictable because the network attempts to discover the optimal solution independently (black box modeling), and it also

tends to be slower to train. In binary classification, the assignment of an observation to one possible outcome is based on the output threshold of ≥0.5 probability of the observation belonging to that class[64, 65]. For more details, please refer APPENDIX C: MODELING Artificial Neural Networks, page 166.

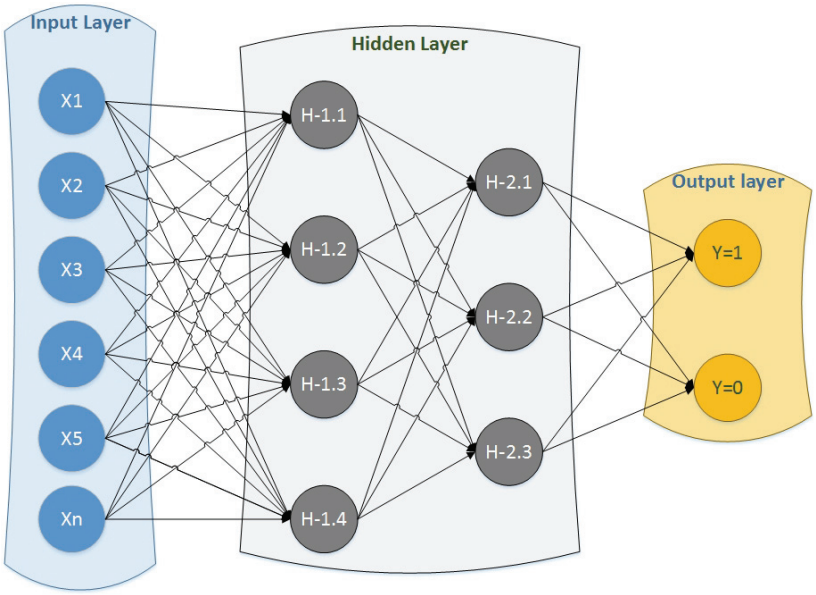

Figure 4-2: ANN Architecture

4.4.3. Bayesian Belief Networks

Probability refers to the likelihood that something is going to happen in the future. It is usually represented on a scale between 0 and 1, and can be represented mathematically with the following equation:

$$Probability\ of\ an\ outcome = \frac{Number\ of\ times\ that\ outcome\ was\ observed}{Total\ number\ of\ all\ possible\ outcomes}$$

In decision-making, probabilities, because they have an inherent uncertainty to them, are usually thought of as guides rather than absolutes[153]. In Bayesian terms, probability measures a degree of belief. In the case of multiple attributes, it assumes that all of the attributes are independent from one another (class-conditional independence). In reality, however, this is rarely true. In biological systems, for example, there are many interconnected and simultaneous interactions that are heavily influenced by earlier events. BBNs were introduced, in part, to mitigate the effects of class-conditional independence [154, 155].

BBNs decompose the problem space to multiple smaller subspaces, identifying a joint conditional probability distribution that is relevant to each subspace and constructing links (arcs) between these subspaces that are based on a probabilistic dependency pattern.

A BBN has 2 main components: a directed acyclic graph (DAG) and a conditional probability table (CPT) for each attribute[64, 65]. A directed graph represents the causal relationship between a parent attribute and its children [67, 125, 155]. The network topology may either be imposed by a domain expert or inferred from the data [64, 67, 125]. Each attribute will have a CPT that is only based on the dependencies that it has with its antecedents/parents[67, 125, 155].

The main advantages of this approach are that it acknowledges dependencies between attributes, provides a simple but elegant graphical representation of the relationships, handles noisy data, provides causal or evidential relationships, and can be easily interpreted by humans and machines; in addition, model parameters have a clear semantic interpretation[67, 125, 156]. Its limitations relate to its heavy dependence on the quality of the data, its inability to handle continuous data (i.e., it requires discretization), and its

requirement of complete data (no missing data, otherwise it auto imputes). For more details, please refer to APPENDIX C: MODELING Bayesian Belief Networks.

4.5. Forecasting Delirium in Cardiac Surgery

Logistic regression is the solely preferred modeling approach used for predicting delirium in the medical literature[8, 10, 11, 13, 15, 17-20, 22-25, 27-29, 35-37, 51, 52, 75, 76, 79]. In this work, the first step was to develop a LR model from the "Alive training" dataset. This model will be used as a reference model that will be compared to the machine learning models. Following that, 4 more models were generated using the same "Alive training" dataset using ANNs and BBNs. Each model was then applied on a test dataset and their prediction performance was compared.

Each classifier general performance was assessed with: ROC-AUC, sensitivity, and specificity [57-59, 64, 65, 125]. The Hanley and McNeil repeated measures ROC test was used to compare the ROC-AUC of the different models[132]. The McNemar's test was used to verify the ROC-AUC test results[62, 133]. Three core performance measures (Kappa statistics, precision, and F1-score) were used to differentiate equivalent models[57-59, 64, 65, 125].

The classification task consisted of 4 main experiments:

- In experiment 1, the original training data was used to develop the initial models and compare their performance (un-manipulated training subset). The goal of this experiment is to develop models from a dataset that resembles the actual prevalence of the outcome class without any manipulation and compares their performance to discern the best performing model that is capable of detecting patients that are prone to develop delirium (positive cases).
- Experiments 2-4 were conducted to explore different recommended methods for mitigating the effect of class imbalance on model performance. It is not practical to change the model every time a new observation is added, and a common practice is to re-evaluate the model based on the variation in the data or strategic

initiatives and decisions[157-162]. The main goal was to evaluate the impact of changing the classifier's environment upon its performance and distinguish which method will significantly improve the classifier ability of detecting positive cases without compromising other metrics

- ✓ Experiments 2: a data-level manipulation method, over-sampling with SMOTE, was applied on the training dataset
- ✓ Experiments 3: a data-level manipulation method, under-sampling with SpreadSubSample, was applied on the training dataset
- ✓ Experiments 4: an algorithm-level manipulation method, cost sensitive classification, was applied on the classifiers

Classification techniques were carried out using the open source Waikato Environment for Knowledge Analysis (WEKA), SAS V9.3, and R V3.0.1[101, 102, 104].

The systematic approach used for all of the experiments included dividing the final dataset into an "Alive training" and "Alive test" subsets. The "Alive training" set was used to construct the models using the mentioned algorithms. Afterward, the models performance was evaluated on the "Alive test" set.

In order to ensure the uniformity and reproducibility of the results, all data-level manipulation techniques were applied only on the "Alive training" subset. To avoid the data manipulation on the developed models, the "Alive test" subset was not subjected to any of the manipulation techniques. All of the reported measures were based on the performance of the developed models on the "Alive test" set. Figure 4-3 and Figure 4-4 summarize the steps taken in the experiment.

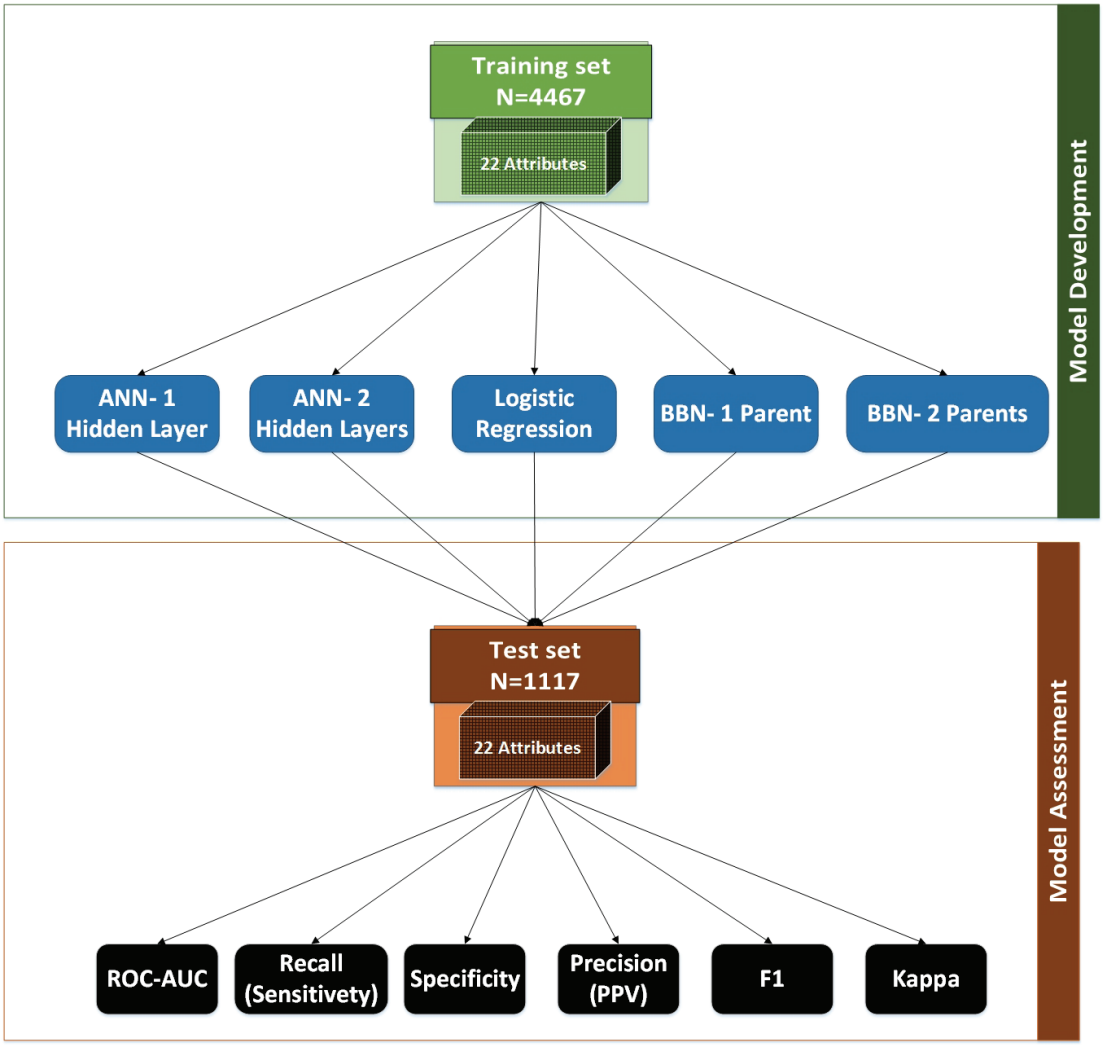

Figure 4-3: Experiment 1 - Original data with class imbalance

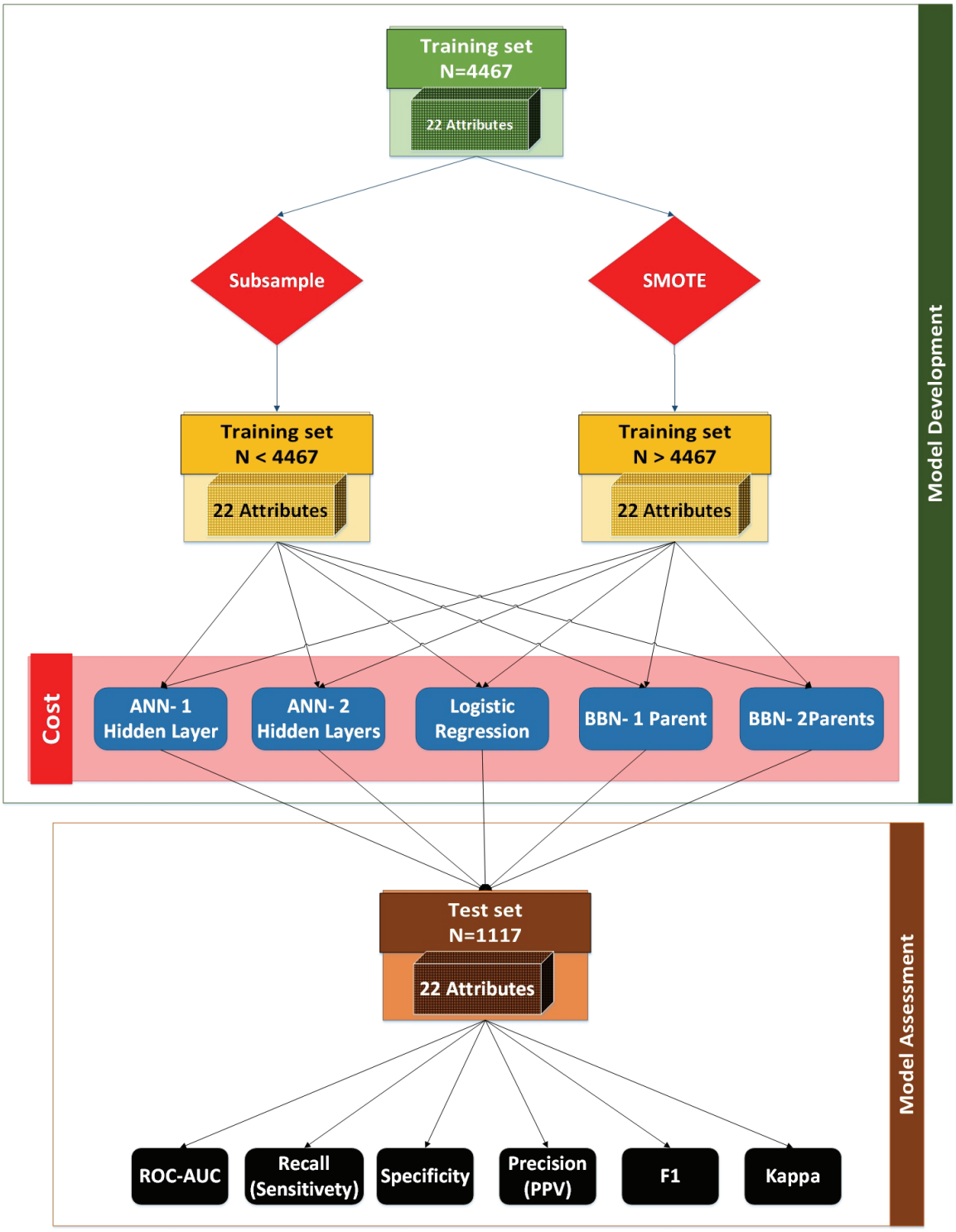

Figure 4-4: Experiments 2, 3, and 4 – Class imbalance manipulation techniques

4.5.1. Experiment 1: Original Training Set

Stepwise logistic regression was used to develop the reference model, in addition to 2 ANN models and 2 BBN models were developed. The general and core performances of these 5 models were then assessed.

4.5.1.A. Experiment 1- Logistic Regression

This model was developed in SAS V9.3. The SAS "Proc Logistic" function was used, with an alpha=0.05. Figure C-17 demonstrates the final model SAS code.

The stepwise approach started by creating dummy attributes for the categorical data, identifying an intercept with a starting -2Log Likelihood=3,059.85. Via 8 steps, the analysis converged with an ROC-AUC=81.7% for the training set and identified 8 attributes as predictors of delirium, with a -2Log Likelihood=2,485.9 for the final model. The Hosmer and Lemeshow goodness-of-fit test, which tests a null hypothesis where there is no difference between the observed and predicted values of the outcome, was applied. The result of this test was significant (Chi2=30.2, p-value=<0.05), such that the null hypothesis can be rejected and the model is considered to be a rational clarification of the available data. The pseudo R2 was 12.5% and max-rescaled R2 was 24.5%.

The multivariate stepwise LR analysis deemed 8 out of the 22 candidate attributes to be important independent predictors of post-operative delirium, 4 of which were preoperative and 4 of which were intra- and post-operative (Table 4-3). Figure 4-5 demonstrates a forest plot of the odds ratios of developing delirium in the developed model. Odds ratio (OR) is a measure of association between an exposure and an outcome. For example, in this model, the odds of developing delirium (outcome) in a patient that had prolonged ventilation (exposure) are 60% higher than the ones who did not with an effect ranging from 17% to 117% in the studied population. Patients who stayed in the ICU for more than 72 hours had a 500% higher chance of developing delirium compared to the ones who did not.

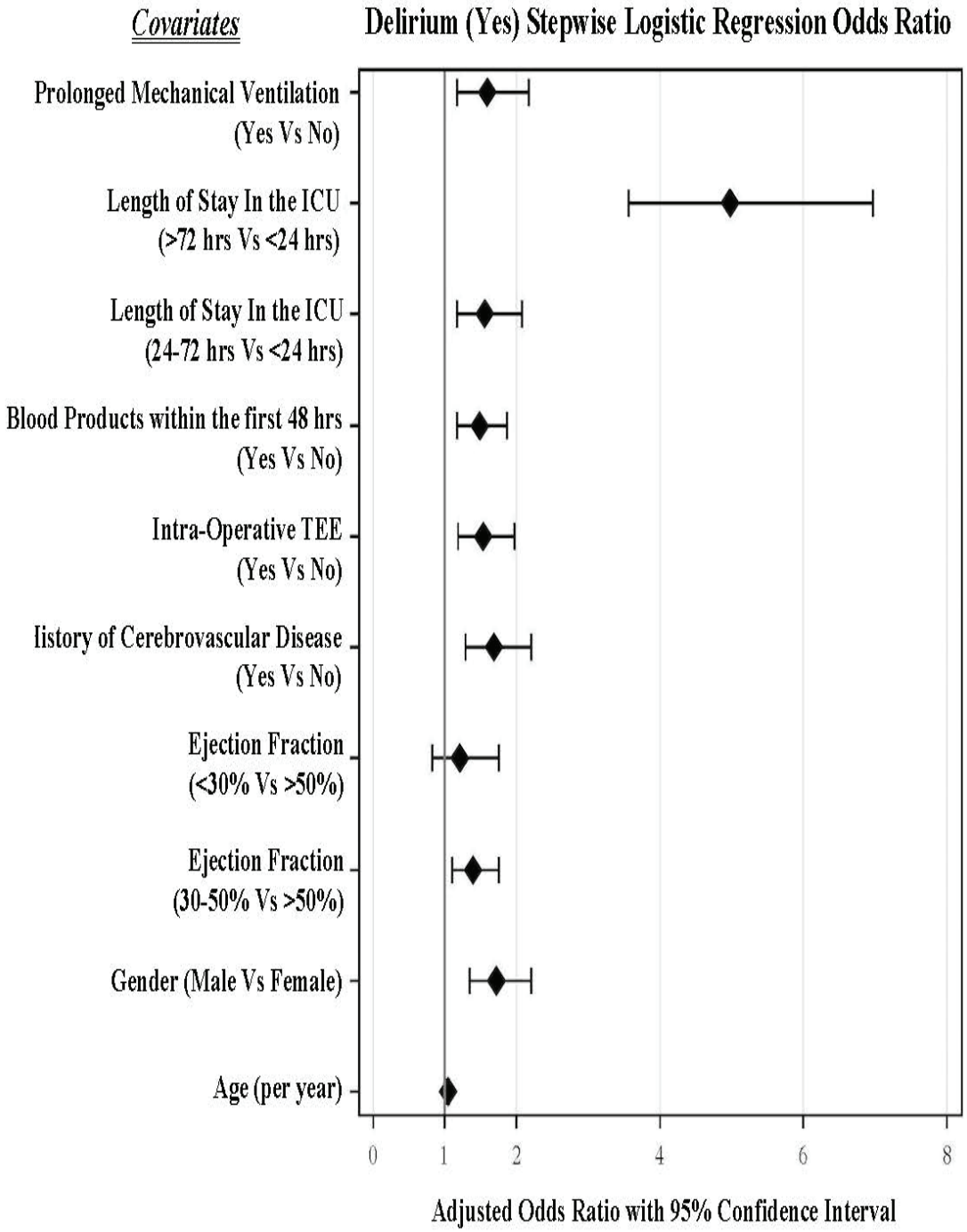

Figure 4-5: Odds Ratio Forest Plot of the Original Logistic Regression Model

Table 4-3: Experiment 1- Predictors of Post-operative Delirium in Multivariate Analysis

Parameter	Value	Coefficient	Standard Error	Odds Ratio (95% CI)	p-value
Intercept		-6.6924	0.4254		< 0.05
Age	Per year	0.0411	0.00555	1.042 (1.03-1.05)	< 0.05
Gender	Male	0.5445	0.1266	1.724 (1.35-2.2)	< 0.05
Ejection Fraction	<30%	0.1871	0.1911	1.2 (0.83-1.75)	0.3275
Ejection Fraction	30-50%	0.3299	0.1196	1.39 (1.1-1.76)	< 0.05
Cerebrovascular Disease	Yes	0.525	0.1361	1.69 (1.3-2.21)	< 0.05
Intra-Operative TEE	Yes	0.4244	0.1309	1.53 (1.18-1.98)	< 0.05
Blood Product transfusion within the first 48 hours	Yes	0.3968	0.1181	1.49 (1.18-1.87)	< 0.05
Length of stay in the ICU	24-72 hours	0.4447	0.1471	1.56 (1.17-2.08)	< 0.05
Length of stay in the ICU	>72 hours	1.6051	0.1714	4.98 (3.56-6.97)	< 0.05
Mechanical Ventilation >24 hours	Yes	0.4684	0.1573	1.6 (1.17-2.17)	< 0.05

The resulting model was validated using the test set, with an ROC-AUC of 77.7% indicating good discriminative power. The specificity (97.8%) and the negative predictive value (89.2%) of the model were good, whereas sensitivity (8.6%) and the positive predictive value (33%) were poor. The F1-score for the positive class (delirium=yes) was 13.8%. The Kappa statistic was 9.5%.

4.5.1.B. Experiment 1 – Artificial Neural Networks with 1 Hidden Layer

The same training dataset was used to generate an ANN model that included all the candidate attributes. WEKA 3.7 has a built an ANN that uses the back propagation algorithm and allows the user to adjust and manipulate several parameters when constructing the network. All the nodes in this algorithm have a built-in default sigmoid function, such that the output nodes for a task involving regression of a numeric outcome will be as linear units. However, this algorithm was selected here because of its

previously documented successful application to various problems in medicine [69, 144, 149, 150, 163-165].

Most experts advocate using a single hidden layer, as most classification problems can be solved this way, and increasing the number of hidden layers has the potential to over-train the network and lead to over-fitting. The learning rate (weight adjustment at each cycle) is usually between=0.1–0.3, and the momentum (how fast weight changes affect current weight changes, by smoothening the path and decreasing oscillations) is typically between=0–0.2[65, 124, 144, 148-150].

After several iterations, optimal settings using a single hidden layer were established (Figure C-19). The optimal number of nodes in the hidden layer was 3 with a learning rate of 0.275. The decay was set to false and momentum to "0" to allow us to identify the optimal structure without any regularization or smoothing. The *NominalToBinaryFilter* was set to false as the multinomial attributes were ordinal (e.g.: EF category and length of stay in the ICU) and turning off the filter is usually recommended[65]. Based on these settings, the training set was used to determine the network's structure and weight. The model's performance was then evaluated using the test dataset. The ROC-AUC was 76.7% with an F1-score=34.5%. These particular settings were used throughout the thesis for the generation of ANNs with 1 hidden layer models.

4.5.1.C. Experiment 1 – Artificial Neural Networks with 2 Hidden Layers

In this experiment, the effect of adding another hidden layer was explored. Although, adding a second layer for a binary class problem might not improve performance and will increase computational complexity, it might help clarifying the decision boundary. The same iterative process was applied in order to determine the optimal structure for a network with 2 hidden layers. After several iterations and adjustments, optimal settings were established (Figure C-20). In this network structure, performance was improved by dividing the ordinal attributes into binary. Based on these settings, the training set was used to obtain the network's structure and weight. These settings were used in the generation of ANNs with 2 hidden layers throughout this thesis. In this experiment, the

model's performance was evaluated using the test dataset. The ROC-AUC was 76.9% with an F1-score=37.2%.

4.5.1.D. Experiment 1 – Bayesian Belief Network with a Single Parent

The same training dataset was used to generate a simple BBN model that included all of our candidate attributes. For more information about the BBN setting options, please refer to APPENDIX C: MODELING C.VI.iv Bayesian Belief Network, page 176.

In this thesis, the default setting in WEKA was used, with the training set used to obtain the network's structure and probabilities (Figure 4-6). Then, the model performance was evaluated on the test dataset. The ROC-AUC was 75.7% with an F1-score=37%. This produces a model that is almost identical to a Naïve Bayesian model. This structure is very interesting, easy to understand and simple; it minimizes the complexity of the problem space.

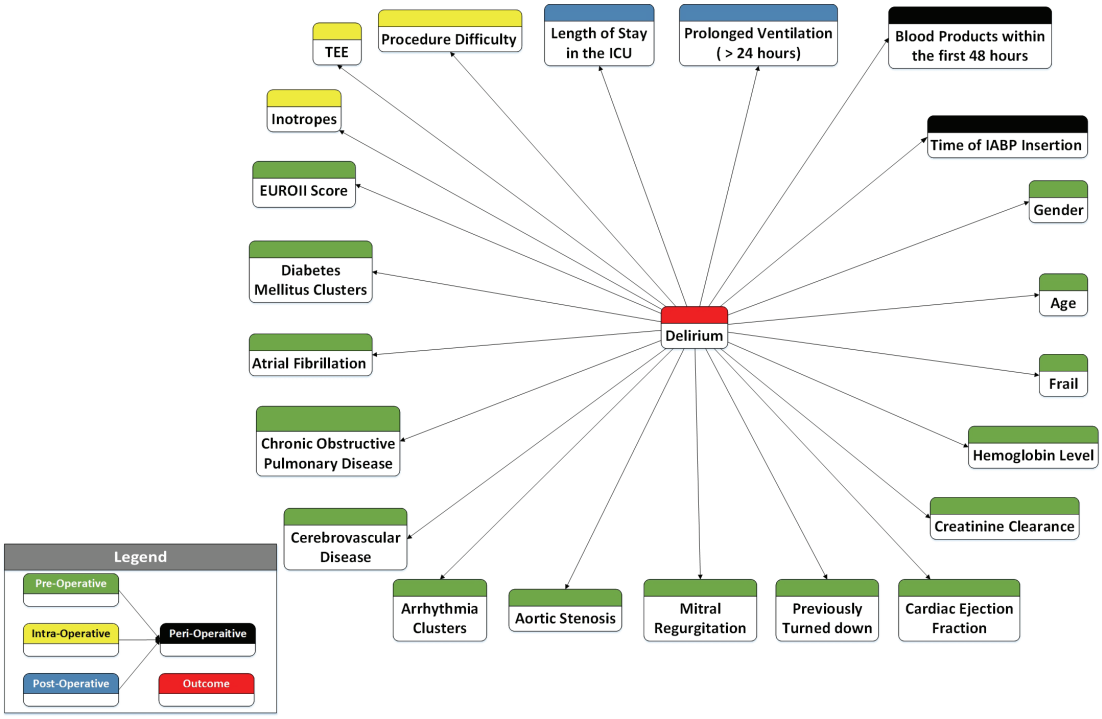

Figure 4-6: Experiment 1- BBN with 1 Parent

4.5.1.E. Experiment 1 – Bayesian Belief Network with 2 Parents

Based on the results obtained from the single parent BBN in the previous experiment, we explored implementation of a more complex structure using an advanced search algorithm. The LAGD (Look Ahead in a Good Direction) HillClimbing algorithm logically studies and chooses the best sequence, which will result in the highest score and a better network; instead of examining all possible moves at each step. LAGD Hill Climbing algorithm is a stochastic process that adds and deletes connections, disregarding the order of the attributes, but can look ahead a specified number of steps, examining several structures and choosing the best one. The WEKA version of LAGDHillClimbing algorithm first finds a subset of the best possible moves and then explores the graph subspace for the best sequence among them[166-168].

Setting the number of parents to a node in a BBN is a hard task, especially if the network structure is not known, and increasing the number is associated with increased complexity[169]. Several authors advocate restricting the number of parent nodes (in-degree), as it will decrease the computational expense and time[169, 170]. Unfortunately, there is no formal way to calculate the best number of parents, most of the time it is based on a combination of heuristics, trial and error, domain knowledge, and the best model[65, 67, 125, 170].

In this thesis, the LAGDHillClimbing algorithm was used and several experiments were conducted. The default setting in WEKA was used. The best results were obtained when the number of parents was restricted to 2, Figure C-23. Alpha was set to a smaller value (Alpha=0.1), to ease its effect in an effort to approximate the results to simulate maximum likelihood estimates[67, 125, 171].

The training set was used to obtain the network structure and probabilities. Then, the model performance was evaluated on the test dataset. The ROC-AUC was 76.4% with an F1-score=39.2%. Figure 4-7 is an illustration network topology using the above setting in WEKA.

In this representation, the BBN algorithm has discovered some interesting relationships between some attributes, and some of the attributes have no direct connection to delirium.

This displays the BBN's distinct ability of extracting important and interesting relationships directly from the data, and representing these relationships in a human interpretable and machine-readable fashion without the need of prior domain expertise or an input from the analyst.

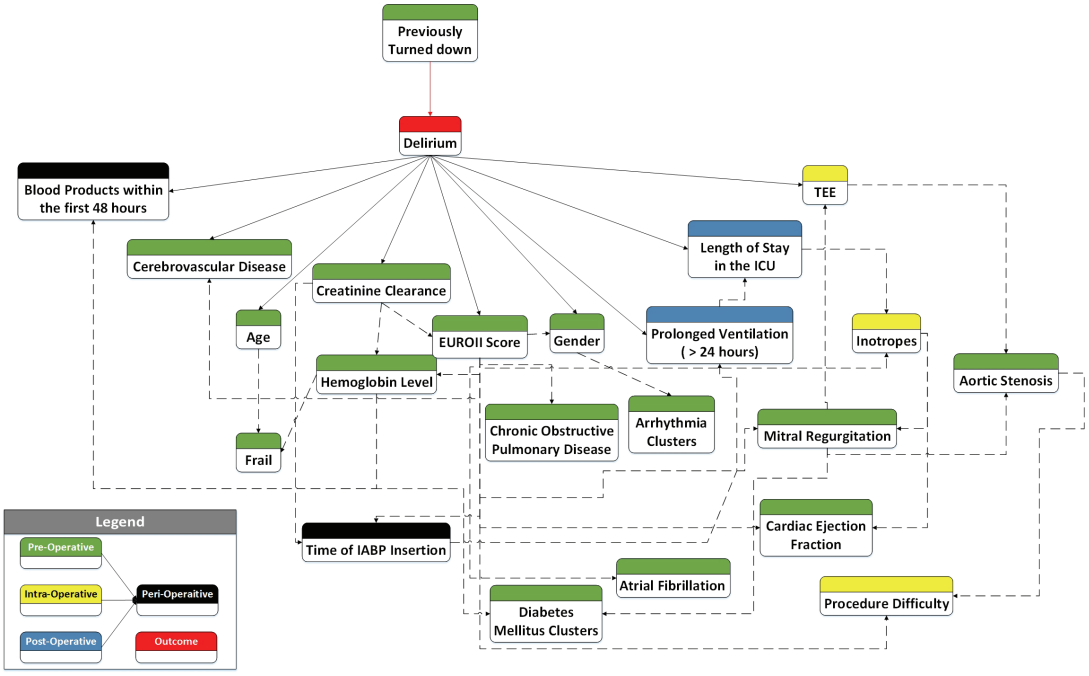

Figure 4-7: Experiment 1- BBN With 2 Parents
Solid black arrows: Direct causal relationship, Solid red arrow: Direct influential relationship, Dashed black arrows: Indirect causal relationship.

4.5.1.F. Summary of Experiment 1

The results from experiment 1 are summarized in Table 4-4. Figure 4-8 illustrates the ROC-AUC and the F1-score of all 5 models. The ROC-AUC was primarily used to assess the classifier's general performance. The primary goal is to determine the best classifier to accurately identify patients who will develop delirium. The F1-score for the positive cases was used, as it is the most reliable measure because it displays the worst-case scenario for the classifier.

Table 4-4: Summary of Experiment 1 Results

	ROC-AUC	Kappa	Recall	Specificity	Precision	F1-Score
LR	**77.7** †	9.5	8.6	**98.7** †	33.3	13.8
ANN (1 Hidden)	76.7	28.4	27.6	**95.9** ‡	**46.1** †	34.5
ANN (2 Hidden)	**76.9** ‡	**29.01** ‡	**37.8** ‡	91.6	36.6	**37.2** ‡
BBN (1 Parent)	75.7	26.5	**49.6** †	84.7	29.4	37
BBN (2 parents)	76.4	**31.8** †	37	93.3	**41.5** ‡	**39.2** †

**Measurements are in %*

† Best performing model, ‡ Second best performing Model

Logistic Regression had the best ROC-AUC and specificity, but was the worst in all other measures. BBN with 2 parents had the best F1 measure and Kappa statistics. The ANN with 2 hidden layers had the second best scores in 4 measures (ROC-AUC, Kappa, Recall, and F1-score).

Logistic regression uses maximum likelihood estimates by Fisher scoring (modified Newton-Raphson method) find the local minima in the sample space, by processing the data in a modified batch-gradient decent format (iterative maximum likelihood reweighted least squares algorithm)[53, 54, 172, 173]. In WEKA 3.7, ANN uses a modified form of Stochastic Gradient Descent (SGD) that is computationally more efficient, but still much slower than batch gradient descent (BGD). BBN in WEKA 3.7 uses a similar form of SGD but the difference here is that it searches for the optimal network topology before it starts looking for the local minima.

Graphically, Figure 4-9 illustrates that the ROC-AUC's for all of the models were very close to each other and overlapping (for more details about the ROC, please refer to APPENDIX C, page156). This indicates that there is no difference in the performance of any of the models.

The Hanley and McNeil repeated measures ROC test was applied. The correlation coefficient "r," using the Kendall tau, was =0.12-0.13. Alpha was set at =0.05 and we applied a 2-tailed test (null hypothesis: the 2 ROCs are not different). In both cases there was no statistical difference between any of the methods compared to LR. The general performance of all 5 models (discriminative power, which refers to the ability of a model to distinguish between positive cases from negative ones) was comparable (

Table 4-5). The McNemar's test was used to discern which model had a superior performance. In Table 4-6, BBN with 1 parent had a statistically significant worse performance compared to LR. ANN with 2 hidden layers was slightly worse than ANN with 1 hidden layer and BBN with 2 parents had an equal performance compared to LR, with ANN with 1 hidden layer making less mistakes and BBN with 2 parents making slightly more mistakes.

The choice of classifier will depend on its key role. In this case, all classifiers had an equivalent general performance. When another test was applied to evaluate their general performance (McNemar's test), only 3 had a comparable performance. Since the best measure to assess and generate the worst-case scenario for a classifier is the F1-score, BBN with 2 parents had the best performance. BBN with 2 parents other core performance measures (Kappa and precision) were also on the high side of the scale compared to the other models. In summary, the general performance of ANN with 1 hidden layer and BBN with 2 parents was equivalent to LR. The BBN with 2 parents showed the best core performance results (Table 4-4, Figure 4-8, Figure 4-9).

Table 4-5: Hanley and McNeil Repeated Measures ROC Test of Experiment 1 P-Values

	ANN (1 Hidden)	ANN (2 Hidden)	BBN (1 Parent)	BBN (2 parents)
Logistic Regression	0.38	0.41	0.28	0.35

**Two-tailed Test with an Alpha=0.05*

Table 4-6: McNemar's Test Results of Experiment 1

		Logistic Regression		McNemar's Test		Compared to Logistic Regression
		Success	Failed	Chi^2	p-value	
ANN (1 Hidden)	Success	947	37	0.23	0.63	=
	Failed	32	101			
ANN (2 Hidden)	Success	915	40	5.09	0.024	↓
	Failed	64	98			
BBN (1 Parent)	Success	850	52	31.9	<0.05	↓
	Failed	129	86			
BBN (2 parents)	Success	929	42	0.53	0.47	=
	Failed	50	96			

↓: Worse performance, ↑: Better Performance, =: Same performance

4.5.1.G. Experiment 1 Conclusion

The results from Experiment 1 have highlighted several interesting observations:

Relying on a single measure, the ROC-AUC, to evaluate a predictive model performance might falsely indicate that all 5 models have the same capabilities. When evaluating a predictive model, clinicians need to assess the performance from several standpoints using different measures (e.g.: F1-score and Kappa), to correctly determine its ability in accurately predicting the outcome of interest.

Based on the McNemar's test, ANN-1 hidden layer and BBN-2 parents had similar performance to LR. The ROC-AUC test indicated that all 5 models are equal. Although, looking at their performance from a different perspective, using the McNemar's test revealed that 3 out of the 5 models actually had a comparable performance.

Logistic regression was superior in identifying patients who will not develop delirium (high specificity) but had a very poor performance in distinguishing the patients that are at risk (low precision, recall). This is reflected in the low F1-score. This is most likely due to the imbalance class representation and the use of batch gradient decent that is influenced by the class distribution.

ANN-1 hidden layer and BBN-2 parents demonstrated a superior ability of distinguishing patients at risk of developing delirium without compromising their specificity, which is reflected in the high F1-score. A lesser effect of class distribution imbalance was noticed, most likely due to algorithms ability of accommodating and establishing non-linear relationships. The BBN-2 parents generated the best model, and it also had an elegant representation that can be symbolized and read by machines, yet comprehended by domain experts.

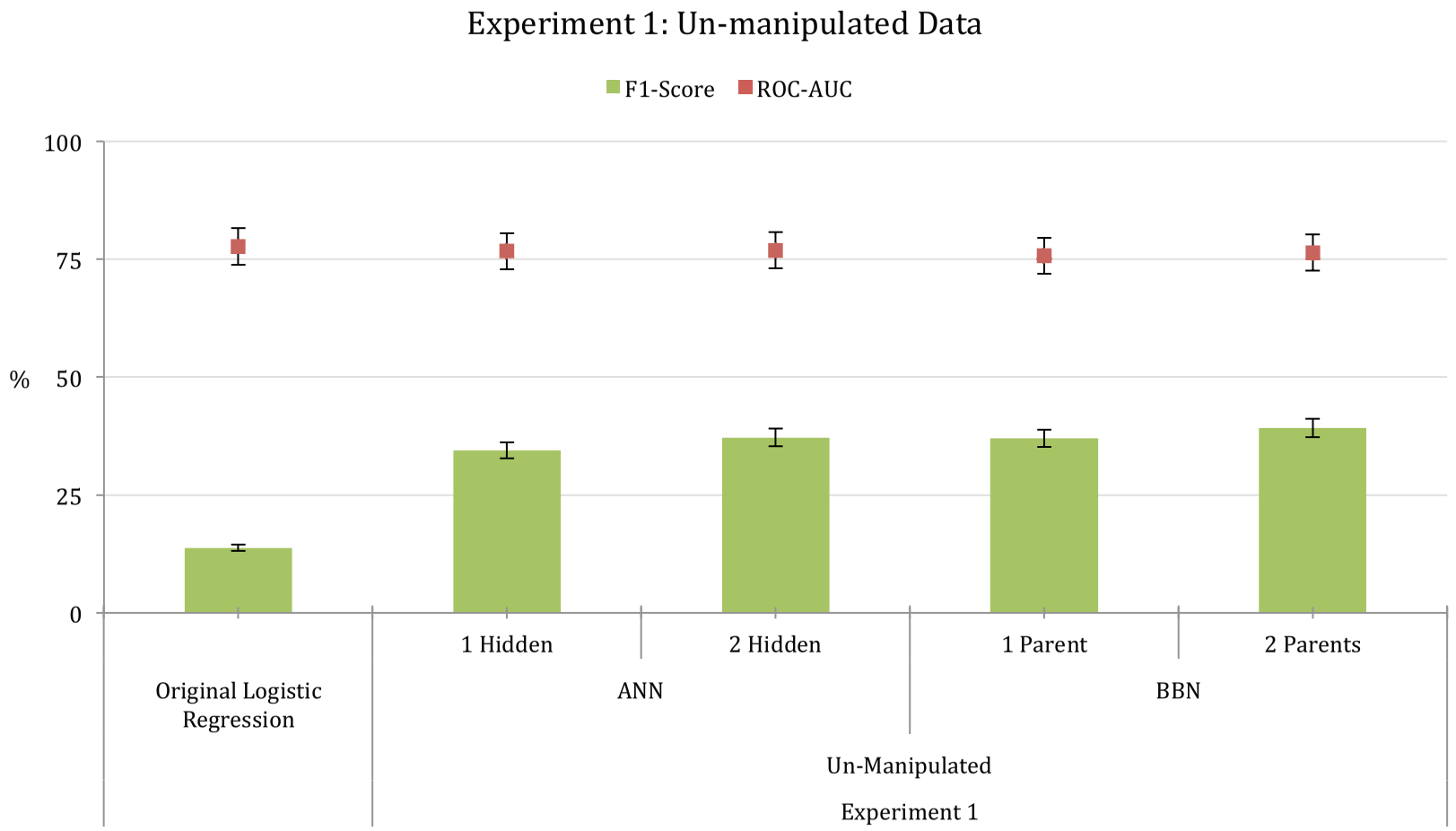

Figure 4-8: Experiment 1 results – Original training set. *(Error bars represent the 95% confidence intervals)*

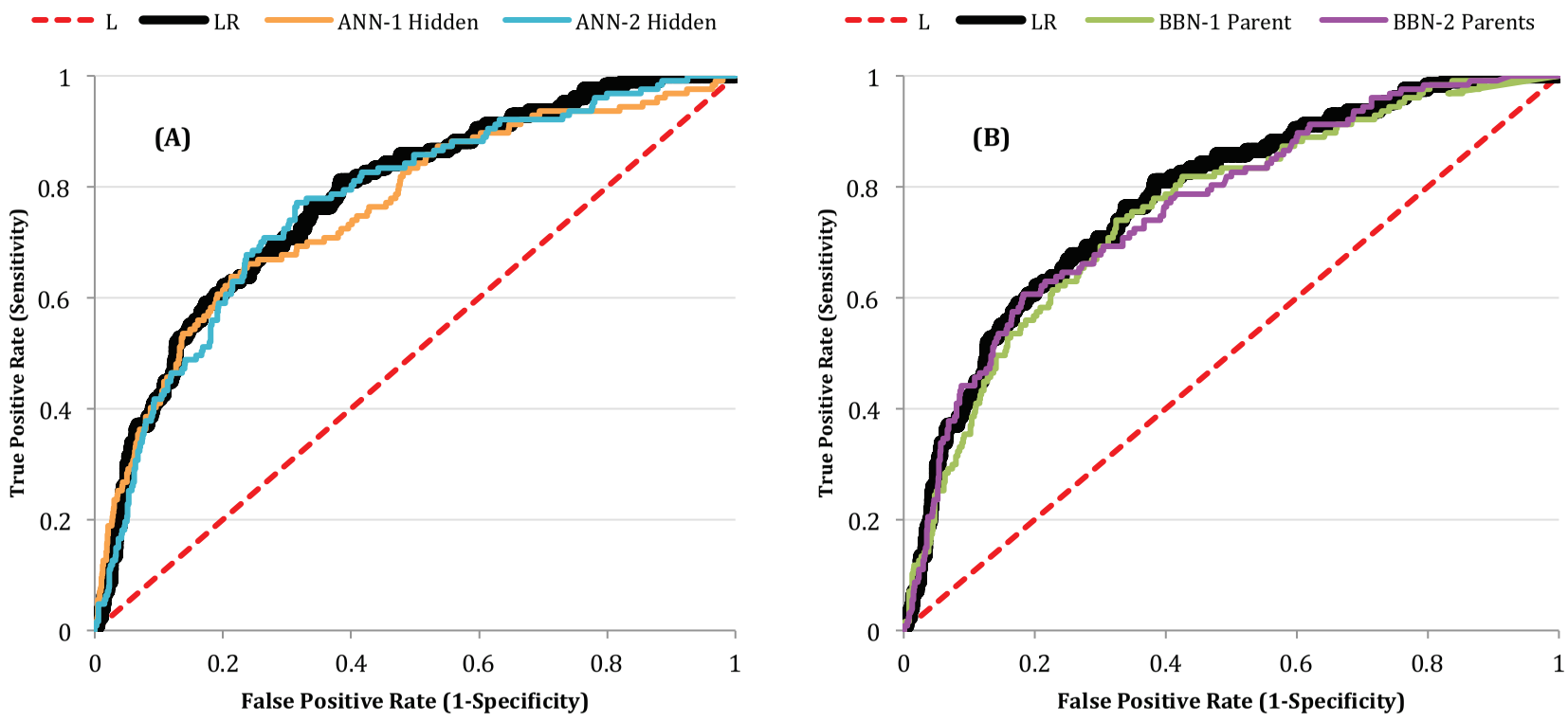

Figure 4-9: Receiver Operator Characteristics Curves
(A) ANN versus LR, (B) BBN versus LR
L: Line of zero discrimination, AUC=0.5.
(Models AUC: LR= 77.7%, ANN-1 Hidden= 76.7%, ANN-2 Hidden= 76.9%, BBN-1 Parent= 75.7%, BBN-2 Parents= 76.4%.)

4.5.2. Experiment 2: Synthetic Minority Over-sampling Technique

In this experiment, the impact of creating artificial observation on performance of the created models was explored. The models settings were kept the same, as described in Experiment 1. The random initialization of weights in ANN and probabilities in BBN might be different. The experiments were repeated with different attribute arrangement, and the results were consistent.

If the primary goal is to assess the effect of manipulating the training dataset sample size on several models, the recommended practice in the machine learning literature is to maintain the model parameters constant throughout the experiments, avoiding the inherent bias associated with adjusting the parameters[56-59, 65, 66, 70, 174-176]. If the primary goal is maximizing a particular classifier performance, adjusting the parameters is appropriate[56-59, 177, 178].

In this section, the main goal was to evaluate the impact of increasing the amount of data available for learning and expanding the algorithm sample search subspace, diverting its attention away from the majority class, on the developed models performance to examine the stability of the models on the same test data.

The SMOTE approach [58], which is available in WEKA under filters was used. Models were compared to the reference, un-manipulated "Alive" data LR model. For more information about SMOTE, please refer to page 162.

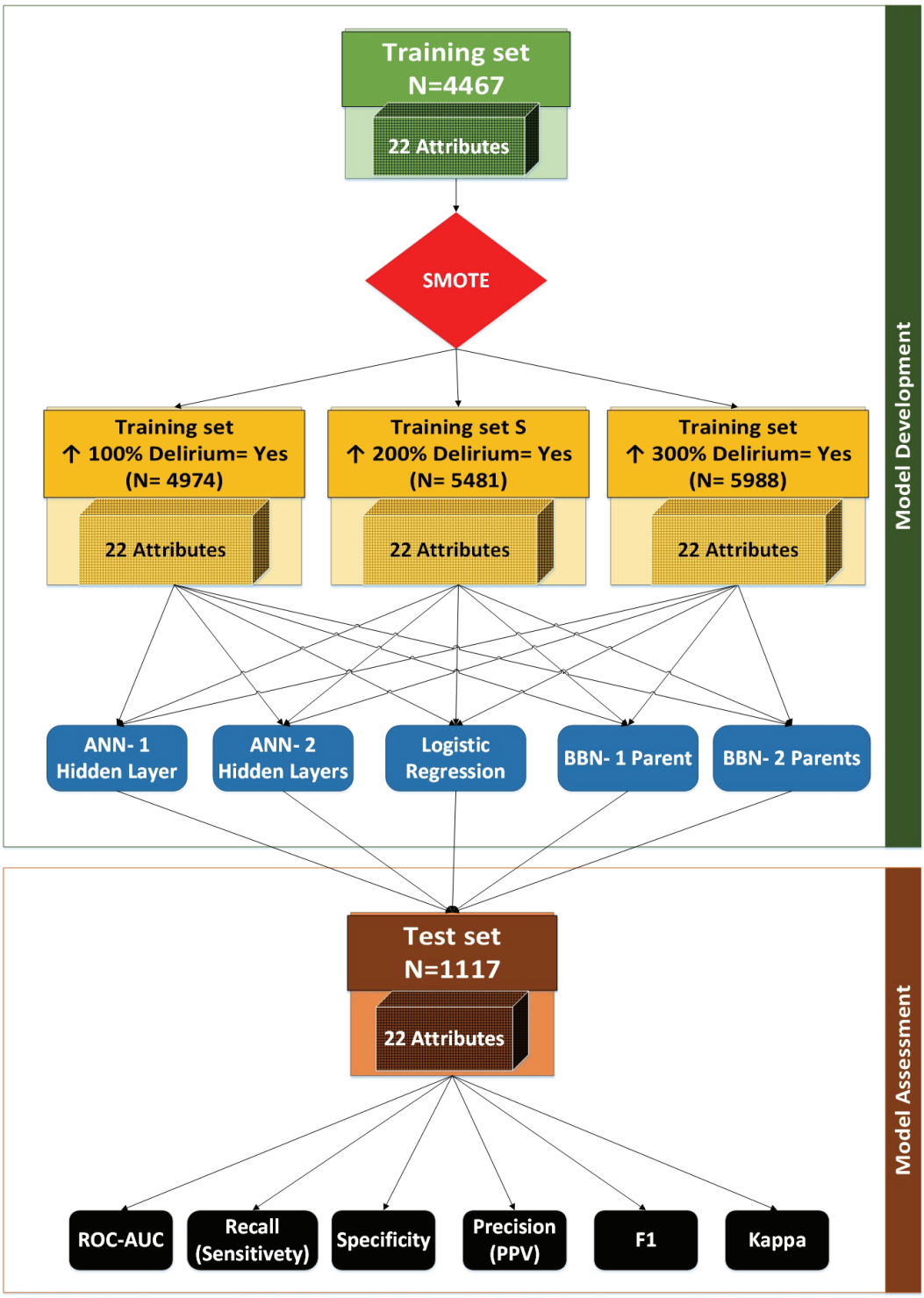

Figure 4-10: Data Level Manipulation with SMOTE – Experiment 2

In this work, the original setting of the algorithm in WEKA was used and over-sampling was done at 3 different levels: 100%, 200%, and 300%. Figure 4-10 represents a summary of experiment 2.The generated models had the exact setting that was used in Experiment 1: ANN with 1 hidden layer: Figure C-19, ANN with 2 hidden layers: Figure C-20, BBN with 1 parent: Figure C-21, Figure C-22, and BBN with 2 parents: Figure C-23.

4.5.2.A. Experiment 2 Results

At 3 different levels of over-sampling (100, 200, and 300%), SMOTE was applied , which generated 3 new training datasets. SMOTE with 100% generated a sample of 4,974 patients with an increase of the minority class from 507 to 1,014 cases (100%↑). SMOTE with 200% generated a sample of 5481 patients with an increase of the minority class from 507 to 1,521 cases (200%↑). SMOTE with 300% generated a sample of 5,988 patients with an increase of the minority class from 507 to 2,028 cases (300%↑). In all 3 datasets, the majority class observations number was constant at 3,960 cases. This produced an incremental increase of the minority class distribution from 11.4% in the original dataset to 34% in the SMOTE with 300% dataset (Figure 4-11).

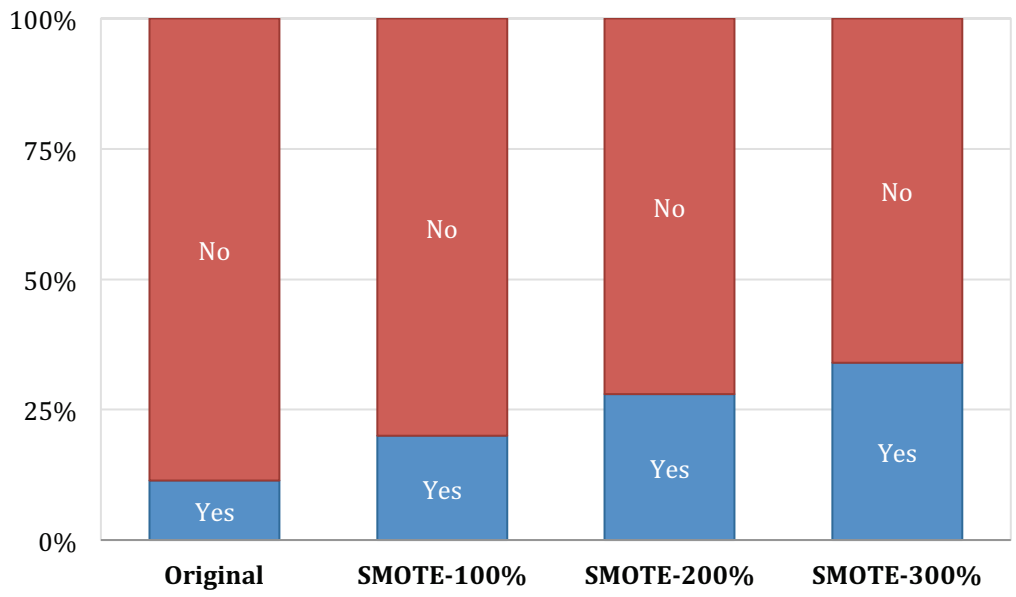

Figure 4-11: Outcome Class (Delirium) Distribution across SMOTE Datasets

Table 4-7, Figure 4-12, Figure 4-18, and Figure 4-19 illustrate a summary of Experiment 2 results with SMOTE. Compared to the performance of the original LR model; there was an incremental improvement, correlated with the sample size, in the LR model recall, precision, Kappa, and F1-score on the expense of ROC-AUC and specificity. The ANN models suffered a sizable decrease of performance, most likely due to the change in the optimal network for the present data. The BBN were the most resistant to change, although the BBN with 2 parents sustained more damage, most likely due to the change in the network structure.

Table 4-7: Summary of Experiment 2 Results-SMOTE

	ROC-AUC	Kappa	Recall	Specificity	Precision	F1-Score
Original LR	**77.7** †	9.5	8.6	**98.7** †	33.3	13.8
SMOTE 100%						
LR	73.6	23.5	27.6	**93.6** ‡	**35.7** †	31.1
ANN (1 Hidden)	72.9	13.5	19.7	92.5	25.3	22.1
ANN (2 Hidden)	67.7	15.7	20.5	93.3	28.3	23.7
BBN (1 Parent)	**75.4** ‡	**26.7** †	**52.8** †	83.6	29.3	**37.6** ‡
BBN (2 parents)	74.5	**24.7**	**36.2** ‡	90	31.7	**33.8**
SMOTE 200%						
LR	71.6	**24.6**	**33.9**	**91.1**	**32.8** ‡	**33.3**
ANN (1 Hidden)	71.6	12.6	22	90.4	22.8	22.4
ANN (2 Hidden)	70	16.8	23.6	92.1	27.8	25.5
BBN (1 Parent)	**75.2**	**25** ‡	**48.8** ‡	84	28.3	**35.8**
BBN (2 parents)	72.2	21.4	**33.9**	89.2	28.7	31
SMOTE 300%						
LR	70.5	**24.7**	37	89.6	**31.3**	**39** †
ANN (1 Hidden)	68.1	11.3	22.8	88.9	20.9	21.8
ANN (2 Hidden)	72.1	19.7	**34.6**	87.7	26.5	30
BBN (1 Parent)	**75.2**	**24**	**47.2**	84.1	27.6	**34.9**
BBN (2 parents)	72.3	20.6	30.7	**90.4**	29.1	29.9

**Measurements are in %*

† Best performing model, ‡ Second best performing Model

The addition of synthetic observations from the feature sub-space produced a noticeable improvement in the general performance (ROC-AUC) and F1-score of LR; secondary to the robustness and stability of the algorithm due to the uniform approach it takes to find the local minima in the sample space.

Figure 4-13 and Figure 4-14 demonstrates the effect of applying SMOTE on the original classifiers ROC-AUC and F1-scores, respectively. Both ANN classifiers sustained the most damage because they are dependent on learning rate, which might indicate that this is not the optimal network structure for the offered sample. Of note, the performance of ANN with 1 hidden layer got worse as the sample size increased. In the case of ANN with 2 hidden layers, as the sample size increased the performance improved. The BBN models were the most stable and had less profound changes. The F1-scores of ANN and BBN have improved; but compared to the original models, they are worse.

4.5.2.B. Conclusion of Experiment 2

Increasing the sample size with SMOTE has dramatically improved LR performance, but had an opposite effect on the other models. ANN sustained more damage compared to BBN that was more stable, most likely due to the change in the sample space, which lead to a drastic re-configuration of the optimal network.

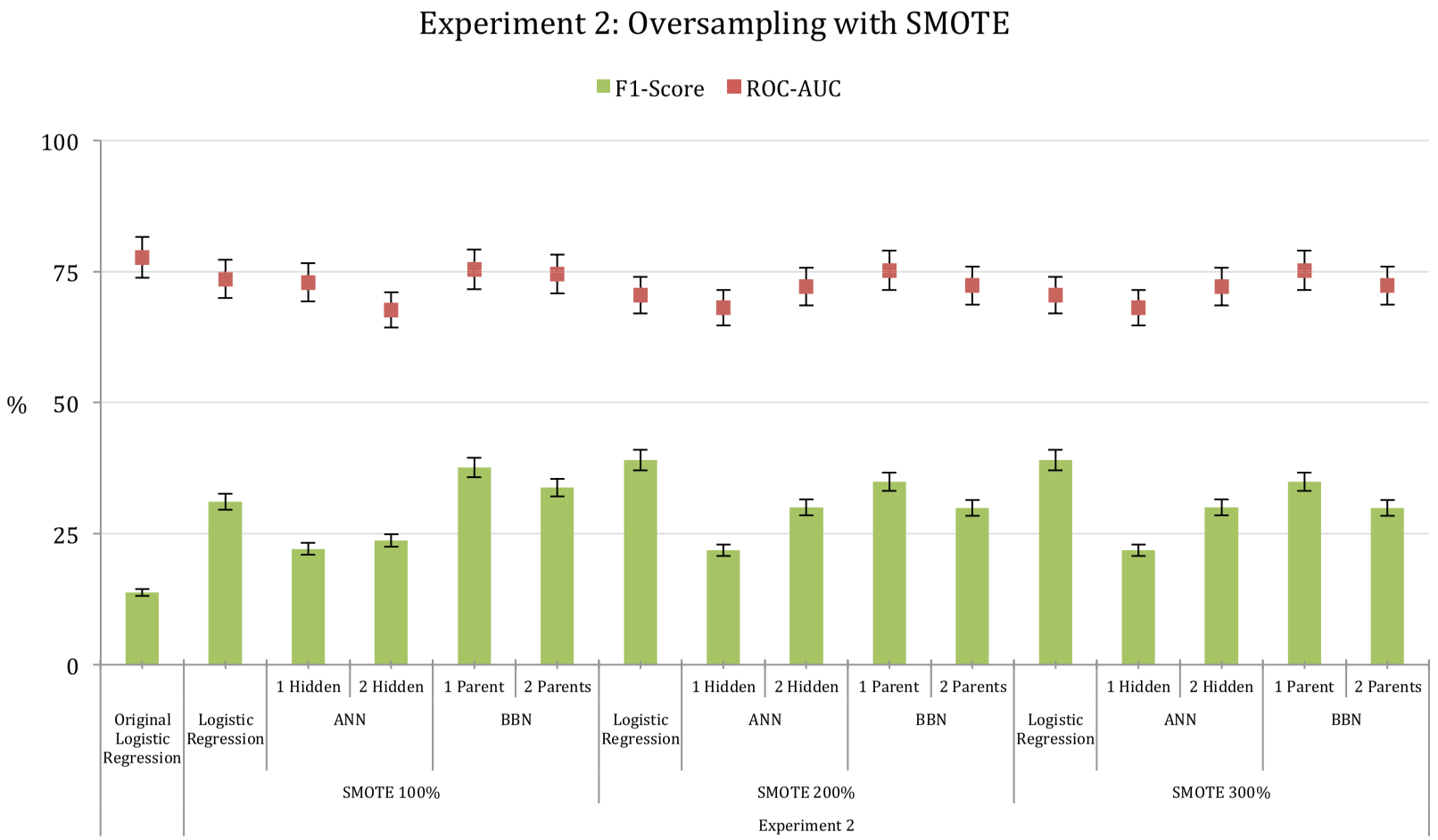

Figure 4-12: Experiment 2 Results – Over-sampling With SMOTE. *(Error bars represent 95% confidence intervals)*

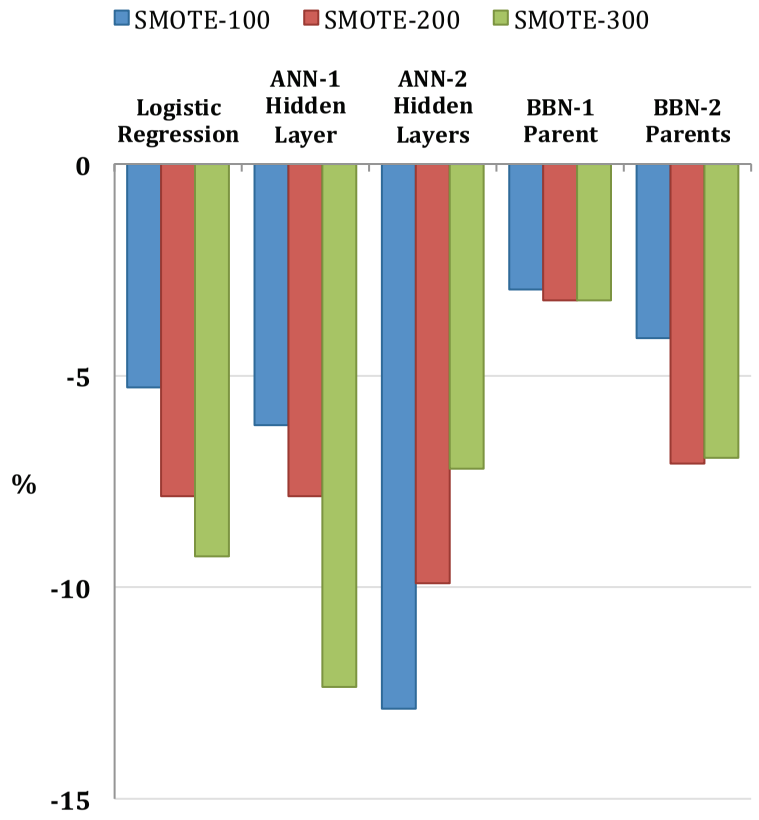

Figure 4-13: Experiment 2 ROC-AUC Compared to the Original Logistic Regression
*(Original LR ROC-AUC=77.7%)

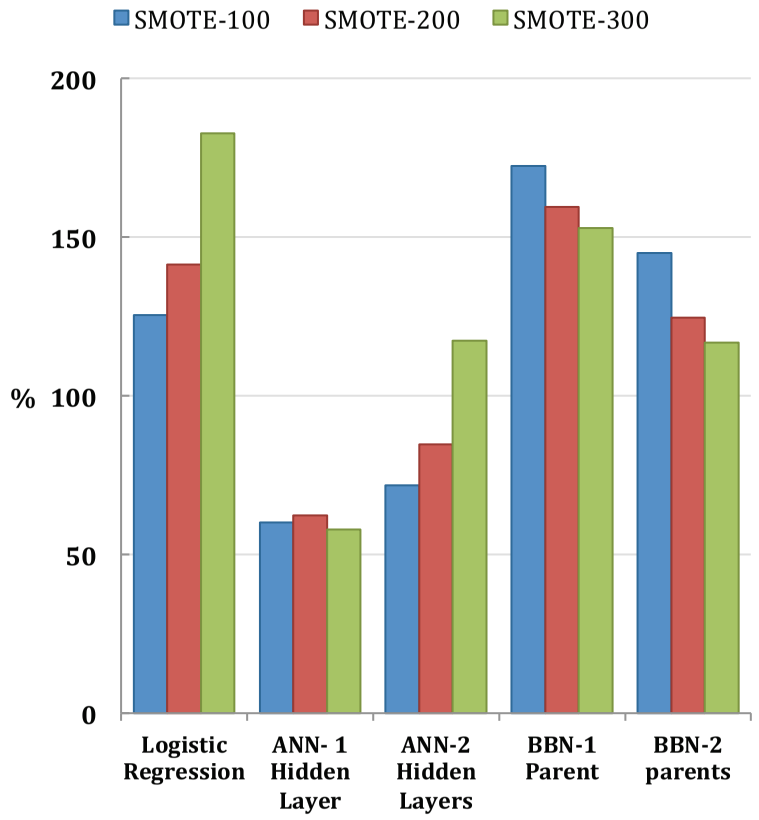

Figure 4-14: Experiment 2 F1-Measure Compared to the Original Logistic Regression
*(Original LR F1-Measure=13.8%)

4.5.3. Experiment 3: Spread Sub-sample

In this experiment, the impact of under-sampling on performance of the created models will be explored. The models settings were kept the same, as described in Experiment 1. The random initialization of weights in ANN and probabilities in BBN might be different. However, the experiments were repeated with different attribute arrangement and the results were consistent.

As the primary goal was to evaluate the effect of training data sample manipulation on the model development, the recommended practice in the machine learning literature is to maintain the model parameters constant throughout the experiments and to avoid the inherent bias associated with adjusting the parameters[56-59, 65, 66, 70, 174-176].

In this section, the main goal was to evaluate the impact of decreasing the amount of data available for learning and restricting the search subspace, directing it toward the minority class, on the developed models performance to examine the stability of the models on a consistent test data.

In WEKA, the SpreadSubSample option under the supervised instance filter was used. This will produce a random sub-sample of the original dataset. This filter allows the analyst to specify the maximum “spread"” between the minority and majority class (difference in class frequencies). The original settings of the algorithm in WEKA were maintained and the distribution of spread (majority : minority) was manipulated at 3 different levels: 1:1, 1.5:1, and 2.5:1. Figure 4-15 represents a summary of Experiment 3. The generated models had the exact setting that was used in Experiment 1: ANN with 1 Hidden layer: Figure C-19, ANN with 2 Hidden layers: Figure C-20, BBN with 1 Parent: Figure C-21, Figure C-22, BBN with 2 Parents: Figure C-23.

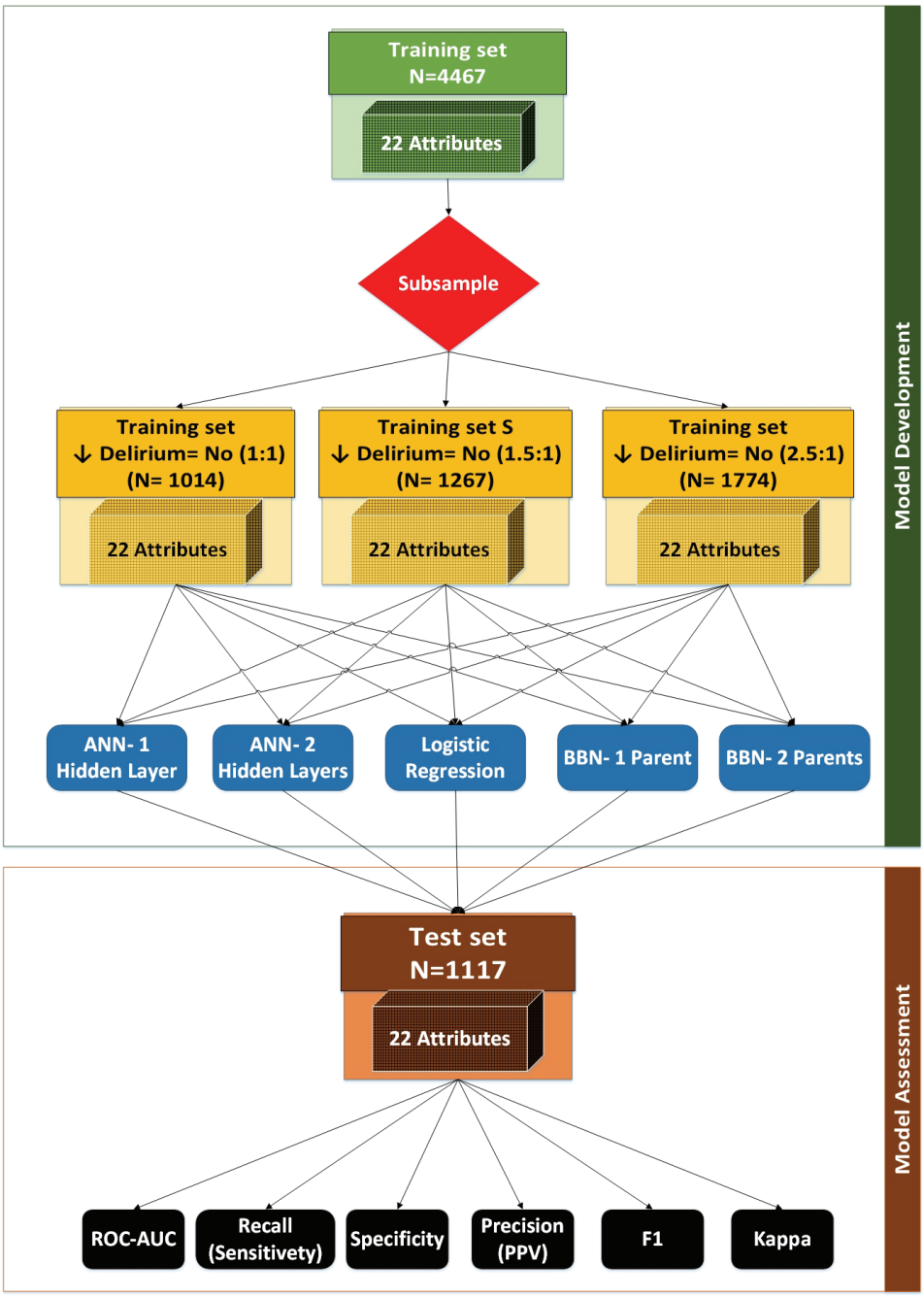

Figure 4-15: Data Level Manipulation with Spread Sub-sampling – Experiment 3

4.5.3.A. Experiment 3 Results

Spread sub-sampling was applied at 3 different ratios of majority: minority class distribution (1:1, 1.5:1, and 2.5:1), which generated 3 new training datasets. Sub-sample at 1:1 created a sample of 1,014 patients with 507 positive cases (minority) and 507 negative cases (majority). Sub-sample at 1.5:1 created a sample of 1,267 patients with 507 positive cases (minority) and 760 negative cases (majority). Sub-sample at 2.5:1 created a sample of 1,774 patients with 507 positive cases (minority) and 1,267 negative cases (majority). In all 3 datasets, the minority class was constant, 507 cases. This lead to a steady decrease in the minority class distribution from 50% in the sub-sample 1:1 dataset to 29% in the sub-sample-2.5:1 dataset (Figure 4-16).

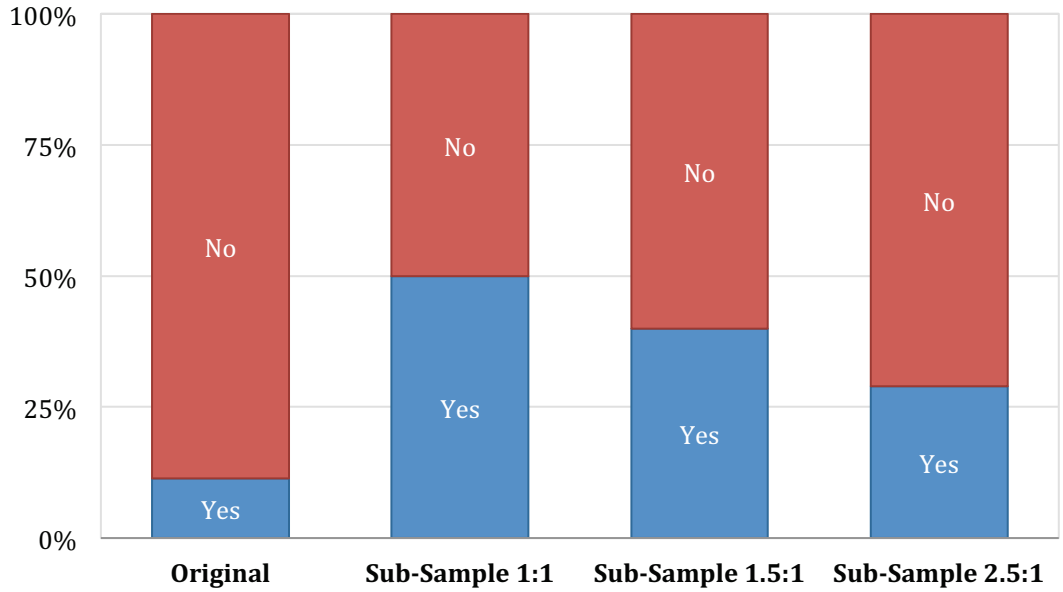

Figure 4-16: Outcome Class (Delirium) Distribution across Sub-sample Datasets

Table 4-8, Figure 4-17, Figure 4-18, and Figure 4-19 represent a summary of Experiment 3 results with spread sub-sampling. Compared to the original LR model performance; reducing the sample space did not affect the general performance of the LR (ROC-AUC) but improved recall, precision, Kappa, and F1-score on the expense specificity. The LR model with a 2.5:1 ratio of sub-sampling has produced the highest F1-score so far (40.5%), with very reasonable performance in other measures.

The use of sub-sampling had some moderate negative effect on the ANN with 2 hidden layers, most likely due to the change in the optimal network structure. The other models had minor decline in their performance.

Table 4-8: Summary of Experiment 3 Results – Spread Sub-sample

	ROC-AUC	Kappa	Recall	Specificity	Precision	F1-Score
Original LR	**77.7**	9.5	8.6	**98.7** *†*	33.3	13.8
Spread Sub-sample 1:1						
LR	**76.8**	**22.8**	63.8	**74.8**	**24.5**	**35.4**
ANN (1 Hidden)	75.3	16.7	73.2	62.7	20.1	31.6
ANN (2 Hidden)	74.1	20.2	**74** *†*	66.8	22.2	34.2
BBN (1 Parent)	75.8	21.2	**70.1** *‡*	70	23.1	**34.7**
BBN (2 parents)	74.9	19.4	68.5	68.8	22	33.3
Spread Sub-sample 1.5:1						
ANN (1 Hidden)	76.7	**26**	55.1	**81.8**	**28** *‡*	**37.1**
ANN (2 Hidden)	75.6	**23.1**	**65.4**	74.3	24.6	**35.8**
BBN (1 Parent)	74.9	20.7	**66.9**	71.1	22.9	34.1
BBN (2 parents)	75.2	22.2	63.8	74.3	24.1	35.1
Spread Sub-sample 2.5:1						
ANN (1 Hidden)	**77.1** *‡*	**31.4***†*	48.8	**88.2** *‡*	**34.6** *†*	**40.5** *†*
ANN (2 Hidden)	76.3	**27.5***‡*	**62.2**	79.7	28.2	**38.8** *‡*
BBN (1 Parent)	70.6	22.6	55.1	79.1	25.3	34.7
BBN (2 parents)	75.7	23.7	**59.8**	77.7	25.6	35.8

**Measurements are in %*

† Best performing model, ‡ Second best performing Model

4.5.3.C. Conclusion of Experiment 3

The use of Spread Sub-sample has dramatically improved LR performance but had a minor effect on the other models. The improvement in LR might be due to the maximization of the provided information in the sub-sampled dataset with more focused class boundaries. The minor to moderate decline of the other models performance might be due to the sub-optimal parameters setting of the networks. If the network parameters were optimized, the performance of these algorithms would have been better. But in this case, it will create a new network structure that is learned in a different setting, and comparing their performance will be challenging.

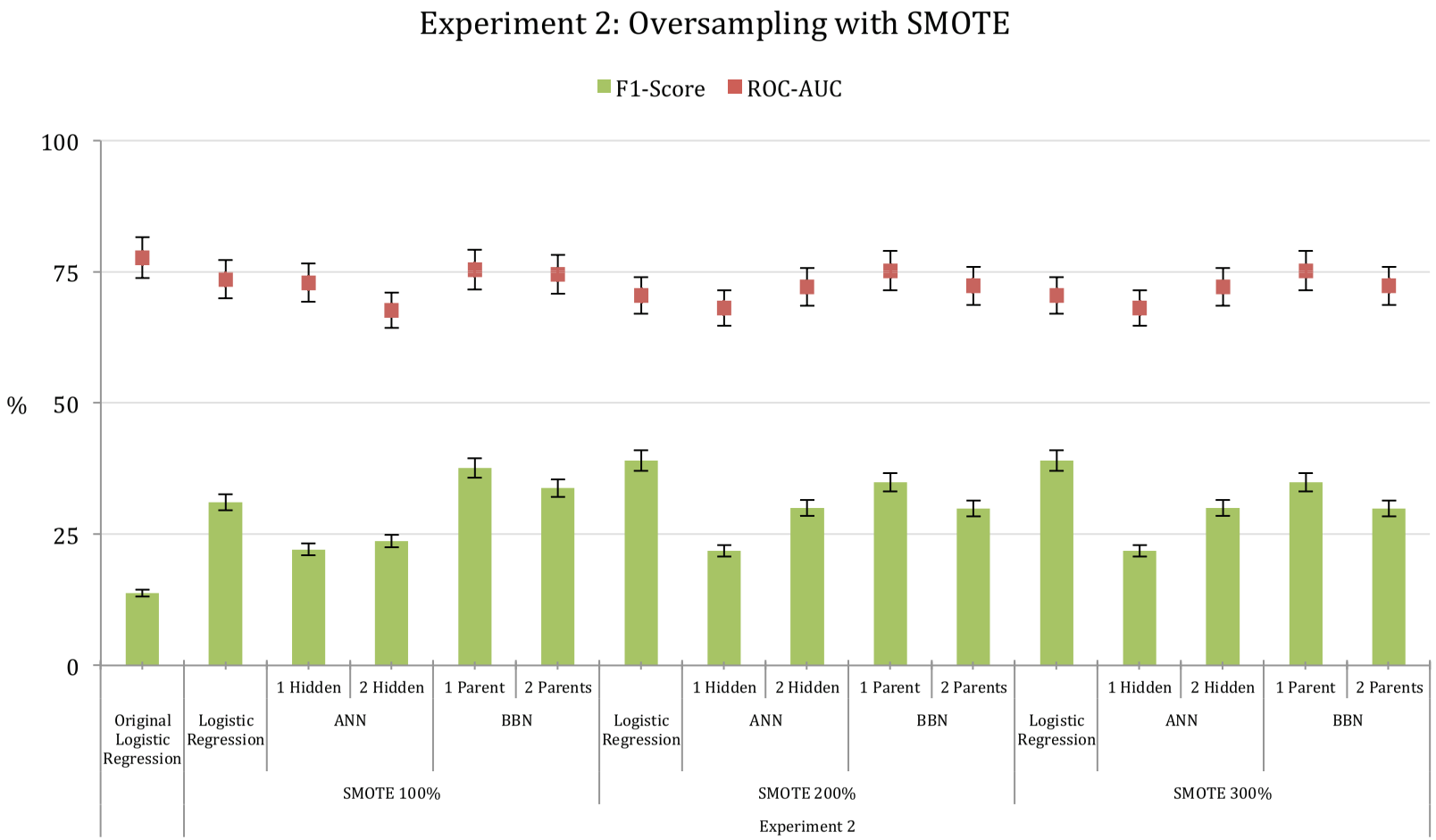

Figure 4-13: Experiment 2 Results – Over-sampling With SMOTE. *(Error bars represent 95% confidence intervals)*

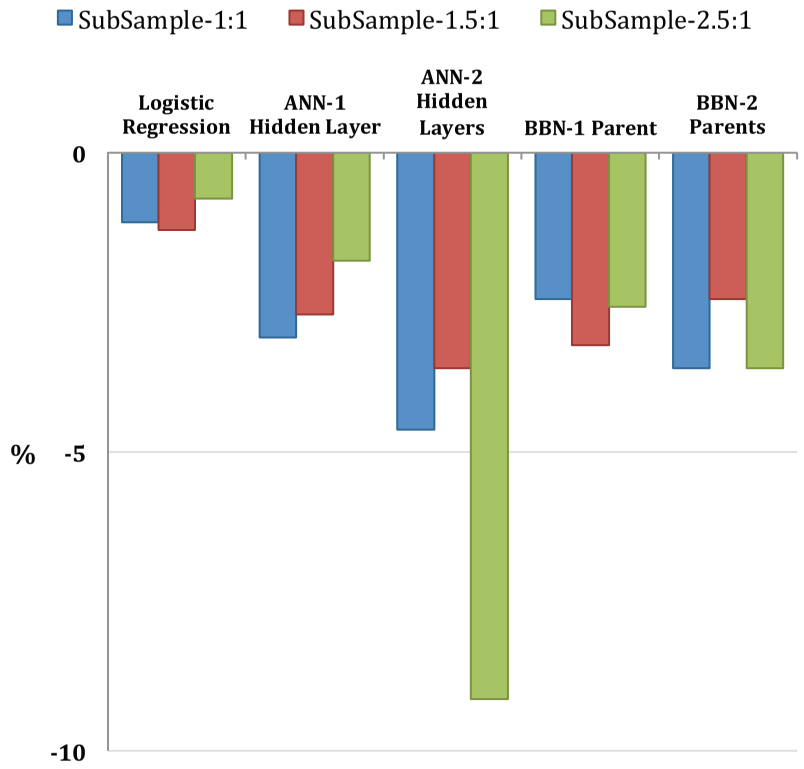

Figure 4-18: Experiment 3 ROC-AUC Compared to the Original Logistic Regression
*(Original LR ROC-AUC=77.7%)

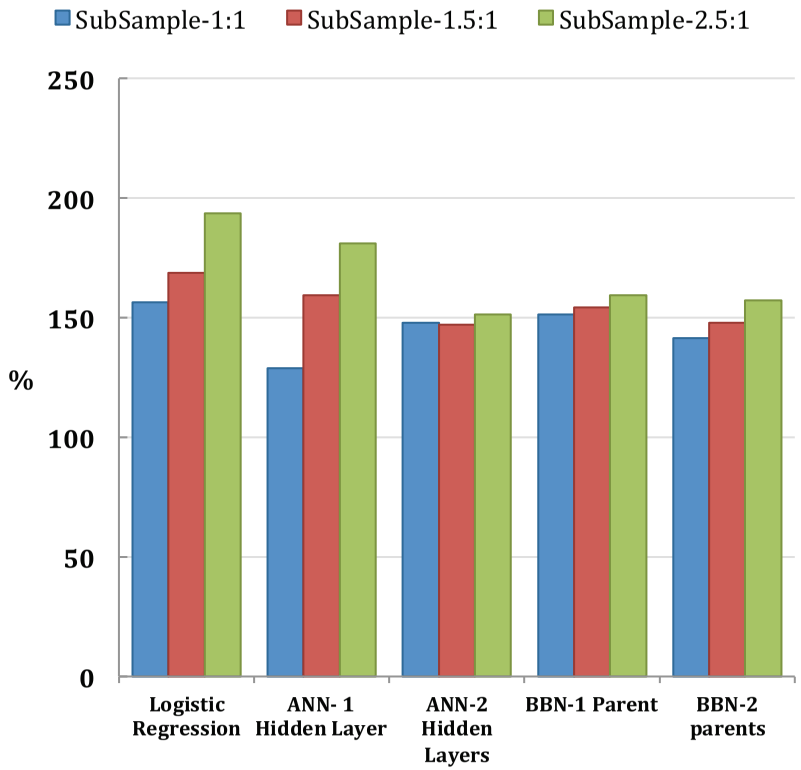

Figure 4-19: Experiment 3 F1-Score Compared to the Original Logistic Regression
*(Original LR F1-score=13.8%)

4.5.4. Experiment 4: Applying Cost

In this section, the main goal was to evaluate the impact of attaching a cost to the classification task, advising the classifier of the uneven ramification of each class, and rewarding it for choosing the minority class correctly, evaluating the models performance on the test data.

Since the price misclassification is not equal, cost is not constant and attaching a cost to a medical complication is hard. Also, delirium is a multifactorial process that can be initiated through different triggers. Identifying a single cost that can be attached to post-operative delirium after cardiac surgery requires a full understanding of the process that leads to its development. Unfortunately, the triggers of post-operative delirium are not clear yet.

In this experiment, cost sensitive classification was applied during the development of the models. In this case, the development of the models involved imposing a penalty/reward of choosing the wrong/right answer, pressing the algorithm to pick wisely. The models settings were kept the same, as described in Experiment 1. The random initialization of weights in ANN and probabilities in BBN might be different. Although, the experiments were repeated with different attribute arrangement, the results were the same.

If several candidate classifiers exist and choosing one is the task, the recommended practice in the machine learning literature is to maintain the model parameters constant throughout the experiments and to avoid the inherent bias associated with adjusting the parameters[56-59, 65, 66, 69, 70, 174, 176]. In reality, changing the cost is a less expensive task than re-developing the classifier from scratch and requires minimal change to the model infrastructure[56-59, 65, 66, 69, 70, 159, 174-177, 179].

In WEKA, applying cost to a classifier can be done in the meta-classifier tab. The classifier is called “CostSensitiveClassifier” that makes its base classifier cost-sensitive. Two methods can be used to introduce cost-sensitivity: first, reweighting training instances according to the total cost assigned to each class; or second, predicting the class with minimum expected misclassification cost (rather than the most likely class). In this work we used the reweighting technique.

Applying the reweighting technique involves deep understanding of the domain and some sense of the attached cost to the misclassification. In this work, several cost matrices were applied with the goal to maximize the F1-score. Most classification algorithms' main goal is to minimize the error rate. They assume that the distribution of categories among the outcome class is balanced and that all misclassification errors have an equal effect (cost)[57]. In real life applications, these assumptions are not true.

Figure 4-20 demonstrates the setting of 3 different cost matrices that were used. The classifiers were rewarded for correctly labeling the minority class. For example, looking at cost matrix 4, when correctly identifying a patient with delirium it, the classifier receives 4 points, but when it correctly identifies a patient who did not develop delirium, it only receives 1 point. By adjusting the reward/penalty (Cost) matrix, it will be necessary for the algorithm to balance the probabilistic estimates or decision thresholds, directing the algorithm attention towards the class of interest, the minority class.

Cost Matrix 1.5		
Predicted→	Yes	No
Class (Yes)	1.5	0
Class (No)	0	1

Cost Matrix 4		
Predicted→	Yes	No
Class (Yes)	4	0
Class (No)	0	1

Cost Matrix 6.5		
Predicted→	Yes	No
Class (Yes)	6.5	0
Class (No)	0	1

Figure 4-20: Experiment 4 Cost Matrices

The cost sensitive classification approach that was used in experiment 4 is summarized in Figure 4-22. In these experiments, the same training dataset from Experiment 1 was used. No manipulation was applied to any of the data sets. The proportion of delirium (positive class) was 11.4%. Models had the exact setting that was used in Experiment 1: ANN with 1 hidden layer: Figure C-19, ANN with 2 Hidden layers: Figure C-20, BBN with 1 Parent: Figure C-21, Figure C-22, and BBN with 2 Parents: Figure C-23.

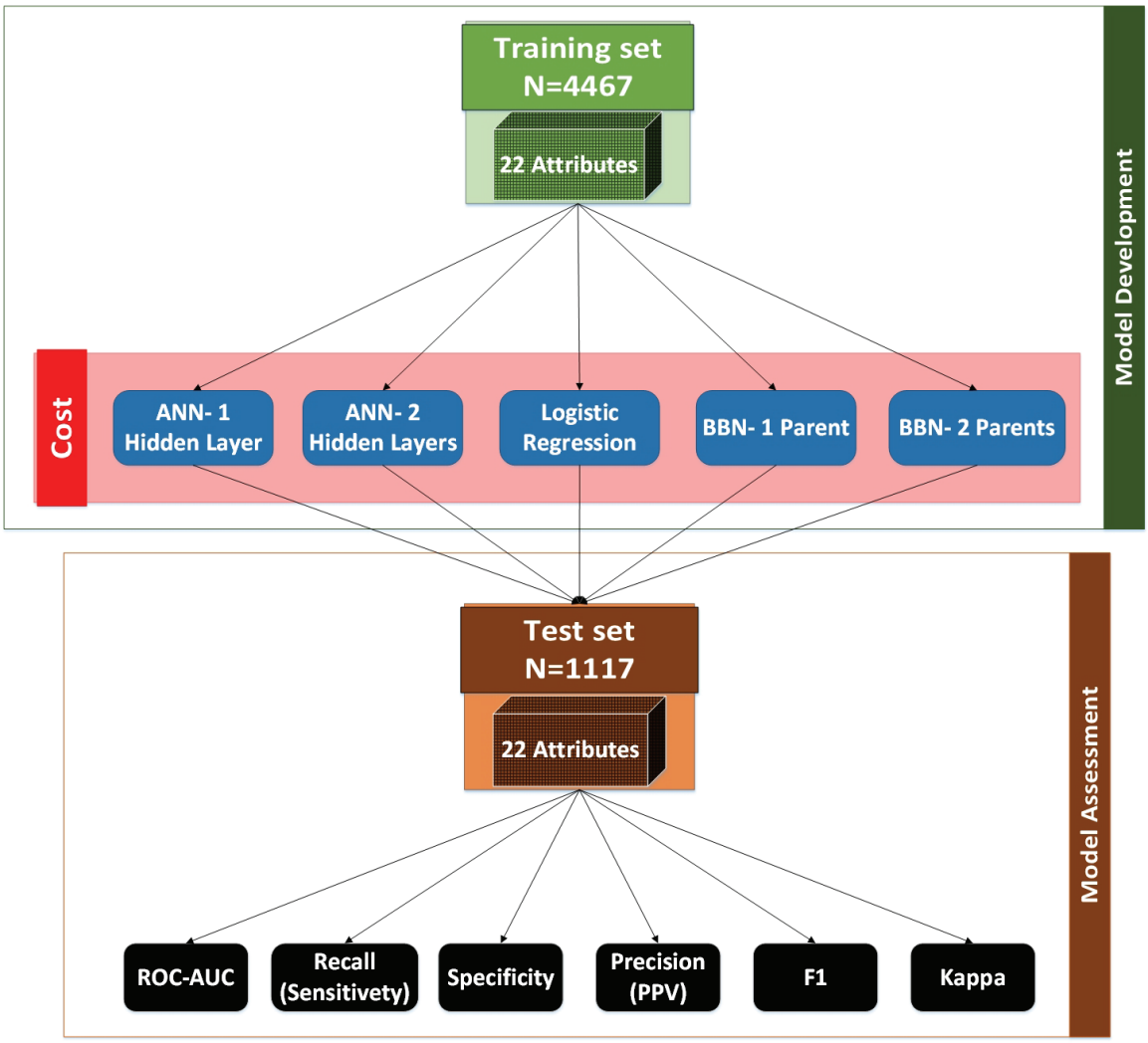

Figure 4-21: Algorithm Level Manipulation by Applying Cost– Experiment 4

4.5.4.A. Experiment 4 results

Three different ratios of minority : majority class recognition costs were applied (1.5:1, 4:1, and 6.5:1). 507 cases belonged to the minority class and 990 cases were majority, Table 4-9 and Figure 4-23 represent a summary of Experiment 4 results with cost sensitive classification applied. Compared to the performance of the original LR model, applying cost did not affect the general performance of most of the models (ROC-AUC) but improved recall, precision, Kappa, and F1-score with no major effect on specificity (Figure 4-23 and Figure 4-24). The ANN with 2 hidden layers with a cost ratio of 4:1 has produced the highest F1-score (41.8%), with second best Kappa, sensitivity, and precision.

Table 4-9: Summary of Experiment 4 Results – Cost Sensitive Classification

	ROC-AUC	Kappa	Recall	Specificity	Precision	F1-Score
Original LR	**77.7**	9.5	8.6	**98.7** †	33.3	13.8
Cost 1.5:1						
LR	**77.9** ‡	26.4	26.8	**95.4** ‡	**42.5** †	32.9
ANN (1 Hidden)	74.4	26.1	34.6	91.5	34.4	34.5
ANN (2 Hidden)	71.4	22.4	29.1	92.2	32.5	30.7
BBN (1 Parent)	75.7	**27.1**	**54.3**	83	29.1	**37.9**
BBN (2 parents)	76.5	**32.6** †	**41.7**	91.2	**39.3** ‡	**40.5** ‡
Cost 4:1						
ANN (1 Hidden)	**78**	**30.5**	51.2	**86.7**	**33**	**40.1**
ANN (2 Hidden)	77.1	22.4	**62.9**	74.8	24.3	35.1
BBN (1 Parent)	75.9	**31.3** ‡	**64.6** ‡	81.5	30.9	**41.8** †
BBN (2 parents)	75.5	22.2	58.3	77.2	24.7	34.7
Cost 6.5:1						
ANN (1 Hidden)	**78**	**26.6**	61.4	**79.4**	**27.6**	**38.1**
ANN (2 Hidden)	**78.5** †	**26.4**	**66.1** †	77	26.9	**38.3**
BBN (1 Parent)	75.2	25.5	59.8	79.2	27	37.2
BBN (2 parents)	75.8	22.4	63	74.8	24.3	35.1

**Measurements are in %*
† Best performing model, ‡ Second best performing Model

4.5.4.B. Conclusion of Experiment 4

Applying cost sensitive classification has generally improved core performance without compromising general measures (Table 4-9 and Figure 4-23). ANN with 2 layers has demonstrated the best results. The BBN models were robust and resistant to manipulation throughout the experiments, specifically BBN with 2 parents (Figure 4-23 and Figure 4-24).

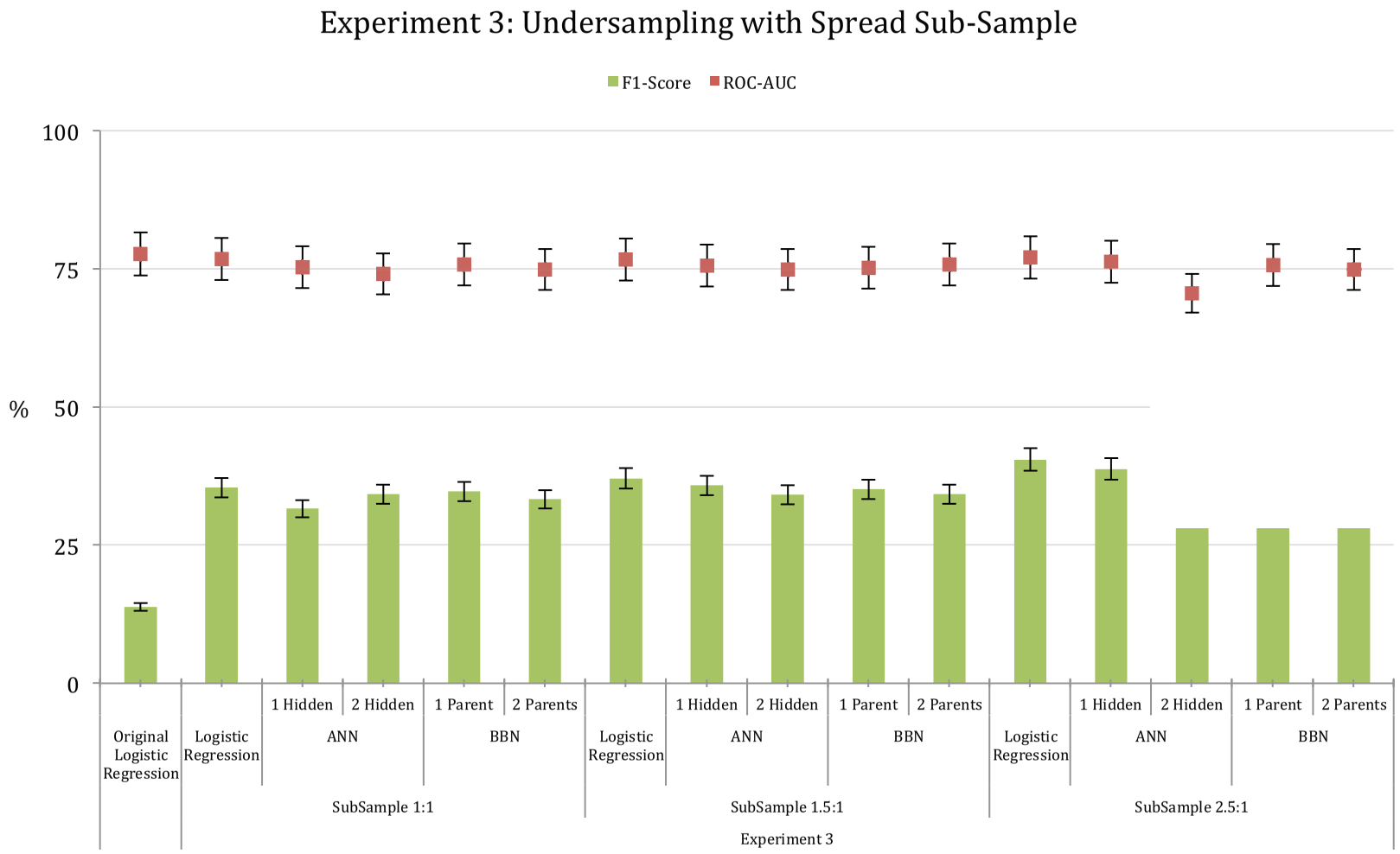

Figure 4-18: Experiment 3 Results – Spread Sub-sample. *(Error bars represent the 95% confidence intervals)*

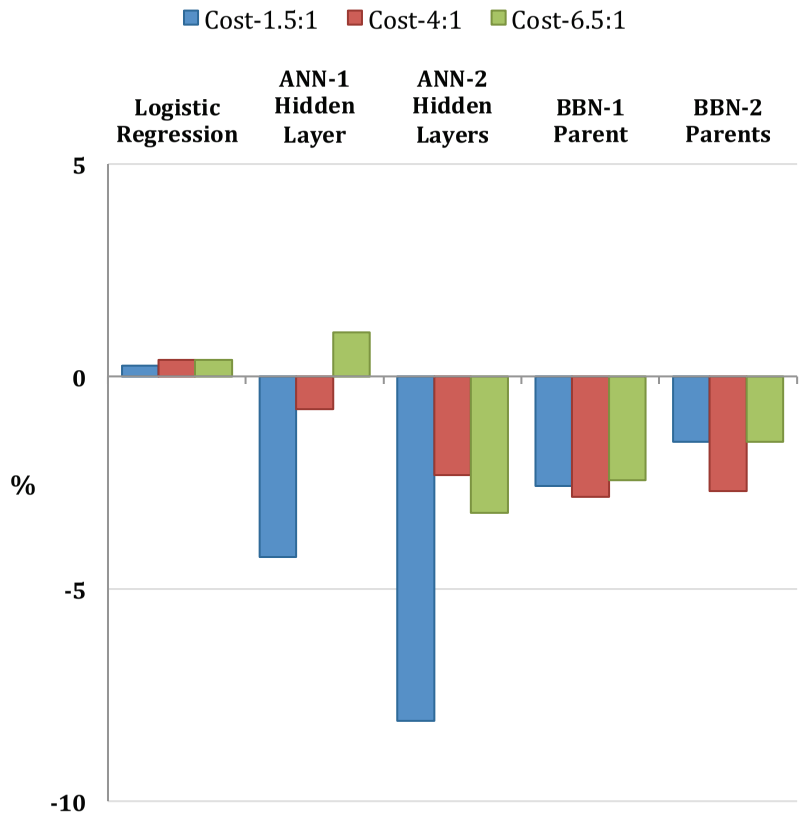

Figure 4-23: Experiment 4 ROC-AUC Compared to the Original Logistic Regression

**(Original LR ROC-AUC=77.7%)*

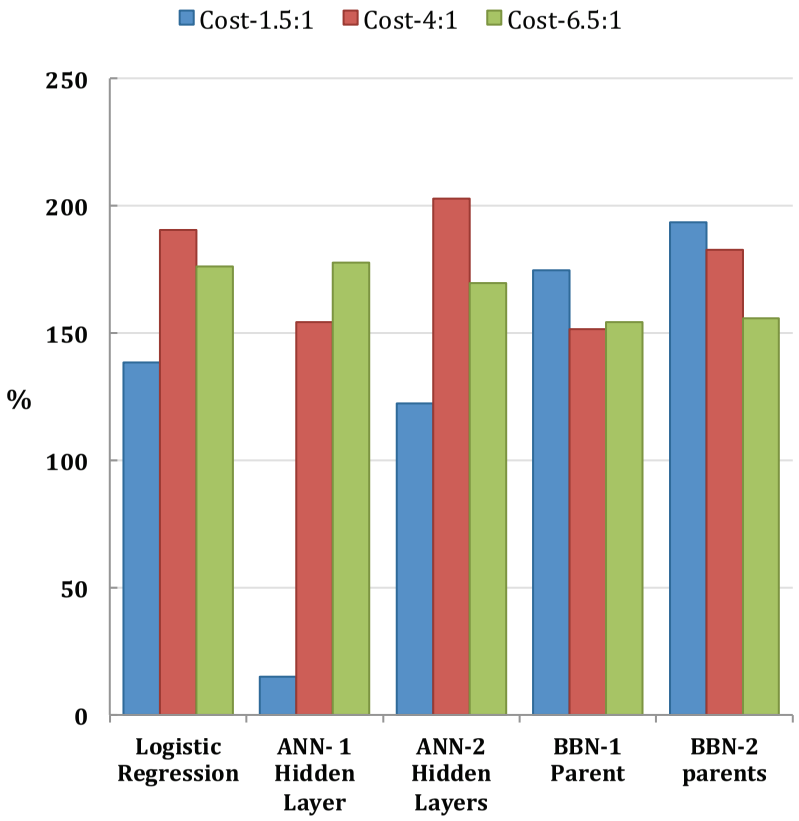

Figure 4-24: Experiment 4 F1-Score Compared to the Original Logistic Regression

*(Original LR F1-score=13.8%)

4.6. Chapter Summary

Developing a predictive model requires a good understanding of the problem in hand, and a systematic approach. In binary classification, the most commonly applied approach in medical literature is LR. The un-manipulated "Alive" training dataset was used to develop a reference model by applying LR. The ROC-AUC was 77.7%, specificity was high (99%) but sensitivity was low (9%). The poor sensitivity can be attributed to the class imbalance in the problem space (11.4% positive cases). The reference model has identified 8 out the 22 candidate attributes as significant (see Figure 4-5) with good discrimination (-2 Log Likelihood: Intercept=3059.85 → Model=2485.9, and a significant Hosmer and Lemeshow goodness-of-fit test).

After that, 2 ANN models and 2 BBN models were developed. The same training set was used to generate the models and the developed models were tested on the same independent test set. The Hanley and McNeil repeated measures ROC test showed that all of the models had an equivalent general performance. McNemar's test was used to corroborate these results, while the ANN (with 1 hidden layer) and the BBN (with 2 parents) had an equivalent performance when compared to the reference model. The core performance of the BBN (with 2 parents) generated the best F1-score (39.2%) and Kappa (31.8%), and the ANN (with 1 hidden layer) had the best precision (46.1%) (See Table 4-4: Summary of Experiment 1 Results).

Several experiments were conducted to investigate the effect of data and algorithm level manipulation on the performance and evaluate the stability of the developed models. The use of SMOTE has dramatically improved the performance of LR but caused a noticeable negative effect on the performance of ANN models. The BBN models exhibited some decline but were more stable. Reducing the amount of data available for training and using spread sub-sampling revealed a similar trend. The performance of LR has dramatically improved with a less profound effect on the other algorithms.

Next, the effect of imposing cost on the algorithm was evaluated. Cost sensitive classification was implemented by rewarding the algorithms for properly labeling the positive class. All of the models' performance measures have improved, especially sensitivity, albeit at the expense of specificity. Cost served as a more realistic approach and can be rationalized and replicated in nature. Throughout the experiments, BBN with 2 parents had the most stable general and core performance. Although, ANN with 2 parents generated the best F1-score, it had a very unstable performance.

4.7. Conclusion

Cognitive decline after surgery is a major concern, especially to patients and their families. Patients undergoing cardiac surgery are at higher risk of developing delirium, which have been linked to post-operative cognitive and functional decline. With the changing demographics and medical profile of patients undergoing cardiac surgery, increasing their vulnerability to adverse events; preventive strategies will become a fundamental part of the process of care.

The ultimate goal is to construct a model that is capable of identifying patients that are at risk of developing delirium after cardiac surgery in an attempt to initiate preventive measures that can divert the patient trajectory away from developing delirium. Because delirium is a complex multidimensional problem, relying on a simple solution may not generate satisfactory results. That was apparent from the original LR model results from Experiment 1. Although the ROC-AUC of the LR is considered to be good (~78%), it was mainly influenced by the presence of too many negative examples that resulted in very high specificity (~99%), yet very poor recall (~9%) and a low F-1 Score (~14%). Looking at a single measure, the ROC-AUC, to evaluate the model performance might lead to false conclusions. Clinicians and scientists need to be aware of the shortcomings of using a single measure, such as the ROC-AUC, when evaluating a predictive model. Examining the performance of predictive models requires a deeper understanding how to evaluate their performance, and considering different angles (measures such as the F1-

score and tests like the McNemar's test) to determine their usefulness in predicting the outcome of interest.

In terms of ANN-1 hidden layer and BBN-2 parents, they had similar general performance (ROC-AUC and McNemar's test), but superior ability of distinguishing positive cases (having a high recall, precision, Kappa, and F1-score). Class imbalance had a smaller effect on their performance, mainly due to their ability to distinguish non-linear relationships and to model them.

To mitigate the expected negative effect of inadequate representation of positive events (class imbalance), several data mining class imbalance techniques were explored on the classifiers developed in Experiment 1. All of the applied techniques have improved the performance of the LR model, demonstrated by the significant increase in the F1-score. Increasing the number of positive cases or decreasing the sample search space directed the algorithm focus to the positive class. In the case of applying cost, we assume that because we informed the algorithm of the inequality of error and its significances, the cost function and the search of local minima were more fitting to the problem.

Although ANN-1 hidden layer relies on SGD to identify its local minima, the change of the training sample space negatively affected its performance. We believe this is due to the high variance in the optimal network structure that is influenced by the available input data. Applying cost had minimal effect on the ANN-1 hidden layer. This might be due to that fact that ANN evaluates each variable at each step and applying cost will not affect the performance because the algorithm had adjusted the cost through the built in back-propagation adjustments of weights.

BBN-2 parents, like ANN-1 hidden layer, rely on SGD to find its local minima in the sample space. Increasing the number of positive cases or decreasing the sample search space had a less negative effect on the BBN-2 parents in comparison to ANN-1 hidden layer. Applying cost did not improve the BBN-2 parents performance because it optimized the network topology based on set scoring criteria (e.g.: MDL, AIC, Bayes), which searches for the best network structure before calculating probabilities.

Throughout experiments 2–4, manipulating the data or the algorithm parameters did not improve the ROC-AUC. It was very clear that a tradeoff between recall (sensitivity) and

specificity. But increasing recall (increasing the classifier ability of correctly labeling that did develop delirium) did not improve the algorithm precision, which indicates that the algorithm was making more mistakes by labeling negative examples as positive (false positive).

When comparing all of the experiments, the un-manipulated BBN-2 parents provided a simple but elegant graphical representation of the problem space that can be interpreted and validated by domain experts, and also can be applied by a machine to predict delirium.

Table 4-10: Summary of All Experiments

		ROC-AUC	Accuracy	Kappa	Recall	Specificity	Precision	F1-Score
Original Logistic Regression		77.7	**87.7** ‡	**9.5** ¡	**8.6** ¡	**98.7** †	33.3	**13.8** ¡
Experiment 1 (Un-manipulated)	ANN (1 Hidden)	76.7	**88.1** †	28.4	27.6	**95.9** ‡	**46.1** †	34.5
	ANN (2 Hidden)	76.9	85.5	29.01	37.8	91.6	36.6	37.2
	BBN (1 Parent)	75.7	80.8	26.5	49.6	84.7	29.4	37
	BBN (2 parents)	76.4	86.9	**31.8** ‡	37	93.3	41.5	39.2
Experiment 2 (SMOTE 100%)	Logistic Regression	73.6	86.1	23.5	27.6	93.6	35.7	31.1
	ANN (1 Hidden)	72.9	84.2	13.5	19.7	92.5	25.3	22.1
	ANN (2 Hidden)	**67.7** ¡	85.1	15.7	20.5	93.3	28.3	23.7
	BBN (1 Parent)	75.4	80.1	26.7	52.8	83.6	29.3	37.6
	BBN (2 parents)	74.5	83.9	24.7	36.2	90	31.7	33.8
Experiment 2 (SMOTE 200%)	Logistic Regression	70.5	84.6	24.7	37	89.6	31.3	39
	ANN (1 Hidden)	68.1	82.6	11.3	22.8	88.9	**20.9** ¡	21.8
	ANN (2 Hidden)	72.1	85.1	19.7	34.6	87.7	26.5	30
	BBN (1 Parent)	75.2	80.1	24	47.2	84.1	27.6	34.9
	BBN (2 parents)	72.3	82.9	20.6	30.7	90.4	29.1	29.9
Experiment 2 (SMOTE 300%)	Logistic Regression	70.5	83.6	24.7	37	89.6	31.3	39
	ANN (1 Hidden)	68.1	81.4	11.3	22.8	88.9	20.9	21.8
	ANN (2 Hidden)	72.1	81.7	19.7	34.6	87.7	26.5	30
	BBN (1 Parent)	75.2	80	24	47.2	84.1	27.6	34.9
	BBN (2 parents)	72.3	83.6	20.6	30.7	90.4	29.1	29.9
Experiment 3 (Sub-Sample 1:1)	Logistic Regression	76.8	73.6	22.8	63.8	74.8	24.5	35.4
	ANN (1 Hidden)	75.3	**63.9** ¡	16.7	**73.2** ‡	**62.7** ¡	**20.1** ¡	31.6
	ANN (2 Hidden)	74.1	67.6	20.2	**74** †	66.8	22.2	34.2
	BBN (1 Parent)	75.8	70	21.2	70.1	70	23.1	34.7
	BBN (2 parents)	74.9	68.8	19.4	68.5	68.8	22	33.3
Experiment 3 (Sub-Sample 1.5:1)	Logistic Regression	76.7	78.8	26	55.1	81.8	28	37.1
	ANN (1 Hidden)	75.6	73.3	23.1	65.4	74.3	24.6	35.8
	ANN (2 Hidden)	74.9	70.6	20.7	66.9	71.1	22.9	34.1
	BBN (1 Parent)	75.2	73	22.2	63.8	74.3	24.1	35.1
	BBN (2 parents)	75.8	73.8	21.3	59.8	75.6	23.9	34.2
Experiment 3 (Sub-Sample 2.5:1)	Logistic Regression	77.1	83.7	31.4	48.8	88.2	34.6	**40.5** ‡
	ANN (1 Hidden)	76.3	77.7	27.5	62.2	79.7	28.2	38.8
	ANN (2 Hidden)	70.6	76.4	22.6	55.1	79.1	25.3	34.7
	BBN (1 Parent)	75.7	76	23.7	59.8	77.7	25.6	35.8
	BBN (2 parents)	74.9	79.1	24.3	50.4	82.8	27.4	35.5
Experiment 4 (Cost 1.5:1)	Logistic Regression	77.9	87.6	26.4	26.8	95.4	**42.5** ‡	32.9
	ANN (1 Hidden)	74.4	85.1	26.1	34.6	91.5	34.4	34.5
	ANN (2 Hidden)	71.4	85.1	22.4	29.1	92.2	32.5	30.7
	BBN (1 Parent)	75.7	80	27.1	54.3	83	29.1	37.9
	BBN (2 parents)	76.5	86	**32.6** †	41.7	91.2	39.3	**40.5** ‡
Experiment 4 (Cost 4:1)	Logistic Regression	**78** ‡	82.6	30.5	51.2	86.7	33‡	40.1
	ANN (1 Hidden)	77.1	73.5	22.4	62.9	74.8	24.3	35.1
	ANN (2 Hidden)	75.9	79.6	31.3	64.6	81.5	30.9	**41.8** †
	BBN (1 Parent)	75.5	75	22.2	58.3	77.2	24.7	34.7
	BBN (2 parents)	75.6	79.9	28.3	56.7	82.8	29.8	39
Experiment 4 (Cost 6.5:1)	Logistic Regression	**78** ‡	77.4	26.6	61.4	79.4	27.6	38.1
	ANN (1 Hidden)	**78.5** †	75.7	26.4	66.1	77	26.9	38.3
	ANN (2 Hidden)	75.2	77	25.5	59.8	79.2	27	37.2
	BBN (1 Parent)	75.8	74	22.4	63	74.8	24.3	35.1
	BBN (2 parents)	76.5	74.1	22.8	62.2	75.7	24.7	35.3

**Measurements are in %*

† Best performing model, ‡ Second best performing Model, ¡ Worse Performance

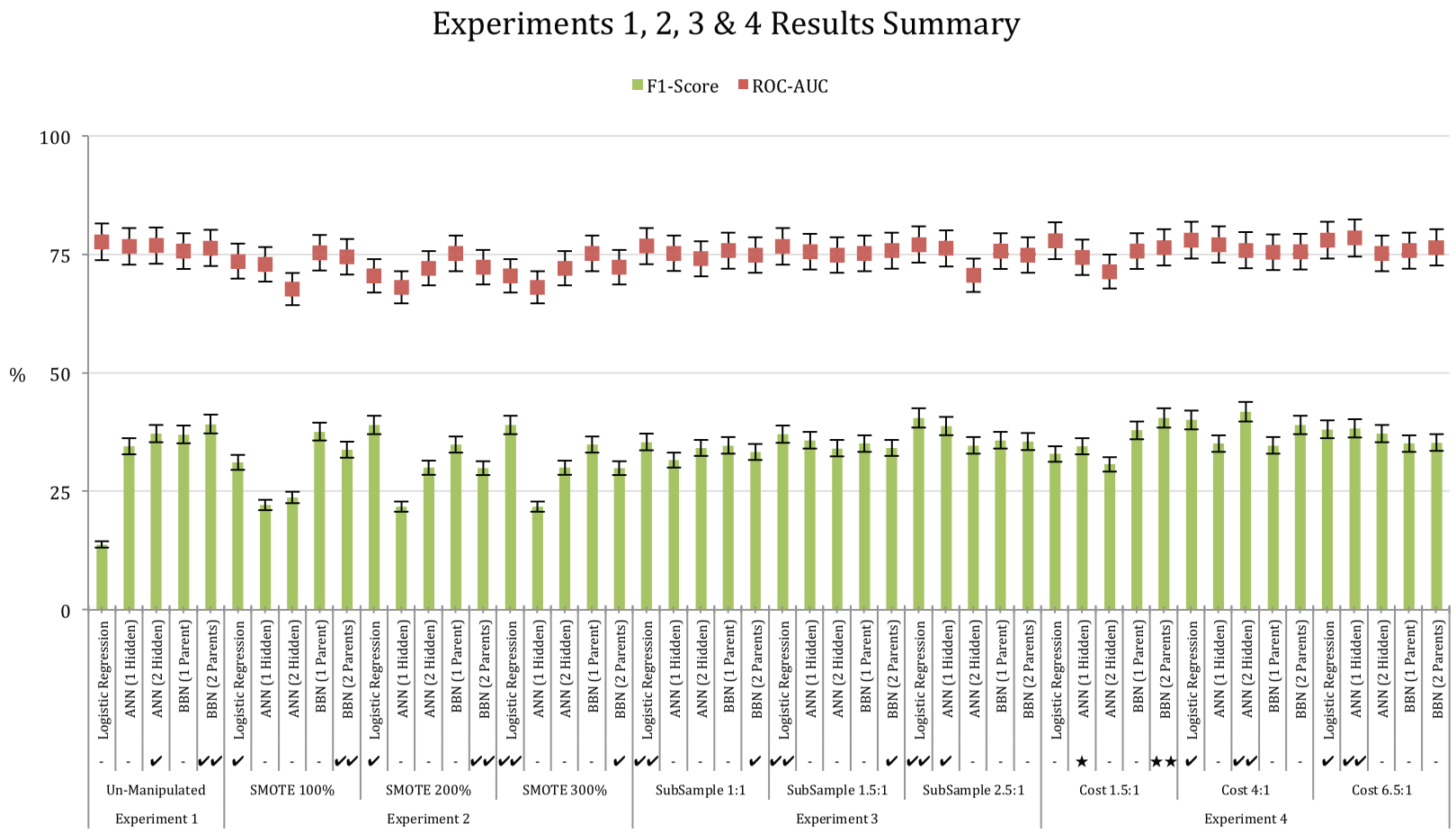

Figure 4-26: Results Summary. *(Error bars represent the 95% confidence intervals)*

✔✔*: Best performance for Experiment (n),* ✔*: Second best performance for Experiment (n),*

★★*: Best overall performance,* ★*: Second best overall performance*

CHAPTER 5: DISCUSSION

Advances in technology have greatly increased the amount of data and information available for clinical decision-making. Analyzing this information will uncover new knowledge, which can be used to create superior predictive models that are geared towards providing personalized plans supported by high quality, evidence-based care. This thesis presents the application of several data mining methods to generate a predictive model for delirium after cardiac surgery.

5.1. Preventing Delirium after Cardiac Surgery

Post-operative delirium is clinically challenging, but can be prevented if approached proactively. Patients undergoing cardiovascular surgical procedures are at higher risk of developing delirium[9, 10, 14]. Therefore, prevention and early recognition of delirium are important. Yet conventional statistical methods have not been able to produce reliable models that can be generalized and used effectively to detect delirium[8, 10, 11, 13, 15, 17-20, 22-24, 29, 51, 52, 75, 79, 180].

Recently there has been a trend towards complementing evidence-based medicine with personalized medicine. In evidence-based medicine, the choice of ideal therapy is based on population studies and clinical trials. This approach benefits from the large sample size and statistical power in a fairly uniform group of patients; but it overlooks the fact that every patient is different. To overcome this limitation, personalized predictive models are being developed. Personalized predictive models exploit all available information from each patient to identify the best intervention for an individual patient[40, 41, 43, 69, 139-141, 144, 181]. This approach improves outcomes, as it is draws a more holistic picture to alleviate uncertainty in predicting outcomes based only on the treatment[40, 41, 43, 139, 140, 142, 182]. Identifying the right group of patients that can benefit from a specific intervention will reduce the costs and improves the results of the intervention. Identifying patients who may not respond to an intervention could prevent undesirable possible

adverse effects and redirect resources. In this work, we demonstrated the machine learning methods are capable of tackling complex medical predictive problems and are able of generating better predictions; with a statistically comparable general performance (ROC-AUC=76.4-77.7%) but a significantly improved ability of identifying positive cases (F1-Score for BBN with 2 parents=39.2% and ANN with 1 hidden layer= 37.2%, compared to the reference LR= 13.8%). These methods can be used to support critical decisions, optimize utilization, and improve patients' outcomes.

5.2. Objectives and Methods Summary

Three objectives were identified in this thesis, to aid the prediction of delirium following cardiac surgery (INTRODUCTION 1.2 Research Objectives, page 4 and Table 1-1 Thesis research objectives, methods and expected outcomes, page 6). Table 5-1 presents the objectives with how they were answered in this thesis, and what the overall contribution of these methods to the development of a predictive model for the detection of post-operative delirium in cardiac surgery patients.

The objectives and achievements from this research are relisted here:

1. The first objective was to identify key features that influence the development of delirium after cardiac surgery. To accomplish this objective, we went through several pre-processing steps to identify the optimal attribute vector space that can optimize the learning process and capture the complexity of delirium. We adapted the conventional statistical approach that is based on statistical significance and domain knowledge to identify candidate attributes. Machine learning feature selection methods were used to corroborate the results of the conventional approach and highlight hidden important attributes. By doing so, we discovered that 5 attributes were picked at least 80% of the time by all methods (EUROII score, blood product transfusion within 48 hours from surgery, prolonged mechanical ventilation, prolonged ICU stay, and history of being turned down) signifying their importance. One of the main contributions of this is that it

highlighted the relationship of blood product transfusion and prolonged stay in the ICU because of delirium.

2. The second objective was to develop predictive models that are capable of capturing the complexity associated with delirium in adult cardiac surgery patients. Five different models were developed using the training set: LR, ANN-1 hidden layer, ANN-2 hidden layers, BBN-1 parent, and BBN-2 parents. Then, we evaluated the performance of these models on an independent test set to identify what worked best. The LR model, the conventional modeling approach according to medical literature, resulted in a good ROC-AUC (~78%) and specificity, but had a poor performance in all other measures. ANN-1 hidden layer and BBN-2 parents produced an equivalent ROC-AUC but had superior results in all other metrics. We were able to show that applying data mining models to a complex medical task will generate superior results when compared to traditional methods, based on a comparable AUC of the LR, ANN and BNN (76.4-77.7%) but a superior ability of the ANN and BBN of identifying positive cases (F1-Score: 37.2 and 39.2) when compared to LR (F1-Score: 13.8).
3. The third objective was to experiment with different approaches that are designed to address the issue of class imbalance. We assumed that the poor performance of the classifiers was due to the outcome class imbalance. To mitigate the effect of class imbalance and evaluate the impact of manipulating the outcome class representation on the model performance, several techniques that are designed to deal with class imbalance were used (either manipulating the class representation at the training data level or the way that the algorithm handles the miss classification error for each class). At the data level, increasing the number of positive examples in the sample space (by using SMOTE or sub-sampling) resulted in a dramatic improvement in the LR performance with some deterioration of the other classifiers (ANN-1 hidden layer >> BBN-2 parents). At the algorithm level, applying cost has also resulted in a dramatic improvement in the LR performance with minimal effect on the other classifiers (ANN-1 hidden layer > BBN-2 parents).

Table 5-1: Contributions of the Methods Developed in this Thesis

Objective	Method(s)	Contribution	Impact on Delirium
-Identify pre-, intra-, and post-operative attributes that influence the development of delirium after cardiac surgery	-Conventional statistical (statistical significance and domain knowledge) -Feature selection methods	-Decreased the feature space from 229 down to 26 key candidate attributes -Identified several critical attributes -Highlighted several pre-operative features -Used machine learning methods to corroborate the conventional approach	-Preoperative interventions can be directed towards optimizing the patient medical and physical condition before surgery -These risk factors will inform the patient's choices -Alert the care team so appropriate measures can be taken
-Develop a model to predict delirium in cardiac surgery patients and evaluate their performance on an independent dataset	-5 different models were developed from the "Alive" training dataset: LR, ANN-1 hidden layer, ANN-2 hidden layers, BBN-1 parent, and BBN-2 parents.	-ANN-1 hidden layer and BBN-2 parents generated superior results in predicting delirium when compared to LR	-Predicting a complex problem, like delirium, will require a complex solution to capture the hidden non-linearity -Machine learning algorithms can be applied to solve complex medical tasks - BBN has a simple but elegant graphical representation that can be understood by machines and humans
-Class imbalance manipulation and its effect on the models	-Two data level techniques (SMOTE and *SpreadSubSample*) and a single algorithm parameter manipulation technique (cost) were applied to the developed algorithms	-Algorithms that are based on BGD (LR) had a significant improvement with all manipulation techniques -Algorithms that are based on SGD (ANN and BBN) suffered some deterioration. This is most likely to change of the optimal network structure.	-Class imbalance manipulation techniques can be used to evaluate the model stability and durability in different environments -BBN-2 parents model have shown a decent stability throughout the experiments, which might indicate its generalizability and validity in predicting delirium when compared to the other models

5.3. Pre-, Intra-, and Post-operative Predictors of Delirium after Cardiac Surgery

One of the contributions of this work is that it highlighted several modifiable pre-operative factors that, if optimized before surgery, might help in preventing delirium. These include: pre-operative hemoglobin, pre-operative intubation pre-operative diabetic control, and frailty. If an intervention can be directed towards optimizing the patient medical and physical condition before surgery and performing surgery electively, we believe that the incidence of delirium will decline. These attributes might inform patient choices and alert the medical team, so appropriate interventions can be initiated in a timely fashion to minimize potential complications.

5.4. Developing a Model to Predict Delirium after Cardiac Surgery

5.4.1. The Traditional Method for Predicting Delirium

The traditional approach in the medical literature to predict delirium would be to use a binary LR model. This technique will identify the important independent attributes. This approach generated a model that included 8 independent attributes (out of 22 candidates) with a reasonable discriminative power in an independent test set (ROC-AUC= 77.7%) and good fit to the data (Hosmer and Lemeshow goodness-of-fit test Chi^2= 30.2, p-value=<0.05). However, recall (8.6%) and precision (33.3%) were poor. This might be secondary to the class imbalance and the approach that LR takes to minimize the error. In this situation we assume that this is the best model, that these are the most important attributes that influence the outcome, and overlook any other possible scenarios. Clinicians and health care professionals need to be aware of the shortcomings of using a single measure, such as ROC-AUC, when evaluating a predictive model.

5.4.2. Alternative Modeling Methods for Predicting Delirium

Data mining techniques have a great potential for discovering hidden important patterns and generating useful information from these patterns that will lead to actionable insight.

The traditional statistical approach generated a model that had a reasonable performance on an independent dataset (ROC-AUC= 77.7%). When the same attributes were applied to 2 different modeling methods (ANN and BBN); a similar general performance was obtained (ROC-AUC: 75.7–76.9%). Although, looking into their core performance, these methods exhibited a better predictive power (see Table 4-4: Summary of Experiment 1 Results, page 76 and Figure 4-8: Experiment 1 results – Original training Set, page 79). We believe that ANN and BBN had better results because they incorporated several attributes into the models and allowed the model to come up with the best prediction that fits the problem space. Some of the included attributes had very low statistical significance; but the ANN and BBN models have identified some hidden potential in them that have resulted in better predictions.

Unfortunately, none of the developed models displayed superior results in predicting delirium (ROC-AUC <85%). This might be due several reasons, which include: the significant class imbalance, the complexity of the problem space, and the possibility of missing some important attributes.

5.4.3. Model stability and Enhancement with Manipulation

To explore the impact of class imbalance manipulation; two different methods were used (sample manipulation and cost sensitive learning). Attempting to adjust the distribution of classes, by either increasing the minority class representation in the sample space, making it more general, or decreasing the over represented majority class and focusing the classifier attention on a smaller sample space, have significantly improved the performance of LR.

Over-sampling with SMOTE has caused a significant deterioration of the ANN models' performance at all levels, most likely due to the change in the optimal network structure because of the change in the sample space. In the case of BBN, some decrease in

performance was also noticed but it was not as significant. In the over-sampling experiments, the methods have overestimated the predictive accuracy although the distribution of classes went from 12:88% to as close as 34:66%. The ROC-AUC gave a more dependable representation of the classifier performance. This supports the notion that is published in the data mining and predictive modeling literature in the case of class imbalance; the use of ROC-AUC is a more appropriate measure for the assessment to compare different models even if the distribution is adjusted with new synthetic samples[56-59, 64, 65, 67, 68, 127, 129].

Under-sampling with SpreadSubSample has dramatically improved all of our models recall but with no much effect on the ROC-AUC. This is mainly due to the noticeable decrease in the specificity. An interesting observation was the tradeoff between sensitivity and specificity; also, increasing sensitivity did not improve precision, Kappa and the F1-score. Under-sampling moved the predictive accuracy closer to the ROC-AUC (see

Figure 4-17: Experiment 3 Results – Spread Sub-sample, page 94). We speculate that this is due to the reduction of noise and restricting the sample space.

Applying cost sensitive classification by rewarding the algorithms for making the right choice had the best results in our experiments. The application of cost improved the core performance without compromising the other measures. Adding cost will cause a certain decrease in specificity but with a noteworthy improvement in the other measures without affecting the ROC-AUC. Applying cost at the ratios of 1.5:1 and 4:1 have produced very balanced and decent models. Although ANN with 2 Hidden layers and LR at a cost of 4:1 generated the best F1-score, the BBN with 2 parents had the best overall performance (Figure 4-25)

5.5.Limitations

Some of the limitations of this work include the latent bias of retrospective studies that is based on observational data that is based on chart abstraction. The quality of the acquired data might profoundly impact the models and their interpretation. The prevalence of

delirium was only 11.4%. This low representation is most likely due to the definition of delirium in the source database (only agitated sub-type). This will limit the ability to generalize the developed models to the other types of delirium other than the agitated type, which only represents 2-15-% of all delirium cases [12, 15, 20, 22, 25, 27]. Hypoactive delirium has a higher prevalence and is associated to postoperative complications and mortality [12, 19, 20, 24, 28, 76, 99]. Because the diagnosis of delirium is based on chart abstraction from the daily progress or nursing notes and is not based on a standard tool, the diagnostic accuracy of delirium in the dataset is questioned and most likely it is overlooking a grater portion of patients who actually developed delirium.

The generated models were tested on a test set and their general performance was considered satisfactory. Even though different methods were used, none of the models had a superior general performance (~78%). This might indicate that there are some missing important features that are not captured by these models. Additional clinical, biochemical, psychological and genetic features such as history of depression, family history of post operative delirium, poly-pharmacy, pre-operative mini mental test, inflammatory biomarkers levels (C-reactive protein and Cerebral fluid dopamine levels), and genetic predisposition to neuropsychological disorders may help explain why some patients are more prone to the development of postoperative delirium.

One of the major drawbacks of ANN is that it is considered as “Black Box Modeling”. It lacks the ease of interpretability by none experts (like medical professionals) that is a big advantage of the logistic regression model. While BBN can provide probabilities that can be easily understood and can uncover interesting interactions between different attributes, but they suffer the skepticism that accompanies Bayesian approaches, like that a model parameter can always be updated by new observations.

We acknowledge that exploring a different network structure or setting might have improved the performance of ANN and BBN. This might indicate that a static model is not the optimal solution and there lies the necessity for a re-learning/re-evaluation cycle with new observation. This also displays some of the disadvantages of using a static ANN in a dynamic process added to the difficulty of establishing etiologic interpretation for the calculated weights. It also illustrates the robustness, consistency, and the demonstration of

a causal/evidential relationship of BBN, and graphical models in general, in the presence of class imbalance and its manipulation.

5.6. Future Work

Existing research acknowledges that delirium is a major concern after surgery, to patients, the health care community, and policy makers. Most of the current research focuses on using conventional statistical methods (e.g.: LR). In this work, we demonstrate that an alternative approach to modeling medical outcomes has an equivalent general performance and a superior internal performance in detecting delirium in adult patients after cardiac surgery. We would like to explore the effect of using an ensemble of models on the performance. We also would like to test the utility of a dynamic graphical model (Markov Chain) on the same dataset.

It will be exciting to embed these models in a clinical decision support system that is connected to an electronic medical record to evaluate their performance. Some of the challenges that face this paradigm shift lie in the end users, health care providers, and their receptiveness to change. Also the adaptation of electronic health systems and patients records is still in its infancy and it has been faced with great skepticism from the health care community. Another challenge that we foresee is the presentation of this information to the healthcare provider and its effect on their adaptation. Finally, what should the health care provider do with this information, what action should they take?

In the present economic environment, the improvement of medical care, and the escalation of healthcare cost; healthcare professionals, policy makers, and patients must use technology to their advantage and make decisions that are based on genuine information. The future of personal health records is a single electronic record inclusive to all patient encounters and metadata. These applications would not only be useful to the patient and the system, but would be a method to collect data in which a dynamic predictive model can transform these complex interactions between several characteristics into a format that can be understood by machines and domain experts, to forecast a potential complication from a specific intervention.

If we can learn from the Internet and agree on a single format, then aggregating several sources of data into a single entity (data warehouse) that is governed yet unconstrained by physical, political, and illogical borders will enable the health care community to create much more superior models that are based on a large volume of heterogeneous data that is more expressive of the inherent variability and diversity in nature. This will prevent the undesirable complication, maximize healthcare efficiency, and improve the personalization of care.

Although delirium is a well-recognized entity, there is no specific treatment and most of the interventions focus on alleviating its negative effects. The key here is identifying these high-risk patients, presenting them with the facts and personalizing their care plan to assure that they obtain best possible outcome with the best quality of life based on their values and personal goals.

5.7. Conclusion

Post-operative complications are a major concern for the healthcare professionals and their patients. Of particular concern to patients is the effect of surgery upon brain functions. Also with the growth in the aging population, health care is facing an interesting shift of patients prospective with an increased attentiveness of the quality of life after the intervention. Several authors agree that delirium is associated with negative effects and advocate early recognitions and preventive measures[7, 9, 10, 12-15, 20, 22, 24, 25, 28, 35, 52, 75, 76, 99]. In the absence of solid evidence; patients and policy makers rely on healthcare professionals to make the right decisions. Most of the decisions that medical professionals make are based on data from large population studies and clinical trials. Every patient is objectively comparable to another one, all of them are getting the same intervention, randomization, and its ability to show efficacy. This approach overlooks the fact that every patient is different, has a limited generalization, and it introduces an artificial environment that is difficult to replicate in nature.

Although there is no lack of data in the medical domain, the use of this data is not optimized to extract all hidden information in it. Vasant Dhar stats that “The era of

limited data and assumption driven modeling is largely over."[182]. He also states that "If a problem is non-stationary and a model is only an approximation anyway, why not build the best predictive model based on data available until that time and just update it periodically"[182]? The use of machine learning to explore interesting questions humans might not consider is a central focus in big data science.

Recently, there has been a recent trend towards complementing evidence-based medicine with personalized medicine[41, 43, 111, 139, 140]. The large volumes of new medical research discoveries (e.g.: genetics, new chemotherapy, new procedures, etc.) and the changing patient population demographics will progressively confront health care professionals and policy makers. With the increased complexity of the generated data, computers will serve as an essential asset in improving the understanding of the complex interactions in the data. Some authors realize that the power in big data and data mining involves its superior ability to forecast the future based on the past (predictive modeling)[142, 182]. The utility of predictive modeling and its capability in preventive medicine cannot be over emphasized.

The findings from this work substantiate the effectiveness of machine learning modeling and data mining methods in predicting medical outcomes, which are inherently complex, and yielding more precise predictions. In the case of delirium, when compared to logistic regression, ANN and BBN had a better capability in identifying positive cases even in the presence of class imbalance. BBN have shown a better stability through out multiple experiments that was not the case for LR or ANN. BBN can provide an easy to understand graphical representation of the attributes relationships, can uncover interesting interactions and can provide probabilities, which can be easily understood by medical professionals. Also, we illustrate the importance of using of multiple measures when evaluating model performance (e.g.: F1-Score). Health care professionals should not rely on a single measure to evaluate a model performance. Using a single measure, like ROC-AUC, might lead to false assumptions that might result in wrong decisions.

The use of novel technologies that supplements high quality, evidence-based medicine with rigorously developed complex predictive models will definitely improve the quality of care, patient's outcomes, and reduce the burden on the health system.

APPENDIX A: ATTRIBUTES

Below is a description of each attribute category:

Demographics: Transfer from Another Hospital, Same Day cases, age at surgery, gender, date of admission, date of discharge, date of surgery, weight (Kg), height (cm), Smoking.

Frailty was defined as impairment in activities of daily living scored by the Katz ADL index[122], which measures diminishing independence in any of the following: Feeding, Bathing, Dressing, Transferring, Toileting, Continence; or limitation in ambulation (patient is using a walking aid or requires assistance for normal daily activity or requires a wheel chair); or a documented history of dementia[26].

Comorbidities: diabetes mellitus, diabetes mellitus control, dyslipidemia (DLP), renal insufficiency (creatinine above 176 μmol/ L), renal failure, renal dialysis, highest preoperative creatinine, HTN, pulmonary hypertension (PHTN), CVD, CVA, CVA timing, peripheral vascular disease (PVD), COPD, pre-operative intubation, history endocarditis and type of endocarditis.

Previous cardiac intervention: percutaneous intervention (PCI), PCI Date, PCI same admission, PCI accident, Emergency Surgery.

Cardiac specific: history of acute coronary syndrome (ACS), timing of ACS, CHF, admission with angina, hemodynamic instability, cardiogenic shock, arrhythmia, type of arrhythmia, preoperative atrial fibrillation (AF), type of AF, Canadian Cardiovascular Society (CCS) Angina class, NYHA dyspnea class and Valve disease (Stenosis, Insufficiency and Etiology).

Pre-operative: Ejection Fraction and urgency for surgery (elective: stable cardiac function in a patient waiting at home to undergo surgery; in-house: the need for hospitalization of a patient awaiting surgery; urgent: the need for surgery within 24 hours to minimize further clinical deterioration; and emergent: ongoing, refractory cardiac compromise, with or without hemodynamic instability, and unresponsive to any form of therapy except cardiac surgery without delay). Pre-operative medications included: ACEI, ARB, Beta Blockers (BB), Calcium Chanel Antagonist (CaCA), intravenous Heparin (IV-Hep),

Warfarin, Inotrops, lipid lowering agents, ASA, ASA last dose, Plavix, Plavix last dose, and Anti-Platelets stopped prior to Surgery.

Intra-operative: operation performed (CABG, CABG ± Valve, Valve Replacement and Valve Repair), Valve operated on, valve primary pathology (Stenosis, Insufficiency and Etiology), Cardioplegia, Cardioplegia Infusion Mode, Clamp Time, Perfusion Time, Return to CBP, Preoperative Hemoglobin, IABP, IABP time of Insertion, IABP Indications, Pacing, Type of Pacing,

Post-operative: Inotropes in admission to CVICU, Blood products use (Type and Timing), Re-operation, reason for Re-operation, Cardiac Complications (Low CO, Valvular, Cardiac Arrest, Heart Block, Permanent PM, AF and New AF), GI Complications (Bleeding, Ischemia, Other, and GI Surgery), Infectious Complications (Sternal wound Infection (Superficial and Deep), Septicemia, UTI, and Others), Pulmonary Complications (Ventilator Prolonged, Days Ventilated, Pulmonary Edema, Re-intubation, Hours Ventilated and Hours in CVICU), CVICU Readmission (Ventilation during Readmission, Reason for readmission, Multiple CVICU readmission, Renal Failure, Type of Renal Failure, Dialysis, Other Complications).

Discharge attributes: Discharge destination, In-Hospital Mortality, Date of Death, Cause of Death

Examples of newly created attributes: Creatinine Clearance, Body Surface Area, Body Mass Index, PCI to Surgery time, Ventilator time, and Return to ICU time.

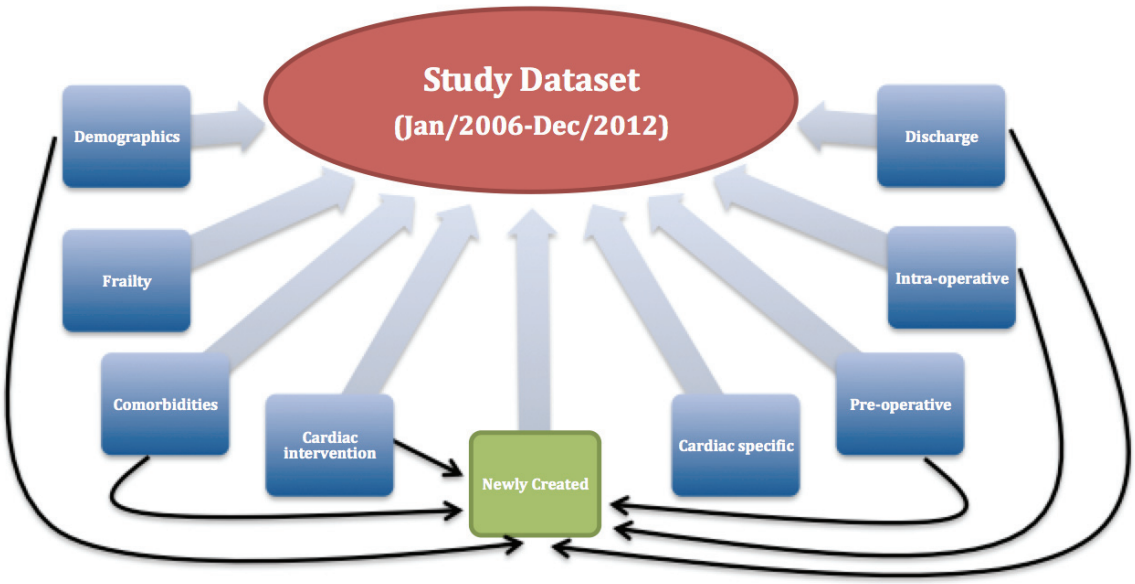

Figure A-1: Study Dataset Attributes categories
Newly Created: indicate attributes that are not present in the MHC registry and were newly derived from other attributes.

A. I Derived Attributes

In addition to these fairly standard elements, a number of newly created ones were also assessed. These included:

- ✓ Body mass index (BMI) (unit= Kg/m^2)

$$BMI = \frac{Weight[in\ Kg]}{Height[in\ m^2]}$$

- Body surface area (BSA) (unit= m^2), using Mosteller formula[183]

$$BSA = \sqrt{\frac{Weight[in\ Kg] \times Height[in\ cm]}{3600}}$$

- Time from Percutaneous Coronary Intervention to surgery (in hours)
- Time on mechanical ventilation (in hours)
- Time to return to the Intensive Care Unit (in hours).

- Estimated Creatinine Clearance (eCrCl) based on approximation of an estimated glomerular filtration rate (eGFR) using either the Cockcroft-Gault and Jelliffe equations:

    a. Cockcroft-Gault Equation (unit= ml/min/1.73 m^2)[184]

$$eC_{cr} = \frac{(140 - Age[in\ years]) \times Weight[in\ kg] \times Constant^{*}}{Serum\ Creatinine[in\ \mu mol/L]}$$

*Where *Constant* is *1.23* for Male and *1.04* for Female.

    b. Jelliffe formula for Creatinine clearance estimation (unit= ml/min)[185]

$$CrCl = \frac{\left\{\left(98 - 0.8 \times (Age[in\ years] - 20)\right) \times \left(1 - (0.01 \times Gender)\right) \times \left(\frac{BSA}{1.73}\right)\right\}}{(SCr \times 0.0113)}$$

CrCl, creatinine clearance (ml/min); Gender: male= 0 & female= 1; BSA, body surface area (Mosteller); SCr, serum creatinine (μmol/l);

    c. Notes:

        i. Although we are aware that these equations are the simplest, least accurate and returns a surrogate of creatinine clearance compared to other more sophisticated estimations; they were chosen for the ease of their implementation and the availability of its components in the dataset.

        ii. We have chosen 2 equations to increase the accuracy of the Creatinine Clearance measurement.

APPENDIX B: PREPROCESSING

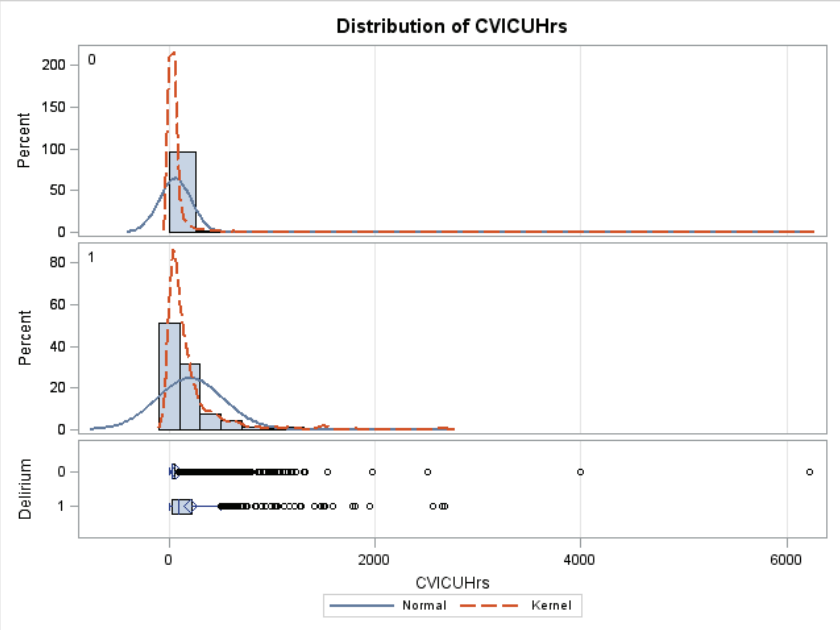

Figure B-2 CVICUhrs Histogram by Delirium Status
(0=Absent and 1=Present)

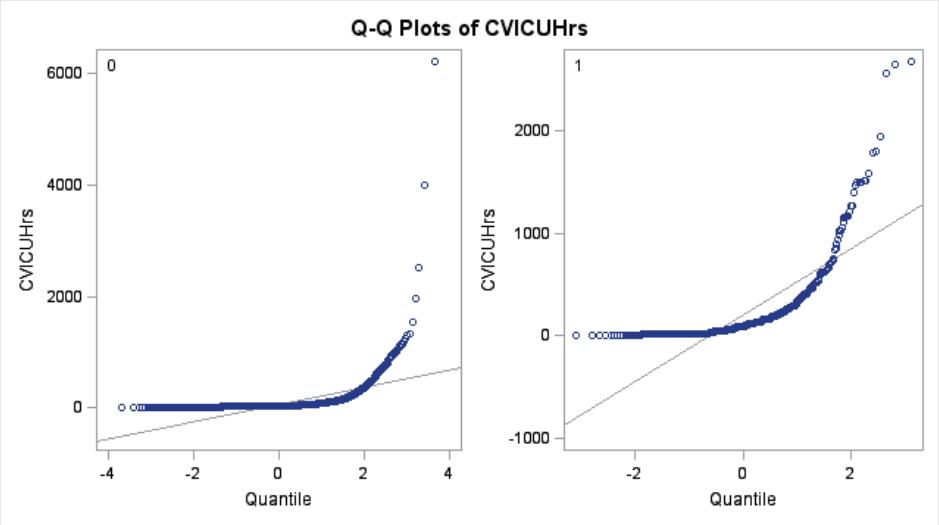

Figure B-3: Quintile - Quintile plot for CVICUhrs by Delirium Status
(0=Absent and 1=Present)

B.I. In-Hospital Mortality, Post-Operative Stroke and Delirium

In this analysis, there was no statistical evidence that delirium is associated with short-term mortality (in-hospital mortality). The incidence of delirium among patients who experienced in-hospital mortality (total=214) was 12.6%, which was not statistically different from the rate of delirium among those who survived (Figure 8). In comparison, the incidence of stroke was much higher in patients who suffered in-hospital mortality compared to those who did not (OR=9.75, CI=7.02-13.51)(Figure B-4).

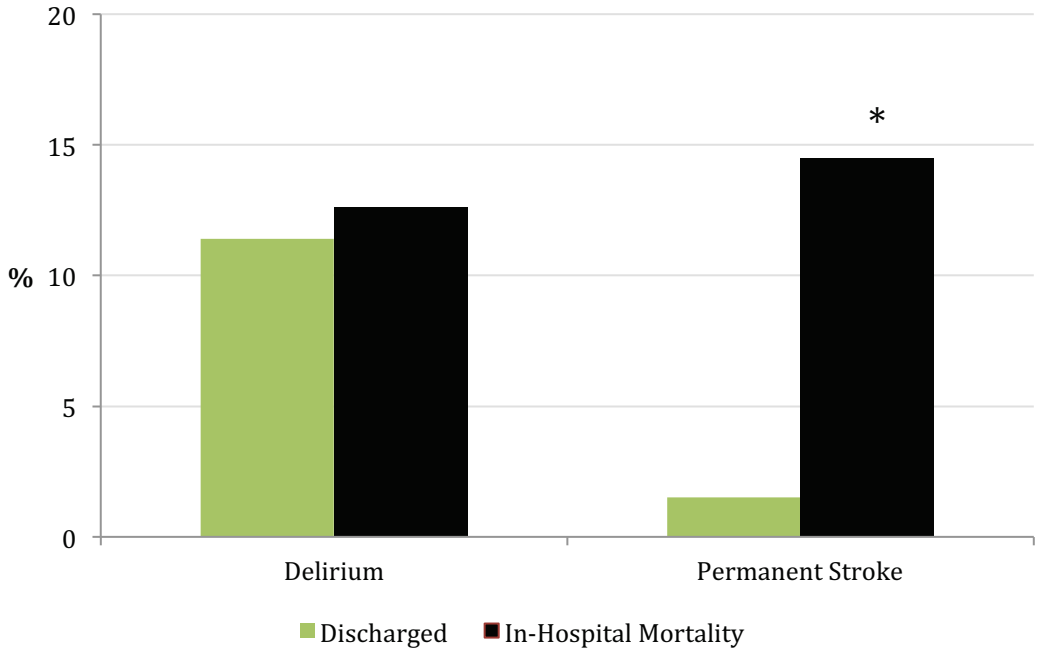

Figure B-4: Permanent Stroke and Delirium by In-hospital Mortality or Discharge
** Indicates: statistically significant difference (α=<0.05)*

As shown in the exploratory univariate analysis of the full dataset, there was a strong association of delirium with several post-operative complications (e.g.: pneumonia, pulmonary edema, sepsis, urinary tract infection (UTI), and discharge to an institution) but not with post-operative stroke or in-hospital mortality. Because stroke is not this study's primary concern in terms of outcomes – and because the incidence of in-hospital mortality was less than 4% and permanent stroke was even less (2%) – we decided to

focus on the patients who developed delirium and were successfully discharged to either home or another institution. Thus, a new dataset, labeled "Alive" with patients who were successfully discharged from the institution, was created. The "Alive" dataset had a total of 5,584 patients over the study period (January 2006 – December 2012).

B.II. Statistical Analysis Tables of Dataset Attributes

Table B-1: Full Dataset - Statistical comparison of pre-operative continuous attributes (Normal Distribution) in the presence and absence of delirium

ATTRIBUTE (UNIT)	MEASURE	DELIRIUM TOTAL= 5798		TEST STATISTICS	
		Yes (n= 661)	No (n= 5137)	Student t-test t-value	p-value
Age (Year)	Mean (SD)	71 (10)	66 (11)	- 12.38	< 0.05
Weight (Kg)	Mean (SD)	82 (18)	85 (16)	4.4	< 0.05
Height (Cm)	Mean (SD)	170 (10)	169 (10)	0.73	0.466
BMI (Kg/m^2)	Mean (SD)	28.5 (6)	29.4 (6)	4.2	< 0.05
BSA (m^2)	Mean (SD)	1.9 (0.2)	2 (0.2)	3.94	< 0.05
Hemoglobin (mg/dl)	Mean (SD)	125 (19)	132 (18)	9.34	< 0.05
CrCl-CGE (ml/min/1.73 m^2)	Mean (SD)	64 (30)	81 (35)	13.5	< 0.05
CrCl-Jel (ml/min)	Mean (SD)	54 (23)	68 (25)	14.3	< 0.05

Table B-2: Full Dataset - Statistical comparison of pre-operative continuous attributes (Non-Normal Distribution) in the presence and absence of delirium

ATTRIBUTE (UNIT)	MEASURE	DELIRIUM TOTAL= 5798		TEST STATISTICS	
		Yes (n= 661)	No (n= 5137)	Wilcoxon-Mann Whitney test Z-value	p-value
EUROOII (%)	Median (1^{st}, 3^{rd} IQR)	4.4 (2-10)	1.7 (1-4)	17.4314	< 0.05
MHC-Mort (%)	Median (1^{st}, 3^{rd} IQR)	4.4 (2-10)	2 (1-4)	15.7124	< 0.05
Creatinine (µmol/L)	Median (1^{st}, 3^{rd} IQR)	103 (87-133)	93 (79-112)	9.2791	< 0.05

Table B-3: Full Dataset - Statistical comparison of pre-operative categorical attributes in the presence and absence of delirium

ATTRIBUTE (YES)	DELIRIUM (%) TOTAL= 5798		TEST STATISTICS			
	Yes (n= 661)	No (n= 5137)	Chi^2	p-value	Cramer's V	OR (CI)
SDA	38.28	44.56	9.5	<0.05	0.04	0.86 (0.79-0.93)
ACEI	48.56	48.92	0.03	0.86	0.0002	0.99 (0.9-1.06)
BB	77.46	77.50	0	0.98	0	1 (0.9-1.03)
ASA	82.30	82.99	0.2	0.7	0	1 (0.96-1.02)
Plavix	34.80	31.61	2.7	0.09	0.02	1.1 (1.02-1.2)
Lipid Ag.	79.58	81.31	1.15	0.28	0.014	0.98 (0.95-1.01)
Anti-Coag.	11.35	5.68	31.7	<0.05	0.07	2 (1.6-2.4)
HTN	78.97	75.24	4.4	0.04	0.02	1.05 (1.01-1.1)
DM	43.12	36.15	12.2	<0.05	0.05	1.2 (1.1-1.3)
COPD	20.88	13.67	25	<0.05	0.07	1.5 (1.33-1.8)
Frail	10.89	6.52	17.2	<0.05	0.05	1.7 (1.4-2.04)
NYHA Class ≥ 3	55.22	41.93	42.1	<0.05	0.085	1.32 (1.24-1.4)
Shock	4.39	1.60	24.3	<0.05	0.065	2.75 (1.9-3.9)
A-Fib	20.57	11.21	48	<0.05	0.09	1.8 (1.6-2.1)
CVD	21.63	11.66	52.2	<0.05	0.095	1.85 (1.6-2.1)
Renal Impairment	42.06	27.06	64.3	<0.05	0.11	1.6 (1.4-1.7)
IABP	15.58	7.38	52	<0.05	0.094	(1.8-2.5)
Redo	10.89	6.70	15.5	<0.05	-0.05	1.6 (1.3-1.99)

Table B-4: Full Dataset - Statistical comparison of intra-operative continuous attributes in the presence and absence of delirium

ATTRIBUTE (UNIT)	MEASURE	DELIRIUM TOTAL= 5798		TEST STATISTICS	
		Yes (n= 661)	No (n= 5137)	Student t-test t-value	p-value
Clamp Time (Min)	Mean (SD)	93 (45)	83 (39)	- 5.6	< 0.05
Pump Time (Min)	Mean (SD)	139 (67)	123 (52)	- 6.16	< 0.05
Core Temp (°C)	Mean (SD)	31 (2.7)	31.7 (2.6)	5.5	<0.05

Table B-5: Full Dataset - Statistical comparison of intra-operative categorical attributes in the presence and absence of delirium

ATTRIBUTE (YES)	DELIRIUM (%) TOTAL= 5798		TEST STATISTICS			
	Yes (n= 661)	No (n= 5137)	Chi^2	p-value	Cramer's V	OR (CI)
Lt Main Disease	23.9	23.0	0.232	0.63	0.006	1 (0.92-1.2)
CABG	77.0	78.7	1.05	0.31	0.013	0.98 (0.94-1.01)
AVR	35.9	25.7	31	<0.05	0.07	1.4 (1.3-1.5)
MVR	5.6	44.8	5.3	0.02	0.03	1.5 (1.1-1.99)
OR Inotrops	48.1	29.9	89.6	<0.05	0.12	1.6 (1.5-1.7)
Intra-Op TEE	79.4	62.7	70.4	<0.05	0.11	1.26 (1.2-1.3)

Table B-6: Full Dataset - Statistical comparison of post-operative categorical attributes in the presence and absence of delirium

ATTRIBUTE (YES)	DELIRIUM (%) TOTAL= 5798		TEST STATISTICS			
	Yes (n= 661)	No (n=5137)	Chi^2	p-value	Cramer's V	OR (CI)
Blood Product Within 48hrs	53.6	27.5	187.6	<0.05	0.1799	1.95 (1.8-2.1)
pRBC	60.7	29.1	264	<0.05	0.214	2.1 (1.96-2.2)
FFP	24.4	10.3	111	<0.05	0.14	2.4 (2.1-2.7)
Platelets	19.7	9.5	62.7	<0.05	0.104	2.1 (1.8-2.4)
Mechanical Ventilation >24hrs	48.7	14.0	481	<0.05	0.29	3.5 (3.2-3.8)
CVICU Stay >72hrs	56.3	15.6	603	<0.05	0.322	3.6 (3.4-3.9)
New A-Fib	43.0	31.5	35.2	<0.05	0.078	1.37 (1.26-1.48)

Table B-7: Full Dataset - Statistical comparison of post-operative categorical complications in the presence and absence of delirium

ATTRIBUTE (YES)	DELIRIUM (%) TOTAL= 5798		TEST STATISTICS			
	Yes (n= 661)	No (n= 5137)	Chi^2	p-value	Cramer's V	OR (CI)
New A-Fib	43.0	31.5	35.2	<0.05	0.078	1.37 (1.26-1.48)
Tamponade	6.5	2.2	42.8	<0.05	0.09	3 (2.3-4)
Pneumonia	21.6	5.4	231	<0.05	0.201	4 (3.5-4.7)
Pulmonary Edema	22.2	6.8	179.4	<0.05	0.18	3.3 (2.8-3.8)
GI Bleeding	2.6	1.1	9.9	<0.05	0.04	2.3 (1.5-3.6)
Post-Op Dialysis	6.1	2.5	26.8	<0.05	0.07	2.5 (1.8-3.3)
Sepsis	10.1	1.7	159.7	<0.05	0.17	5.9 (4.6-7.65)
UTI	18.0	5.6	137.1	<0.05	0.154	3.2 (2.7-3.7)
Permanent Stroke	3.0	1.8	4.3	0.04	0.03	1.7 (1.1-2.5)
Temporary Stroke	1.8	0.8	6	<0.05	0.03	2.2 (1.3-3.7)
Discharge to Institution	38.1	10.8	365.7	<0.05	0.25	3.54 (3.2.3-3.9)
In-Hospital Mortality	4.1	3.6	0.326	0.57	0.008	1.1 (0.8-1.6)

B.III. Clustering Approaches

In general, clustering approaches are divided into 5 main categories Partitioning, Hierarchical, Density-based, Grid-based, and Sub-space based[64, 97, 186].

- **Partitioning methods:** divides the data into groups such that each group must contain at least one object. These methods usually adopt “exclusive cluster separation”, which means that each object must belong to exactly one and only one group. They are usually Distance-based. It uses an iterative relocation technique to re-assign observations to different groups for the main goal of decreasing the inter-cluster distance between observations from the same class while increasing the intra-cluster distance between different clusters.
- **Hierarchical methods:** find successive clusters using previously established clusters. They create a hierarchical decomposition of the given set of data objects. These algorithms usually are either "Bottom-Up" (agglomerative) or "Top-Down" (divisive). Bottom-Up approaches begin with each element as a separate cluster and merge them into successively larger clusters. Top-Down approaches begin with all of the data and progress to split it into sequentially smaller clusters. They can use distance, density or continuity as a similarity measure.
- **Density-based methods:** the basic idea of using density in clustering is due to that distance based algorithms can only find spherical shapes. In general, the cluster continues to grow as long as the density (number of objects) in the “neighborhood” exceeds some threshold. This way, the cluster can take any shape and different clusters can have different shapes.
- **Grid-based methods:** These methods use a single uniform grid mesh to partition the entire problem domain into cells and the data objects located within a cell are represented by the cell using a set of statistical attributes. Clustering is, then, performed on the grid cells, instead of the database itself. Since the size of the grid is usually much less than the number of the data objects, the processing speed can be significantly improved. However, an over or under saturated grid will require further analysis. Several authors advocate the use of grid-based methods as an

initial step and augmenting the analysis of the saturated grids with another clustering method (Density or Distance based).

- **Sub-Space methods:** look for clusters that can only be seen in a particular projection (subspace, manifold) of the data. These methods thus can ignore irrelevant attributes. They can be combined with any of the other types of clustering methods and focus on a particular dataset projection.

B.IV. Expectation Maximization Algorithm

The EM algorithm is a Probabilistic algorithm that belongs to the partitioning clustering methods. It attempts to construct a "latent" attribute that can be used to maximize the likelihood estimate of the model[64, 115]. This new "latent" attribute is created using available observations. EM is considered to be a natural generalization of maximum likelihood estimation to the incomplete data case. One of the assets of these models is that they are capable of approximating parameters estimates in the presence of incomplete data. Ceppellini et al first introduced it in 1955[187].

As the name suggests, the EM algorithm has 2 steps: The first step, also called E-Step from "Expectation", entitles the calculation of the expected class probabilities values of the missing or latent attribute from the available observed data. This is followed by the second step, also called M-Step from "Maximization", that allows the augmentation of the log-likelihood function by re-estimating the expected obtained values from the E-Step, under the assumption that all missing values have been replaced in the E-Step and there are no missing observations[64-66, 115]. The EM algorithm is iterative in nature, in the sense that E- and M-steps are alternated until the changes in the estimated parameters or the log likelihood are less than some specified threshold.

Compared to other partitioning methods, the EM algorithm provides the allocation probabilities of an observation to a particular cluster[114]. In other words, an observation can be a member in multiple clusters with a membership potential assignment to each cluster, expressed in probability. At the end, an observation gets assigned to a specific cluster based on the greatest probability value of it has (e.g.: Observation x has a membership in Cluster-A with a probability=0.25, Cluster-B with a probability=0.6, and

Cluster-C with a probability=0.15; in the final output, it is assigned to Cluster-B). The beauty of this technique is that the analyst will be able to review the membership probabilities of an observation to multiple clusters. EM algorithm can be applied to categorical and continuous attributes, unlike the classic implementation of k-means that can only accommodate continuous attributes[114].

The EM algorithm will converge; but usually this is not the best possible solution. To obtain better results and reach a superior answer, the whole process needs to be repeated several times, with different initialization. The overall log-likelihood is used to compare several attained patterns and the best one is chosen biased on the model fit and its representation[64-66, 188].

The number of generated clusters can be either pre-determined based on some apriori domain knowledge, but in reality the optimal number is not known. There are several techniques that are available for the analyst based on heuristics, statistical, information theory or other methods. One common and effective technique is to use “v-Fold Cross-Validation”[141, 189]. In general, it starts by dividing the overall sample into a number of v folds. The same type of analysis is then continuously applied to the observations belonging to the v-1 folds (training sample), and the results of the analyses are applied to sample v (the sample or fold that was not used to estimate the parameters to determine the clusters; testing sample) to compute some index of predictive validity (Log-Likelihood). The results for the v replications are averaged and return a single model performance assessment measure (In the case of EM, cluster assignment likelihood). The use of v-Fold Cross-Validation sometimes generate un-realistic number of clusters or assigns a very small portion of observations to a specific irrelevant cluster (Over fit the clusters). To overcome this issue, a reasonable common approach is start with v-Fold Cross-Validation, analyze the clusters, then try to come us with an optimal number of clusters without dramatically affecting model performance assessment measure, the “Elbow method”[190].

Cluster analysis is an unsupervised learning technique, and we cannot observe the (real) number of clusters in the data. However, it is reasonable to replace the usual notion (applicable to supervised learning) of "accuracy" with that of "distance." In general, we can apply the v-fold cross-validation method to a range of numbers of clusters in k-means

or EM clustering, and observe the resulting average distance of the observations (in the cross-validation or testing samples) from their cluster centers (for k-means clustering); for EM clustering, an appropriate measure would be the average (log-likelihood) computed for the observations in the testing samples. In the case of log-likelihood the closer your log-likelihood is to 0 the better the clustering assignment (range= -∞:0).

B.V. Expectation Management Algorithm Clustering Experiments

B.V.i. ACEI and ARB Clustering:

Number of clusters selected by cross validation: 3

Attribute	*Cluster* "0" N= 2738 (49%)	"1" N= 2171 (39%)	"2" N= 675 (12%)
ACEI			
• Yes	2737.0234	1.9994	1.9772
• No	1.9993	2171.0394	675.9612
ARB			
• Yes	93.0582	2.0005	675.9413
• No	2645.9646	2171.0384	1.9971
Total	2739.0227	2173.0389	677.9384

Log likelihood: -1.04486

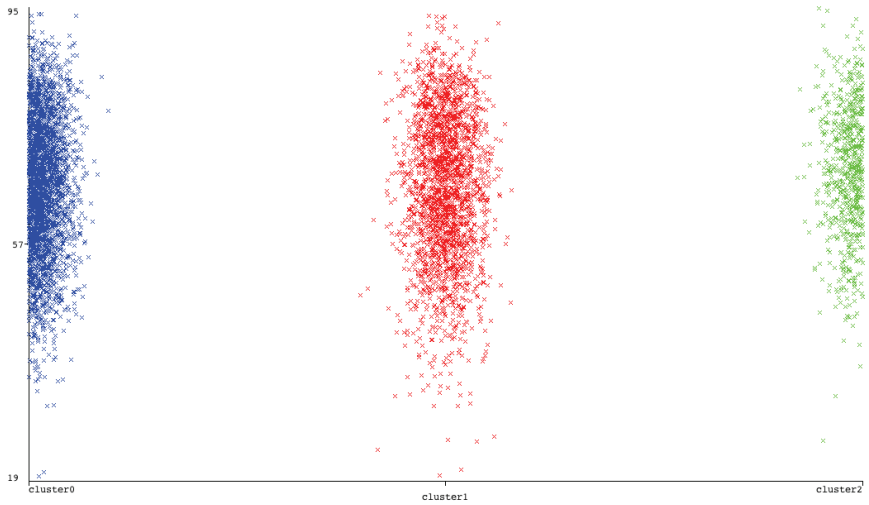

Figure B-5: ACEI-ARB Clusters by Age
X-Axis: ACEI-ARB Clusters, Y-Axis: Age in years

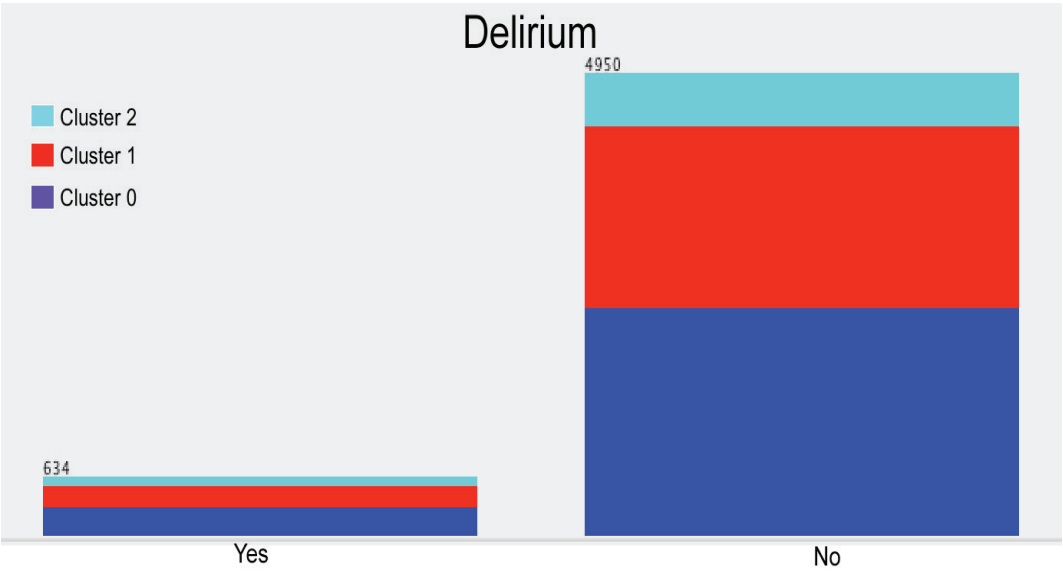

Figure B-6: ACEI and ARB Clusters in Delirium

Although this clustering did not give us any insight, it showed a clear separation that is representative of current clinical practice. We can see, Cluster "0" had mainly patients who are mainly taking ACEI, Cluster "1" had patients who are on none and cluster "2" had patients who are mainly on ARB. Clinically, we usually start with ACEI for the treatment of HTN or HF and switch to ARB if a patient start developing side effects of ACEI (e.g: dry cough or angioedema). This demonstrates that the EM algorithm is capable of identifying a clear existing pattern in the data with minimal input from the data analyst.

B.V.ii. Aspirin and Lipid Lower Agents Clustering:

Number of clusters selected by cross validation: 2

	Clusters	
Attribute	"0" N= 4540 (81%)	"1" N= 1044 (19%)
Lipid Agent		
• Yes	4540.0027	1.9973
• No	1.9993	1044.0007
ASA		
• Yes	4034.3186	595.6814
• No	507.6834	450.3166
Total	4542.002	1045.998

Log likelihood: - 0.89411

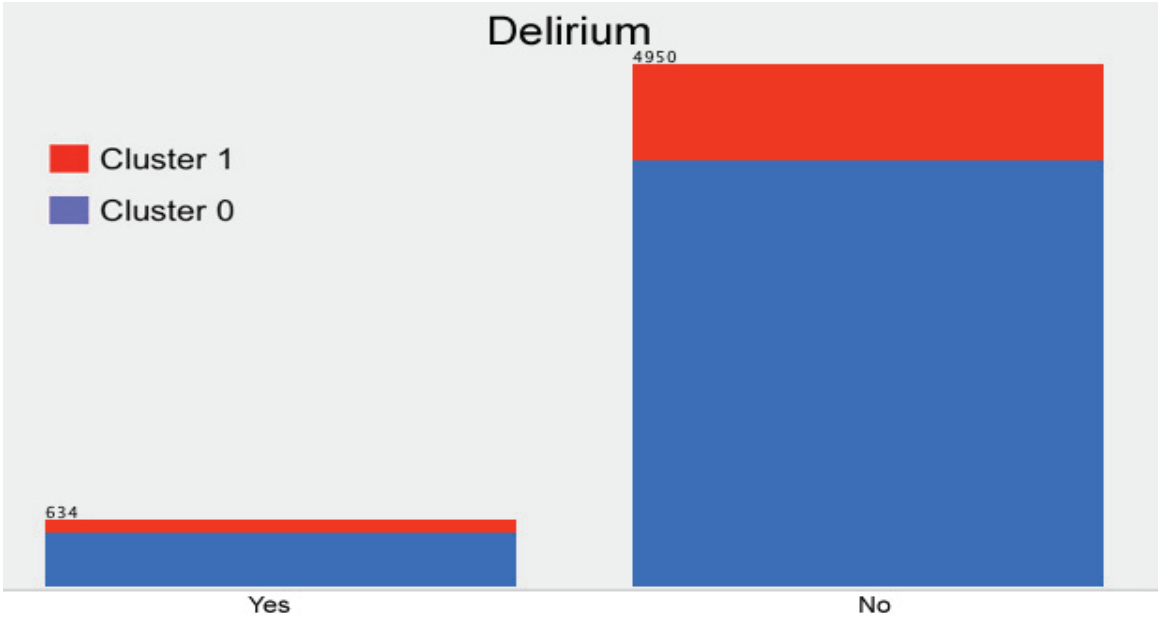

Figure B-7: ASA and Lipid Lowering Agent Clusters in Delirium

Here the algorithm identified a group of patients who take ASA and a lipid-lowering agent in Cluster "0", which represents patients with Ischemic Heart Disease (IHD). What's interesting is that 57% of patients in Cluster "1" were on ASA and almost none were on lipid-lowering agent, which most probably represent pure Valvular Heart Disease patients who are mainly young or with no clinical indication of atherosclerosis.

B.V.iii. DM and DM Control Clustering:

Number of clusters selected by cross validation: 2

Attribute	*Clusters* "0" N= 3556 (64%)	"1" N= 2028 (36%)
DM		
• Yes	19.113	2029.887
• No	3537.9995	1.0005
DM Control		
• None	3556.0957	1.9043
• Diet	1.0056	1025.9944
• Oral	1.0056	294.9944
• Insulin	1.0056	709.9944
Total	3559.1125	2032.8875

Log likelihood: - 1.03753

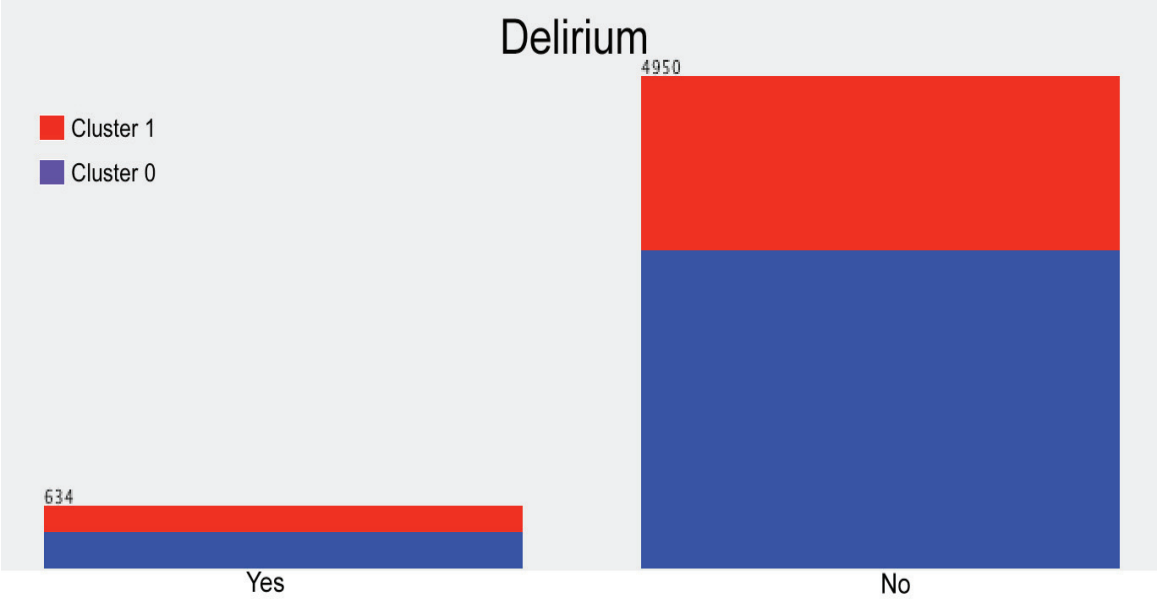

Figure B-8: DM and DM Control Clusters in Delirium

Here, the algorithm wanted to split patients to diabetics and non-diabetics. The treatment modality did not have an influence the clustering.

B.V.iv. Smoking History and Current Smoker Clustering:

Number of clusters selected by cross validation: 2

Attribute	*Clusters* "0" N= 3894 (70%)	"1" N= 1690 (30%)
Smoking History		
• Yes	3894.0011	1.9989
• No	1.7848	1690.2152
Current Smoker		
• Yes	838.9994	1.0006
• No	3056.7865	1691.2135
Total	3895.7859	1692.2141

Log likelihood: - 0.9765

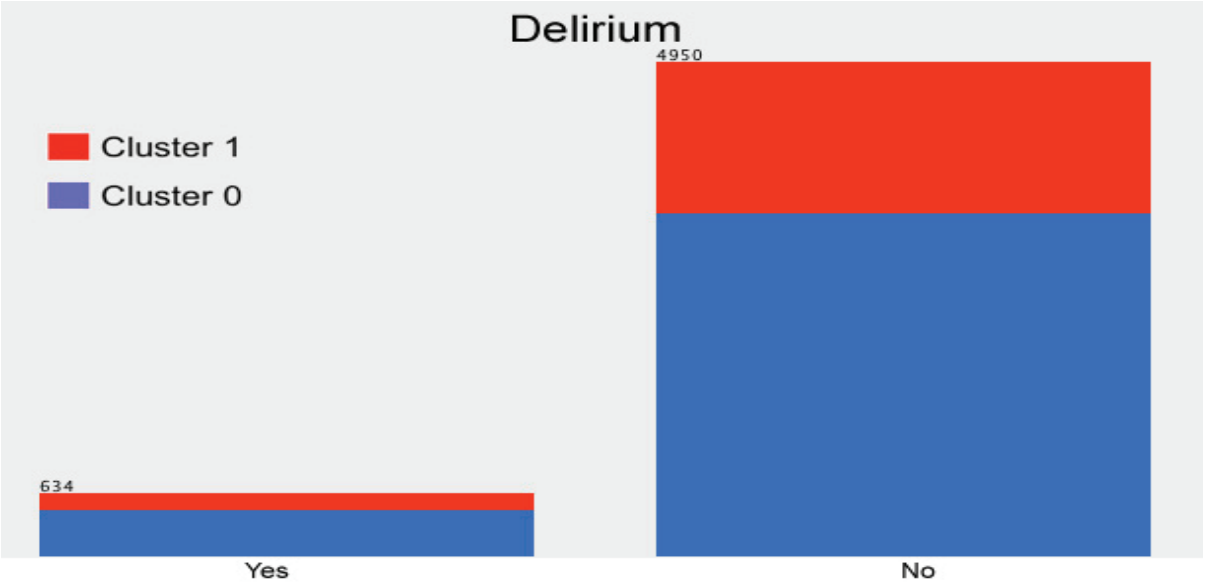

Figure B-9: Smoking History and Current Smoking Clusters in Delirium

Here, the algorithm has decided to split the patients based on their history of smoking. 22% of patients who had history of smoking are still smoking. Although it dose not provide much insight, this information can be used to focus some effort on providing a personalized smoke cessation program to this group of patients that will help in their secondary prevention.

B.V.v. Pre-operative Arrhythmia Clustering:

Number of clusters selected by cross validation: 2

Attribute	*Clusters* "0" N= 1266 (23%)	"1" N= 4318 (77%)
Ventricular Arrhythmia		
• Yes	60.0198	124.9802
• No	1207.9774	4195.0226
Atrioventricular Block		
• Yes	134.045	253.955
• No	1133.9522	4066.0478
Complete Heart Block		
• Yes	43.0229	43.9771
• No	1224.9743	4276.0257
Other		
• Yes	1265.9997	2.0003
• No	1.9975	4318.0025
Total	1267.9972	4320.0028

Log likelihood: - 1.00358

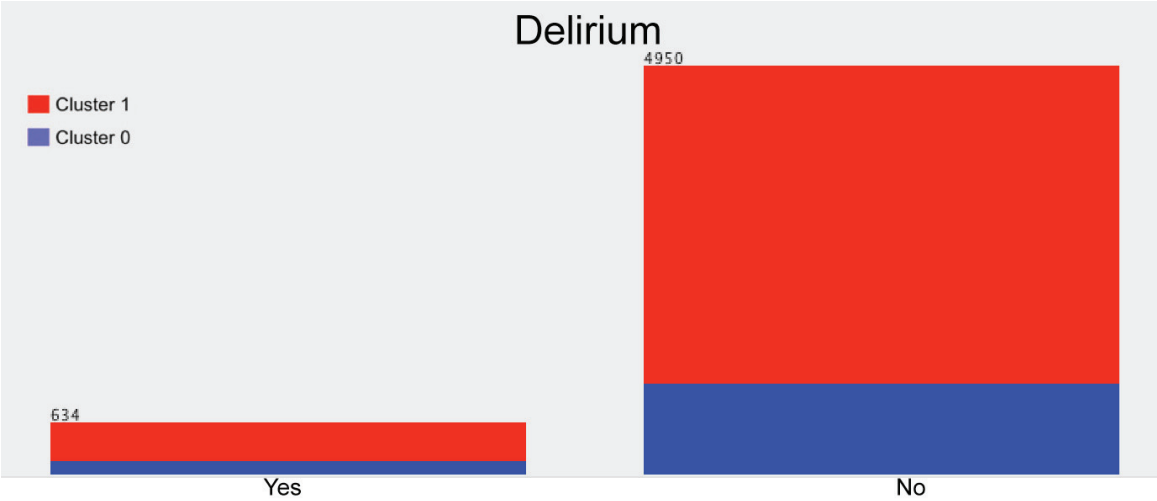

Figure B-10: Pre-Operative Arrhythmia Clusters in Delirium

Here, the algorithm has decided to split the patients into two groups with almost double the representation of arrhythmias in the cluster "0" patients.

B.V.vi. Beta-Blockers and Calcium Channel Antagonists (Ca-Ant)

Clustering:

Number of clusters selected by cross validation: 2

Attribute	*Clusters* "0" N= 3931 (70%)	"1" N= 1653 (30%)
Beta-Blockers		
• Yes	3019.9711	1320.0289
• No	913.0295	334.9705
Ca-Ant		
• Yes	1.9994	1653.0006
• No	3931.0012	1.9988
Total	3933.0006	1654.9994

Log likelihood: - 1.13776

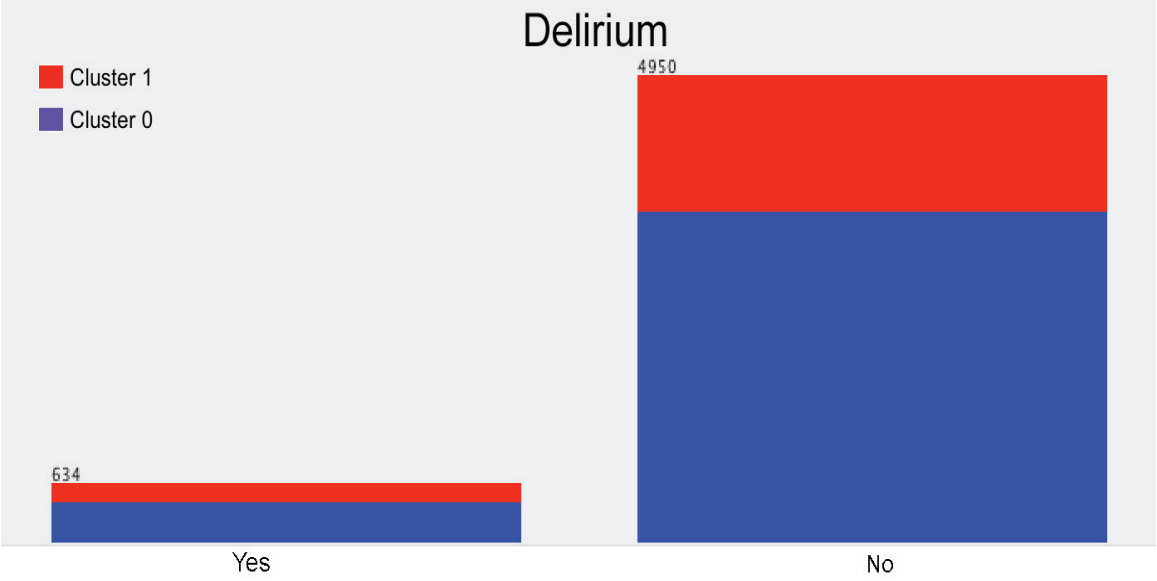

Figure B-11: Beta-Blockers and Calcium Channel Antagonists Clusters in Delirium

Here, the algorithm was able to identify two distinct groups. Cluster "0" represents the patients who are only on Beta Blockers and Cluster "1" represents the patients who are on Beta Blockers and Calcium Channel Blockers. Clinically, Cluster "1" can illustrates the patients who are on combination therapy for resistant HTN.

B.V.vii. Atrial Fibrillation and its Pattern Clustering:

Number of clusters selected by cross validation: 4

Attribute	*Cluster* "0" N= 4926 (88%)	"1" N= 380 (7%)	"2" N= 162 (3%)	"3" N= 116 (2%)
Pre-Op A-Fib				
• Yes	2.0021	380.0229	162.9935	116.9815
• No	4925.3694	1.966	1.3325	1.3321
Paroxysmal				
• Yes	1.673	1.9823	162.3363	1.0084
• No	4925.6985	380.0066	1.9897	117.3052
Persistent				
• Yes	1.0002	9.9869	1.0054	1.0075
• No	4926.3713	372.002	163.3206	117.3061
Permanent				
• Yes	1.6756	1.9893	1.0061	116.3289
• No	4925.6958	379.9996	163.3199	1.9847
Total	4927.3715	381.9889	164.326	118.3136

Log likelihood: - 0.48894

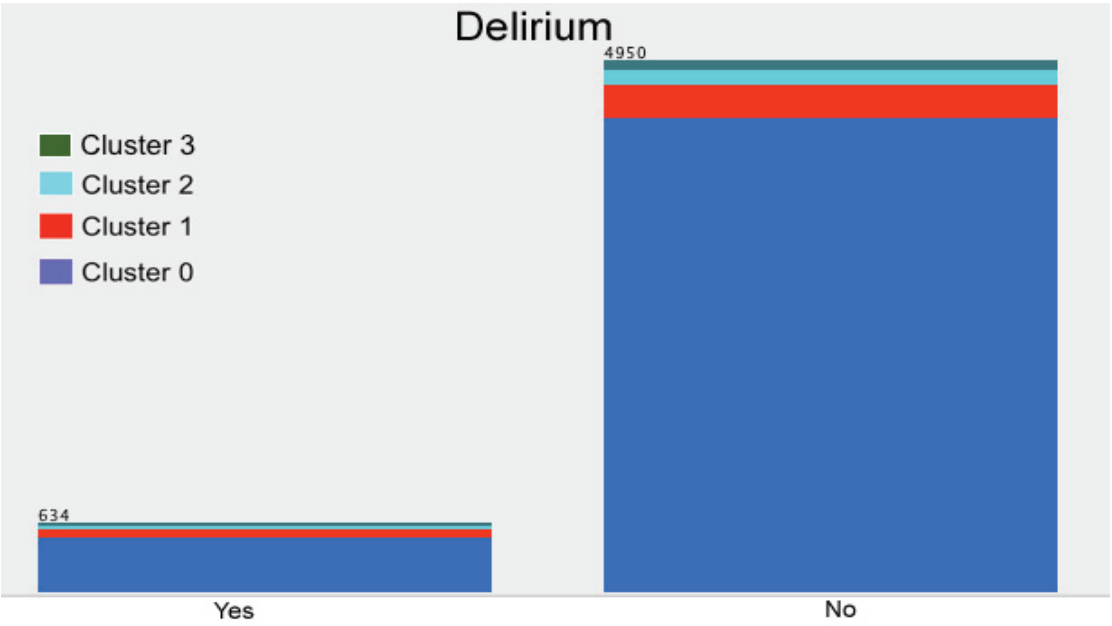

Figure B-12: Atrial Fibrillation and its pattern Clusters in Delirium

In this case, the EM algorithm did not offer any help. Although, the Log-Likelihood was very good and the clusters are clinically relevant; what it actually did is just create a new

attribute that has 4 categories: Cluster "0"= no A-Fib, Cluster "1"= Paroxysmal A-Fib, Cluster "2"= Persistent A-Fib, and Cluster "3"= Permanent A-Fib. A clinician could have done this and it adds no new insight.

B.VI. Manual Selection based on statistical analysis

We applied the same principles that we used to test the full dataset. All date attributes were used to generate length of stay continuous attributes. Statistical measures of central tendency (Mean, Median, Mode, Inter-quartile ranges, standard deviation and others), shape of distribution and outliers were examined. Continuous attributes were examined and when the shape of the distribution was of close to a normal (Gaussian) distribution; mean and standard deviations were used. When the shape was extremely skewed (non-normal distribution); Median, 25% Inter-quartile range (IQR) and 75% IQR were used. Categorical attributes are reported with frequency in percent.

Continuous attributes with normal distribution were tested using the student t-test for 2 levels and Analysis of Variance (ANOVA) for more than 2 levels. Continuous attributes with non-normal distribution were tested using the Wilcoxon-Mann Whitney test for 2 levels and Kruskal Wallis test for ordinal attributes. Categorical attributes were examined using the Chi-Square test. Table B-8 summarizes the normally distributed attributes that were identified as important. We can see that the Age, pre-operative hemoglobin level and pre-operative creatinine clearance had a strong influence on delirium. Although, BMI was statistically significant, it had a very small effect (R2= 0.0023) and we choose to exclude it from the analysis. We also can see that both formulas used to calculate creatinine clearance showed that pre-operative calculated renal function has a significant influence on delirium; we choose to use the Jelliffe formula results.

Table B-8: Alive Dataset - Statistical Comparison of Preoperative Continuous (Normal Distribution) Attributes in the Presence and Absence of Delirium

ATTRIBUTE (UNIT)	MEASURE	DELIRIUM TOTAL= 5584		TEST STATISTICS	
		Yes (n= 634)	No (n= 4950)	Student t-test t-value	p-value
Age (Year)	Mean (SD)	72 (10)	66 (11)	13.06	< 0.05
BMI (Kg/m^2)	Mean (SD)	28.5 (6)	29.4 (6)	3.78	< 0.05
Hemoglobin (mg/dl)	Mean (SD)	126 (19)	133 (18)	8.63	< 0.05
CrCl-CGE (ml/min/1.73 m^2)	Mean (SD)	64 (30)	82 (34)	19.6	< 0.05
CrCl-Jel (ml/min)	Mean (SD)	54 (22)	68 (25)	14.5	< 0.05

When examining the non-normally distributed pre-operative attributes in Table B-9, both the EUROII and MHC scores were both significantly associated with delirium. In this analysis, the EUROII score was chosen as it is a more commonly used and widely available.

Table B-9: Alive Dataset - Statistical Comparison of Preoperative Continuous (Non-Normal Distribution) Attributes in the Presence and Absence of Delirium

ATTRIBUTE (UNIT)	MEASURE	DELIRIUM TOTAL= 5584		TEST STATISTICS	
		Yes (n= 634)	No (n= 4950)	Wilcoxon-Mann Whitney test z-value	p-value
EUROOII (%)	Median (1^{st}, 3^{rd} IQR)	4.2 (2-10)	1.6 (1-3.5)	17.9	< 0.05
MHC-Mort (%)	Median (1^{st}, 3^{rd} IQR)	4.3 (2-9)	2 (1-4)	16.1	< 0.05
Creatinine (μmol/L)	Median (1^{st}, 3^{rd} IQR)	103 (87-133)	92 (79-110)	9.5	< 0.05

Table B-10: Alive Dataset - Statistical Comparison of Categorical Attributes in the Presence and Absence Of Delirium

ATTRIBUTE (YES)	DELIRIUM (%) TOTAL= 5584		TEST STATISTICS			
	Yes (n= 634)	No (n= 4950)	Chi^2	p-value	Cramer's V	OR (CI)
Male	76.3	74.2	1.3	0.25	0.015	1.02 (0.98-1.07)
Anti-Coag.	11.5	5.6	34.1	<0.05	0.08	2.1 (1.7-2.5)
HTN	78.7	75.3	3.5	0.06	0.03	1.05 (1.01-1.1)
COPD	16.4	10.7	18.8	<0.05	0.06	1.5 (1.3-1.8)
Frail	10.1	6.1	14.5	<0.05	0.05	1.6 (1.3-2)
CHF	27.1	15.2	58.5	<0.05	0.10	1.8 (1.6-2)
A-Fib	20.3	10.7	50.5	<0.05	0.10	1.9 (1.6-2.2)
CVD	21.5	11.3	53	<0.05	0.10	1.9 (1.6-2.2)
Renal Impairment	41.5	25.8	69.4	<0.05	0.112	1.6 (1.5-1.8)
IABP	15.0	5.4	86.4	<0.05	0.124	2.8 (2.3-3.4)
Blood Products ≤48 hrs	52.8	25.4	207	<0.05	0.193	2.1 (1.9-2.2)
Ventilation >24 hrs	46.8	11.7	527	<0.05	0.31	4 (3.7-4.4)
LCOS	23.5	5.5	264	<0.05	0.22	4.3 (3.7-5)
Redo	11.2	6.2	22	<0.05	0.06	1.8 (1.5-2.2)

Table B-11: Alive Dataset - Statistical Comparison of Ordinal Attributes in the Presence and Absence of Delirium

ATTRIBUTE	Cramer's V	KRUSKAL-WALLIS TEST Chi^2	p-value
CCS	0.068	1.3	0.19
NYHA	0.125	42.2	<0.05
CVICU length of stay categories	0.34	472.2	<0.05
KATZ Index	0.03	3.05	0.08
EF categories	0.113	69.9	<0.05
Number of Distal Anastomosis	0.039	1.6	0.2
Number of IMA Distal Anastomosis	0.066	17.8	<0.05

Table B-12: Alive Dataset - Clustered Attributes Association Coefficients in the Presence Of Delirium

ATTRIBUTE	CRAMER'S V
ACEI	0.0021
ARB	0.034
Clustered ACEI-ARB	0.041
BB	0.0001
CA-Antagonist	0.025
Clustered BB-CA Antg.	0.025
ASA	0.005
Lipid Lower Agent	0.02
Clustered ASA-Lipid lowering Agent	0.02
DM	0.044
DM Control	0.045
Clustered DM-DM Control	0.043
Smoking History	0.023
Current Smoker	0.02
Clustered Smoking History & Current	0.023
Clustered Pre-op Arrhythmia	0.039

B.VII. Feature Selection Methods

B.VII.i. BestFirst:

Searches the space of attribute subsets by greedy hill climbing augmented with a backtracking facility. Setting the number of consecutive non-improving nodes allowed controls the level of backtracking done. The direction of search can be forward (empty set and then add), backward (full set and eliminate) or bidirectional (start with a random subset and search in both directions).

In this work, the bidirectional approach was chosen with a search termination of 8 attributes and did not specify a start set.

B.VII.ii. Ranker:

Ranks attributes by their individual evaluations.

The default setting was used and the algorithm was allowed to give a complete ranking of the 92 attributes

B.VII.iii. CfsSubsetEval:

Evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low Inter-co-relation are preferred.

Options:

- Debug: *False*
- Locally Predictive: *True* (Identify locally predictive attributes. Iteratively adds attributes with the highest correlation with the class as long as there is not already an attribute in the subset that has a higher correlation with the attribute in question)
- Number of Threads: *1* (The number of threads to use)
- Pre-Compute Correlation Matrix: *False* (Pre-compute the full correlation matrix at the outset, rather than compute correlations as needed during the search)
- Missing Separate: *False* (Treat missing as a separate value)
- Pool Size: *1* (The size of the thread pool)

B.VII.iv. ReliefFAttributeEval:

This method assesses the worth of an attribute by recursively sampling an observation and deliberating the value of the given attribute for the nearest observation of the same and different class. It is highly tolerant of noise, can easily detect feature interactions, and is applicable to both binary and continuous data. However, it cannot discriminate correlated attributes and is affected by the class imbalance.

Options:

- Number of Neighbors: *10* (Number of nearest neighbors for attribute estimation)
- Sample Size: *-1* (Number of instances to sample. -1: indicates that all instances will be used for attribute estimation)

- Seed: *1*
- Sigma: *2* (Set influence of nearest neighbors)
- Weight By Distance: *False* (Weight nearest neighbors by their distance)

B.VII.v. GainRatioAttributeEval:

Measuring the Information gain ratio with respect to the class establishes the influence of the attribute.

$$Gain\ Raio(Class, Attribute) = \frac{(P(Class) - P(Class|Attribute))}{P(Attribute)}$$

Options:

- Missing Merge: *True* (Distribute counts for missing values)

B.VII.vi. SignificanceAttributeEval:

Calculates the probabilistic influence of an attribute as a two-way function.
(Attribute←→Classes association)

Options:

- Missing Merge: *True* (Distribute counts for missing values)

B.VII.vii. SymmetricalUncertAttributeEval:

Assess the impact of an attribute symmetrical uncertainty with respect to the class.

$$Symmetrical\ Uncertinty(Class, Attribute) = 2\times\frac{P(Class) - P(Class|Attribute)}{P(Class) + P(Attribute)}$$

Options:

- Missing Merge: *True* (Distribute counts for missing values)

APPENDIX C: MODELING

C.I Hold-Out or Cross-validation

Several methods are used to test a model performance (e.g.: Cross-Validation, Leave one out, Bootstrapping, Random sub-sampling and Hold-Out). The choice of the method is primarily based on the amount of available data. The goal is to maximize the amount of data available for training to support the algorithm learning process. Cross-Validation starts by dividing the data set in "k" mutually exclusive partitions. These "k" partitions will be the test sets for each of the k models that will be built with the remaining data. The Hold-out method will split the original full data set into two partitions, one for training and another for testing. It is done by randomly selecting observations from the full dataset that are not used in any way in the building of the model. In Cross-Validation several models are built, tested and their performance averaged; in the Hold-out method a single model is obtained and this model is tested on a single test set. Table C-13 illustrates some of the pros and cons of the Hold-out method and Cross validation.

The cost of the holdout method comes in the amount of data kept out for testing. If the model is trained on a small dataset, its error will be exaggerated. Such pessimistic predictions are always better in real life implementations in comparison to an over confident ones[128]. One of the major advantages of Hold-Out sample is that it emulates reality (predicting the future/unknown based on the past/known) and doesn't minimize the effect of uncertainty[191]. The Hold-out method is usually preferred when the dataset is large. The definition of large is not really solid but most authors prefer the use of this method if you have more than 1000 observations in your dataset[56, 64]. In simple Hold-Out, some of the attributes - mainly the prediction class - may exhibit a different distribution across datasets (training versus testing). To avoid this, experts encourage stratification (Stratified Hold-Out sample), this ensures that most attributes - mainly the outcome/class – has a comparable distribution across datasets[65].

The Stratified Hold-Out method was used to ensure equal representation of essential attributes across the datasets (Table 4-1). An 80/20 split was used, 80% for training and 20% for testing (Figure 4-1). This split will maximize the amount of available data for the learning step. That resulted in a training dataset based on 4467 patients and a testing set of 1117 patients. To address the issue of improper representation of the problem space due to random sampling, we made sure that both datasets have a similar distribution of several clinically relevant attributes. By doing so, we assure the generalizability and applicability of our results to the general population. Table 3-4 demonstrates some of the characteristics of the training and test subsets. This final datasets (Training and Test) had a list of 22 candidate attributes (Table C-14).

Table C-13: Pros and Cons of Hold-Out and Cross-validations methods

	Hold-Out	Cross-validation
Pros	• No parametric assumptions • Highly accurate in large sample • Easy Implementation • Simple	• No parametric assumptions • Better if sample size is an issue • Utilizes all observations • Simple
Cons	• Overly cautious • If not applied correctly, results are at risk of contamination • Empirical choice of set size • Some of the observations are not used for learning	• Computationally expensive • Empirical choice of fold size • Somewhat cautious • Error is an average

Table C-14: Final Alive Datasets (Training and Testing) and Attributes Missing Values Frequency Counts

Attribute	Type	Possible Values	Missing	
			Training	Test
			N= 4467	N= 1117
Length of Stay In the ICU	Ordinal	<24, 24-72, >72 hrs	0	0
Prolonged Ventilation	Categorical	Yes/No	0	0
EUROII Score	Continuous	0-100 %	0	0
Age	Continuous	19-95 years	0	0
Procedure Difficulty	Categorical	Single/Combined	0	0
Blood Product Within 48 hrs	Categorical	Yes/No	0	0
Intra-Operative TEE	Categorical	Yes/No	8	8
Timing of IABP	Ordinal	None, Pre-, Intra-, Post-Operative	0	0
Intra-Op Inotrops	Categorical	Yes/No	1	1
Pre-Op Creatinine Clearance	Continuous	4-203 ml/min	13	1
EF categories	Ordinal	<30 30-50, >50 %	19	6
Pre-Op Hemoglobin	Continuous	10-196 mg/dl	9	4
Pre-Op A-Fib	Categorical	Yes/No	0	0
AS	Ordinal	None, Trivial, Mild, Moderate, Critical	106	35
CVD	Categorical	Yes/No	0	0
Clustered DM	Categorical	Cluster 1 or Cluster 2	0	0
Frail	Categorical	Yes/No	0	0
History of Turn Down	Categorical	Yes/No	0	0
MR	Ordinal	None, Trivial, Mild, Moderate, Sever	107	36
Clustered Arrhythmia	Categorical	Cluster 1 or Cluster 2	0	0
COPD	Categorical	Yes/No	0	0
Gender	Categorical	Male/Female	0	0
Delirium	Categorical	Yes/No	0	0

C.II Fundamental Terms Definitions

Most of these definitions were obtained from these resources[53, 54, 60, 61, 63-68, 82, 93, 124, 125, 127, 129, 130, 133, 172, 176, 182, 192-195].

True Positive (TP): the number of observations correctly labeled as belonging to the positive class.

False Positive (FP): the number of observations in-correctly labeled as belonging to the positive class.

True Negative (TN): the number of observations correctly labeled as belonging to the negative class.

False Negative (FN): the number of observations in-correctly labeled as belonging to the negative class.

Confusion Matrix = Contingency Table: is a specific table layout that allows visualization of an algorithm performance. Each column of the matrix represents the instances in a predicted class (by the algorithm), while each row represents the instances in an actual class (from the real data). The name stems from the fact that it makes it easy to see if the system is confusing two classes (i.e. commonly mislabeling one as another).

Table C-15: Confusion Matrix Example

		Predicted Class		
		+	-	
Actual Class	+	**True Positive**	**False Negative**	**Sensitivity (Recall)**
	-	**False Positive**	**True Negative**	**Specificity**
		PPV (Precision)	**NPV**	**Accuracy**

Predictive Accuracy = Recognition Rate = Accuracy: represents the test performance on the entire data irrespective of the classes or their distribution. It relates to a test ability to identify observations assignment correctly.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} = \frac{TP + TN}{Total}$$

True Positive Rate (TPR) = Recall = Sensitivity: relates to a test ability to identify observations with a condition correctly. How inclusive is the test?. In a medical context, it is the probability that a test will be positive in an individual that actually have or will develop the disease.

$$\text{True Positive Rate (TPR)} = \text{ Recall } = \text{ Sensitivity} = \frac{TP}{TP + FN}$$

True Negative Rate (TNR) = Specificity: relates to a test ability to identify observations without the condition correctly. In a medical context, it is the probability that a test will be negative in an individual that actually does not have or will not develop the disease.

$$\text{True Positive Rate (TNR)} = \text{ Specificity} = \frac{TN}{TN + FP}$$

Positive Predictive Value (PPV) = Precision: is the degree of correctness. It is the percentage of actually positive observations that were considered by the test as positive. In a medical context, it is the probability that a patient has the disease if the test is positive.

$$\text{Positive Predictive Value (PPV)} \ = \ \text{Precision} = \frac{TP}{TP + FP}$$

Negative Predictive Value (NPV): It is the percentage of actually negative observations that were considered by the test as negative. In a medical context, it is the probability that a patient does not have the disease if the test is negative.

$$\text{Negative Predictive Value (NPV)} \ = \frac{TN}{TN + FN}$$

Kappa Statistic: is a measure that compares observed accuracy with expected accuracy (random chance)[134].

$$\text{Kappa Statistic } = \frac{Observed\ Accuracy - Expected\ Accuracy}{1 - Expected\ Accuracy}$$

It is used to evaluate a single model or evaluate different models that were used on the same data. In addition, it takes into account random chance (agreement with a random model). It is often used as a measure of reliability two or more independent observers are evaluating the same thing[135]. In machine learning, one rater is the "actual truth" (the actual values of each instance to be classified), obtained from the data, and the other rater is the algorithm used to perform the classification[64-66, 134]. The Kappa statistic is standardized to lie on a scale of (-1,1). A value of 1 indicates perfect agreement, 0 is exactly what is expected if it was due to a stochastic process (random chance), and negative values indicate potential systematic disagreement (Figure C-13)[135].

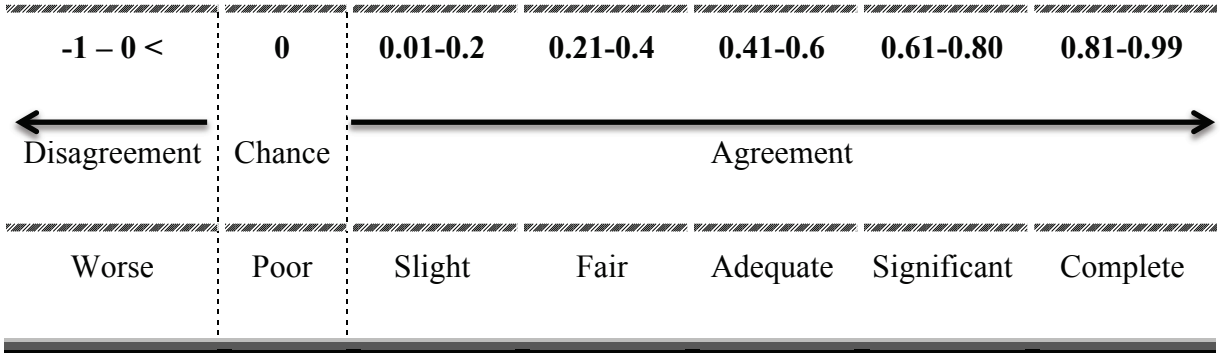

Figure C-13: Kappa Statistic Interpretation Scale

F-measure = F_1 = F-score: is defined as a harmonic mean of precision and recall[64, 65].

$$\text{F} - \text{measure} = F_1 = \text{F} - \text{score} = \frac{2 \times Precision \times Recall}{Precision + Recall} = \frac{2 \times PPV \times \text{Sensitivity}}{PPV + Sensitivity}$$

Is the harmonic mean better than the arithmetic mean?[61, 66] When comparing ratios, the harmonic mean gives a more realistic picture of the true mean compared to the arithmetic mean. The harmonic mean (H_{mean}) is one of the three Pythagorean means (arithmetic, harmonic and geometric). The arithmetic mean (A_{mean}) is basically a summation values (average) over their *n* values. The geometric mean (G_{mean}) specifies the

central tendency of a set of numbers by using the product of their values. The G_{mean} is used to when comparing multiple sets that have different ranges, it "normalizes" the ranges being averaged, so that no range dominates the weighting. Since the H_{mean} of a list of numbers tends strongly toward the least elements of the list, compared to the arithmetic mean, it tends to diminish the influence of outliers and augment the impact of small ones[136].

$$A_{mean} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$G_{mean} = \sqrt[n]{x_1 x_2 \ldots x_n} = \left(\prod\nolimits_{i=1}^{n} x_i\right)^{1/n}$$

$$H_{mean} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

Suppose we have a system that has a sensitivity/recall= 0.95 and a positive predictive value (PPV)/precision= 0.25. The A_{mean} = 0.6, G_{mean} = 0.5, and the H_{mean} = 0.4.

$$A_{mean} = \frac{Precision + Recall}{2} = \frac{0.25 + 0.95}{2} = 0.6$$

$$G_{mean} = \sqrt[2]{Precision + Recall} = \sqrt[2]{0.25 + 0.95} = 0.49$$

$$H_{mean} = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} = \frac{2 \times Precision \times Recall}{Precision + Recall} = \frac{2 \times 0.25 \times 0.95}{0.25 + 0.95} = 0.4$$

As we can see from the example, the H_{mean} is a more practical score that have a more precise and vigilant representation of the system performance, demonstrating the worst case scenario. The A_{mean} is overoptimistic and the G_{mean} is intermediate.

$$min \leq H_{mean} \leq G_{mean} \leq A_{mean} \leq max$$

Receiver Operator Characteristics Curve [126, 129, 130, 132, 196]: developed in the 1950's after World War II as a by-product of research into making sense of radio signals contaminated by noise. So called because radio receiver operators (Electrical and Radar Engineers) invented them after the attack on Pearl Harbor to determine how the US radar had failed to detect the Japanese aircraft. It is a two dimensional graph that illustrates the performance of a model by plotting the False Positive Rate (1-specificity) on the x-axis against the True Positive Rate (sensitivity) on the y-axis for all potential points. In doing so, it demonstrates the trade-off between true and false alarm rates. They allow visualization of performance over a spectrum of different conditions instead of just relying on a point estimate (e.g.: Accuracy)[131].

Figure C-14 illustrates several examples of ROC curves. Curves A and B represents the characteristics of tests more typically seen in routine clinical use. Curve C represent a curve that is ideal or perfect, with a very high sensitivity and specificity. The AUC represents the overall accuracy of a test, with a value approaching 1.0 indicating a high sensitivity and specificity. The line of zero discrimination indicates the test is no better than chance or random guessing, which is equal to 50% or an AUC of 0.5. AUC has been related to the Wilcoxon statistic. Wilcoxon statistic has been defined as an estimate of 'true' area under the ROC curve, area constructed from an infinite sample[131].

Most statistical books and machine learning texts grade the general performance of a model based on the AUC that it covers. There are 6 categories: AUC 0.9-1 is considered "Excellent," AUC 0.8-0.9 is "Good," AUC 0.7-0.8 is "Fair," AUC 0.6-0.7 is "Poor," AUC 0.5-0.6 is "Fail", and AUC < is "Undesirable." The last category indicates that the new test performs worse than random guessing and can be even harmful. An AUC of a classier is equivalent to the probability that the classier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. It is also considered to be non-parametric since it has similar statistical properties to the Wilcoxon test of ranks[127, 130, 132].

When comparing "*n"* ROC-AUC of several models that were developed from the same dataset, Hanley and McNeil[132] developed a method to that takes into account the correlation between the AUC induced by the paired nature (same cases) of the data. They

introduced a correlation coefficient *"r"* that establishes the correlation between the curves based on the relationship of the correlation coefficient of abnormal cases "r_A" and the normal cases "r_N". These correlations can be calculated using the Pearson product-moment method for continuous values or the Kendall tau for categorical attributes. In this project, we were aiming on at least an AUC of ≥ 0.75.

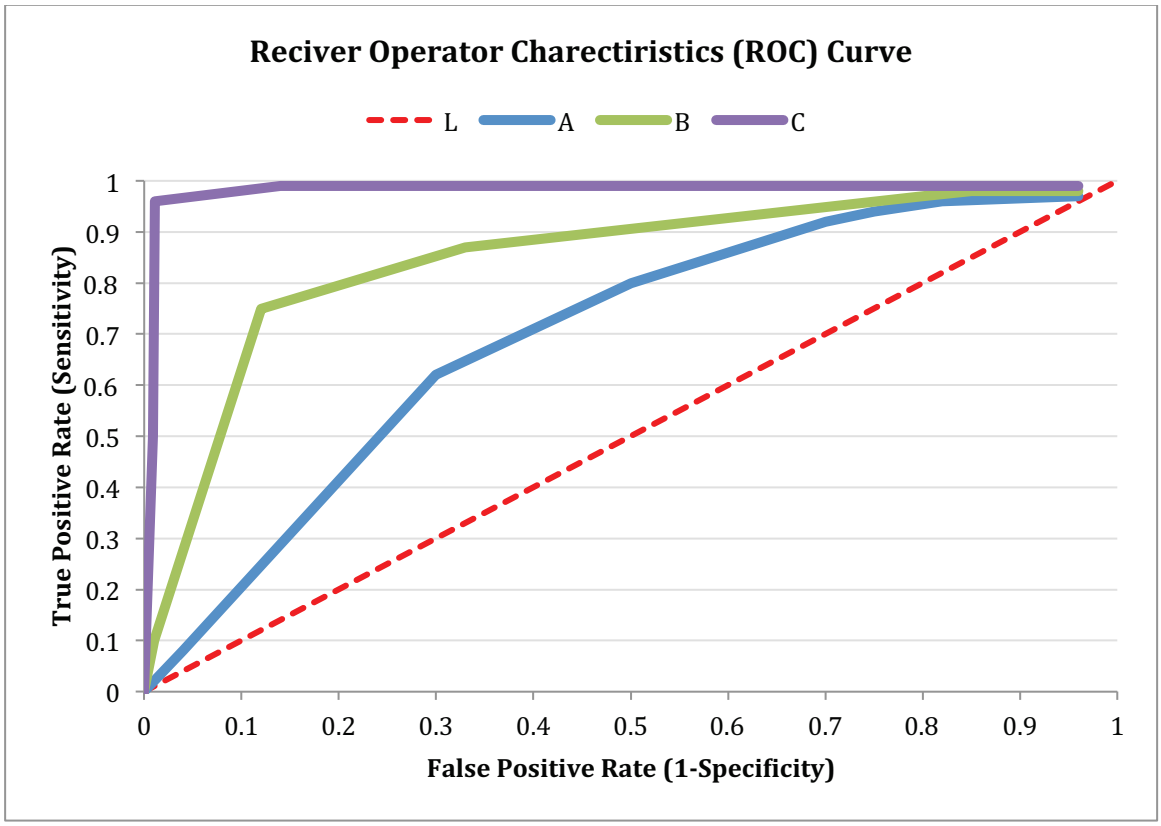

Figure C-14: ROC Curve
(L) Line of zero discrimination (AUC=0.5); (A) a test with an AUC=0.65; (B) a test with an AUC=0.80; (C) perfect test with an AUC=0.99.

McNemar's Test: was developed by Quinn McNemar in 1947[197]. It is applied to 2 × 2 confusion matrix with a binary attribute. Mainly used to analyze matched pairs of data, to determine whether the row and column marginal frequencies are equal. It is considered a variant of the χ^2 test and is considered to be non-parametric[133]. This test have been described and used by several authors in machine learning literature and some authors

advocate it's use as an adjunct, and some consider it superior, to the ROC-AUC for the development and improvement of algorithms[62, 133].

In machine learning, each algorithm will have two possible outcomes: Success = *s*, when the algorithm prediction agrees with the real data results (TP or TN) or Failure (*f*), when the algorithm prediction contradicts what was observed (FP or FN). After that, the outcome of each algorithm is compared to the other one and 4 possible results are generated in a 2 × 2 confusion matrix. Table C-16 illustrates the possible results of comparing two algorithms, using the same dataset, to apply the McNemar's Test[62].

Table C-16: McNemar Test Confusion Matrix for 2 Algorithms

	Algorithm A Success	Algorithm A Failure
Algorithm B Success	N_{ss}	N_{fs}
Algorithm B Failure	N_{sf}	N_{ff}

N_{ff} indicates the failure of both algorithms frequency, N_{ss} indicates the success of both algorithms frequency, N_{sf} denotes the frequency when algorithm A succeeded but Algorithm B failed, and N_{fs} denotes the frequency when algorithm A failed but Algorithm B succeeded. Although interesting, N_{ss} and N_{ff} do not illuminate the difference between the two algorithms performance. On the other hand, the N_{sf} and N_{fs} do show case the performance discrepancies.

The McNemar test uses the Z-score table of Normal distribution because it assumes that the Central Limit Theorem is valid (number of tested observations is ≥ 30).

$$Z = \frac{\left(\left|N_{sf} - N_{fs} - 1\right|\right)}{\sqrt{N_{sf} + N_{fs}}}$$

If the Z-score = 0, the two algorithms have a comparable performance. As this value diverges from 0 in positive direction, this indicates that their performance differs. A z-score of 1.96 indicates a 95% confidence level that there is a difference between the two

algorithms (two-tailed) and if we wanted to test that one is better, then it will have a 97.5% confidence level that one is better (one-tailed). It is capable of detecting a real difference between two scenarios when there is one, which means that it has a low Type-I error (detecting a difference when, in reality, there is no difference).

C.III Screening Test and Performance Measures in the Context

In real life, tests are performed in order to corroborate or refute a hypothesis or suspicion, which is typically based on some prior knowledge of a documented pattern in nature. In medicine, the main reason for performing a diagnostic test is to confirm or eliminate the presence of a specific problem, guiding the actions of professionals and often playing an instrumental role in preventing death and/or further deterioration [60, 63, 193].

Sensitivity, specificity, PPV, and negative predictive values (NPV) are commonly used measures in the evaluation of a diagnostic/screening test[60, 194]. The optimal test is a one that is highly sensitive and specific. Sensitivity and specificity are independent from the disease prevalence, but PPV and NPV are not. In reality, there is always a tradeoff between sensitivity and specificity; they are inversely proportional.

Test with high sensitivity (recall) but low specificity are good at detecting the disease (true detectives), but are sometimes too good in that they also often generate false positives. High sensitivity is of particular importance when the disease is relatively serious, treatable with good survival rates, and the "cure" is relatively inexpensive (e.g., breast cancer, prostate cancer, HIV in blood donors). Tests with a low sensitivity (Recall) but high specificity are good in excluding a disease. High specificity is important when the disease is fatal, relatively expensive (test or intervention), rare, and has a high complication rate (e.g., early cases of AIDS, terminal cases of cancer, limb ischemia requiring amputation).

Sensitivity and specificity usually remain fairly consistent across different populations [193]. It is important to note, however, that sensitivity and specificity do not specify information of an individual patient but rather about the group of people that underwent

the test. Patients, and their doctors, are typically not interested in the probability of having a positive test if they already know that they have the disease. Instead, they are more interested in the opposite: their likelihood of having the disease given a positive or negative test result. To answer this question, PPV (Precision) and NPV are used, because they reflect the applicability of a test in practice. Unfortunately, they also have one major limitation: they are incidence-dependent. Thus, having a highly sensitive test for a rare outcome will result in a low PPV and a very high NPV, because the number of the "false positive" examples will be extremely high and generate a drop in the PPV. This is a common clinical situation, since the number of people without the condition is usually much larger than the number of those with the condition; as a result, even a very good test can easily yield more false positives than true ones [195]. An acceptable complementary approach in clinical medicine is to subject patients who are initially positive to a test that has high sensitivity but low specificity, followed by a second test that has low sensitivity but high specificity. This approach makes certain that almost all of the false positive cases can be appropriately identified as disease free [60, 63, 193, 194].

C.IV Addressing Class Imbalance

Although balance is vitally important in scientific experimentation in order to eliminate random and confounding effects, truly achieving absolute balance is excessively rare. Fortunately, negative events (e.g: medical complications, financial loss, fraud, mechanical malfunction, death….etc) are not that common. In data science, the definition of an imbalanced dataset relates to the representation of the classes/outcomes compared to each other, with imbalance arising when there are many more samples from one class than from the rest of the classes. Specifically, data sets are unbalanced when one or more of the classes represent a small proportion of the set (called the minority class) while other classes make up the majority. In this situation, classification algorithms tends to predict the majority class very well but perform poorly on the minority class, an effect that has been attributed to three main reasons[57-59]: 1) the goal of minimizing the overall error (maximize accuracy), to which the minority class contributes very little; 2) the

algorithm's assumption that classes are balanced; and, 3) the assumption that impact of making an error is equal.

In reality, these assumptions usually do not hold. In the medical example of a patient who presents to the emergency room (ER) with chest pain, sending a patient home when they actually have a heart attack (i.e., missing a heart attack, "FN") has serious consequences (e.g., another heart attack, possible death) and is therefore more costly than keeping a patient in the ER for observation when he or she may not have had a heart attack (FP).

Class imbalance is a popular problem in the data science community[56-59, 137]. As detailed below, the two main strategies for dealing with the issue depend on the level of the intervention (data or algorithm) [56-59].

C.IV.i. Data Level Manipulation

A direct approach to deal with class imbalance is to manipulate the data, over-sampling the minority class, under-sampling the majority class, or doing both. Because data level manipulation methods are considered to be part of preprocessing, they can be applied to any algorithm. Although several authors have shown the ease and advantages of these methods, there are some drawbacks. Under sampling, because it discards some of the observations, creates the risk of losing some potentially important information. In the case of over sampling, artificially increasing the size of the creates the danger of overfitting the results to the training set, such that the algorithm will and it will fail to generalize.

In an attempt to fairly assess the effect of data manipulation on the model's performance, both approaches were independently applied in this work. The two data level manipulation methods that were available and used in WEKA [103] are:

1. ***SpreadSubSample*** produces a random sub-sample by under-sampling the majority class (by either specifying a ratio or the number of observations). Here, decreasing the number of the majority class instances reduces the difference between the minority and the majority class. Under-sampling is considered an effective method for dealing with class imbalance. In this approach; a subset of the majority class is used learn the model. Many of the majority class examples are ignored, the training set becomes more balanced which makes the training more

efficient[64, 65, 176]. The most common type of under-sampling is random majority under-sampling (RUS). In RUS, observations from the majority class are randomly removed. The main disadvantage of this approach is that we might overlook potentially important hidden information[57, 176].

2. ***Synthetic Minority Over-sampling Technique (SMOTE)*** [58], the algorithm oversamples the minority class by creating synthetic instances using a k-nearest-neighbor approach. The user can specify the over-sampling percentage, as well as the number of neighbors to use when creating synthetic instances. Here, artificially increasing the number of the minority class instances reduces the difference between the classes. SMOTE generates new artificial observations from similar observations, based on their feature subspace similarity, in the sample space. Described by Chawla in 2002[198] as a technique to overcome the issue of data imbalance and its effect on the predictive power of a model. Compared to random over-sampling with replacement, which predisposes to overfitting; this technique generates a more general and less specific minority class decision space. The minority class is over-sampled by taking each observation in the minority class and constructing a new simulated observation along the vector subspace that is common to all/any of *"i_{th}"* new observation nearest neighbors. The neighbors from the k-nearest neighbors are randomly chosen, based on the amount of required over-sampling. This will create new simulated observations that will have shared similarities that are common to its sample subspace in the minority class (k-nearest neighbors), rather than making exact copied of the same minority class.

C.IV.ii. Algorithm Parameter Manipulation

Algorithm parameter manipulation techniques involve manipulation of a parameter (e.g., cost of error, probability, threshold, etc.), such that the classifier is penalized for making mistakes and rewarded for making the right choice. The most commonly used algorithmic parameter manipulation is cost-sensitive learning, which takes into account the effect of making an error (cost) on the final performance assessment of the algorithm. It can be categorized into three main types[59]: 1) applying costs to the dataset as a form of data

space weighting, 2) applying cost-minimizing techniques to multiple algorithms (ensemble methods), and, 3) incorporating costs directly into the classifier.

If the cost of misclassification is known, then applying it to the cost matrix is straightforward. If the cost of misclassification is not known or hard to estimate, several authors advocate the use of different techniques to estimate cost. Zadrozny and Elkan argue that when conditional probabilities and cost are not known or the problem feature space is not well defined, cost can be expressed in a range instead of a single value[70]. The presence of class imbalance complicates the situation [57]. The simplest and most direct approach in a two-class problem is to equalize the cost by dividing the number of majority over the number of minority, then applying that ratio to the minority class[199]. The constructed cost-factors will balance potential total cost of the false positives to the potential total cost of the false negatives. Though this technique is attractive, equalizing cost assumes the homogeneity of cost across the population.

In WEKA, applying cost to a classifier can be done in the meta-classifier tab, with "CostSensitiveClassifier" making its base classifier cost-sensitive. Two methods can be used to introduce cost-sensitivity: reweighting training instances according to the total cost assigned to each class, or predicting the class with minimum expected misclassification cost (as opposed to the most likely class). Applying the reweighting technique requires that the analyst has a deep understanding of the domain, as well as a sense of the attached cost to the misclassification. Based on that, the reweighting technique was used in this work.

C.V Predictive Models

C.V.i. Logistic Regression

As the dependent attribute is non-linear in nature (Delirium= Yes/No), the use of linear regression is inappropriate. LR is the most commonly used method in the medical literature[8, 10, 11, 13, 15, 17-20, 22, 23, 28, 29, 35, 37, 51, 52, 75, 79]. LR examines the relationship between a binary outcome (dependent) attribute such as presence or absence of disease and predictor (independent) attribute(s), these independent attribute(s) can be continuous like age, categorical like gender, or even ordinal level of education. The presence or absence of a disease within a specified time period might be predicted from knowing the patient age, past medical history, family history, and any other attributes[53, 54].

Unlike linear regression models that are based on ordinary least squares algorithms, LR makes fewer assumptions. The outcome attribute does not need to be linearly related to any of the predictors but its natural log does, data does not need to fit a Gaussian distribution, no homoscedasticity (the standard deviations of the error terms are constant and is independent from other values), and it allows categorical or ordinal predictor attributes to be used in the model[54].

If $X_1, X_2, X_3, \ldots, X_n$ represents independent attributes (e.g.: Age, Gender, history of CVD, Level of education, etc…), Y signifies the outcome presence (Y=1 or Yes) or its absence (Y=0 or No), the following equation explains the linear relationship between the probability of the outcome existence (p) and the independent attributes:

$$Log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$$

where β_0 is the intercept (constant) and $\beta_1, \ldots, \beta_n$ are the regression coefficients for the independent attributes $X_1, \ldots, X_n$. Because the outcome is binary in nature, a patient can only have one out of the two outcomes. The probability of the outcome (p) can be calculated by mathematically transforming the original linear regression equation to yield the natural log of the odds of being in one outcome category versus the other category.

$$p(Y_{Yes}) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_n X_n}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_n X_n}} \ , where\ 0 \leq p \leq 1$$

Each regression coefficient signifies the magnitude of influence the corresponding independent attribute on the outcome. The effect of the independent attributes on the outcome is usually expressed in odds ratio, which represents the effect by which the odds of an outcome change for a one-unit change in the independent attribute[53, 54, 144].

One of the key steps in building a LR model is independent attribute selection, which is usually based on their statistical significance or clinical relevance. Regardless of applied the selection criteria, the analyst must be aware of the potential role of cofounders. Confounding attributes are those whose relationship to both the outcome and another independent attribute obscures the true association between that independent attribute and the outcome[53].

Although LR models are considered to be less stringent and have less assumptions compared to linear regression, it still has some basic assumptions that must be met[53, 54]:

1. Independence of errors: observations are not related or repeated (no duplicate measurement of the same patient).
2. Continuous independent attributes are linear to the natural logarithmic scale.
3. Absence of co-linearity or redundancy between independent attributes.
4. Lack of outlier effect
5. Interactions between independent attributes: the effect on model performance and complexity should be always acknowledged.

One of the specific goals in LR, and generally in predictive modeling, is to generate the most concise model that best explains the outcome without losing any important information and avoiding overfitting. The usual recommendation is to have ≤10 independent attributes; each one should have at least 10 cases in each outcome and preferably 50 cases in each outcome.

Another important consideration that the analyst must keep in mind is the model building method, candidate attributes inclusion. These methods are all based on the statistical

contribution of the attribute in the improvement of the model fit. Three different methods are available: Direct, Sequential or Stepwise. In the direct approach, all independent attributes are assumed to be equally important and are included in the model. The sequential approach adds or deletes attributes in a systematic manner and observes their effect. Stepwise selection chooses attributes based on predefined statistical criteria, which is influenced by the data. Forward, Backward and Subset selection are the most commonly used types of the stepwise methods. In Forward selection, the model starts with no predictors in the model. It then starts adding independent attributes with the smallest p-value, one at a time. It continues doing this until no more improvement in the model fit. In Backward selection, the model starts with all of the predictors in the model. It then starts removing independent attributes with the largest p-value, one at a time. It continues doing this until the model reaches best fit. In the subset method, different models are generated and compared to determine the best fit[54].

C.V.ii. Artificial Neural Networks

Artificial Neural Networks (ANN) is a computer model inspired by the biological function of the brain. It consist of highly interconnected neurons (nodes), and their overall ability to predict outcomes is determined by the connections between these neurons[64, 66]. They apply nonlinear mathematical models to simulates the brain's own problem-solving process. Just as humans apply knowledge gained from past experience to new problems or situations, a neural network base its new decisions on previously solved examples by building a system of "neurons"[150]. They are considered to be complex non-linear systems that can deal with noisy or incomplete data, allow multiple and simultaneous multilevel interactions, and have a high ability to generalize based on the input[57, 65, 144]. They are mainly used in pattern recognition and forecasting, where a precise computational answer is not necessary and the goal is mainly of approximation. They are also used when the number of attributes is large and the relationship between them is not really well understood.

ANN models are just a simple representation of a real network of neurons. Like a network of neurons in the brain, they learn by adjusting the connection weights between present neurons. Learning from reality (AKA: Adaptive learning), self-organization, parallel

processing, it is dynamic, deals efficiently with non-linear relationships, and high tolerance to redundant information are some of the major strengths of ANN[152]. Some of the drawbacks of ANN are that the model is highly dependent on the quality of the input data; the model is unpredictable as the network try's to find the optimal solution by independently (AKA: Black Box Modeling); and they tend to be slower to train.

Usually, a neural network is organized in layers. Each layer is composed of several nodes (neurons). The layers are usually divided into 3 main categories: input, hidden, and output. The layers are connected to each other in a specific sequence; with the input layer passing a signal to the hidden layer that manipulates that signal with a specified activation function. After that, the new signal either activates or inhibits the output layer. In a classification task; the input layer represents the attributes, and the output layer represents the outcome. The hidden layer represents the actual processing of information and where most of the work happens. There is usually a single level in the input and output layers. In contrast, the hidden layer can be made of ≥1 level. Increasing the levels in the hidden layer will increase the network complexity, but that does not necessarily improve the performance of the network and will lead to over-fitting[64-66].

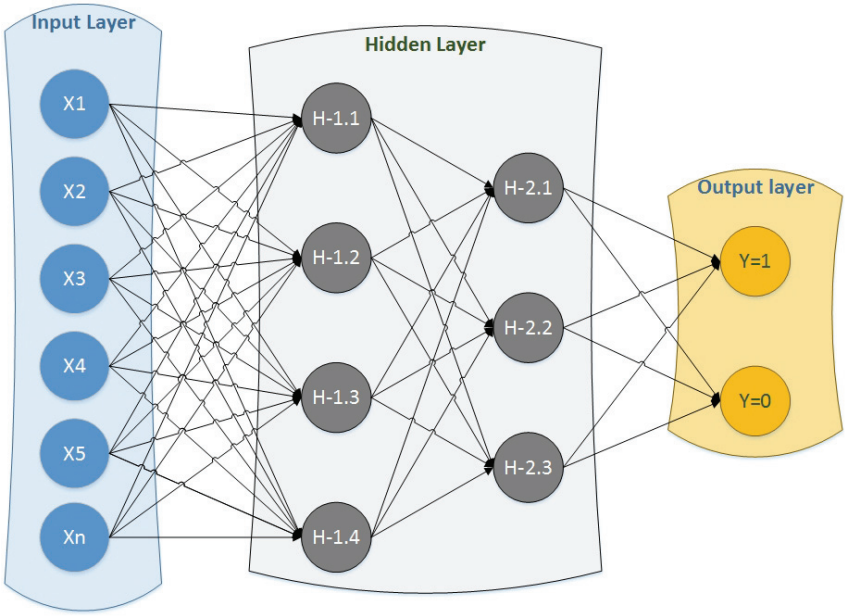

Figure C-15: ANN Architecture

Most ANN will have some form of "learning rule" that modifies the weights of the connections based on the input. There are several types of artificial neural networks, but the most commonly used in a supervised learning task, classification, is the BPNN. In this type learning occurs with each cycle or "epoch" (i.e. each time the network is presented with a new input pattern) through a forward activation flow of outputs, and then backwards error propagation of weight adjustments. In simple terms, the neural network firstly presented with a pattern on which it makes an arbitrary guess as to what it may be. It then compares the its results to the actual data and measures how far is the answer was from the right one, and based on this difference it makes an appropriate adjustment to the weights of its connection. In the case of a binary classification task; the sigmoid, or logistic, function is usually used as an activation function. The use of this function allows the BPNN to model classification problems that are linearly inseparable[64].

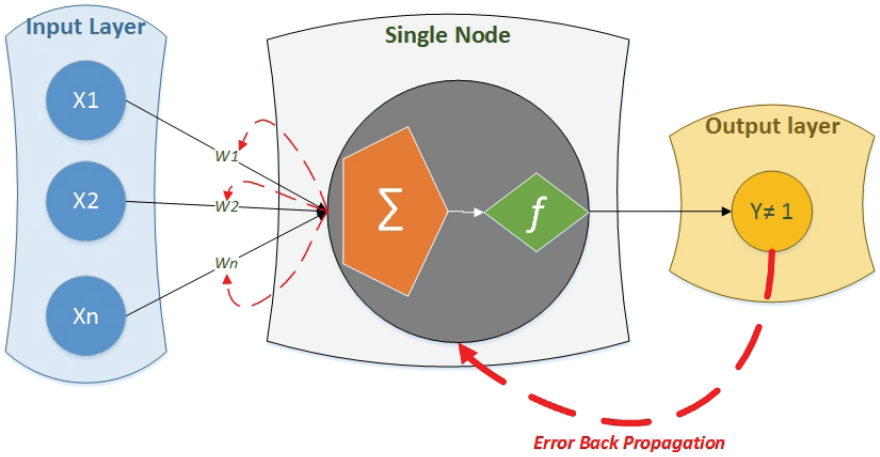

Figure C-16: Back Propagation Algorithm

Some of the major criticisms of BPNN are: the difficulty of directly interpreting the symbolic meaning of the denoted connection weights, what happens inside the "hidden layer" in the network, and the network structure is based on the analyst understanding of the problem in hand. The issue with the network structure is that there is no clear rules that governs its construction, it is mainly an iterative process based on a trial and error strengthened by good domain knowledge. In binary classification, the assignment of an observation to one possible outcome is based the output threshold of ≥0.5 probability of the observation belonging to that class[64, 65].

C.V.iii. Bayesian Belief Networks

Simply put, a Probability how likely is something is going to happen. This means that we will never be absolutely certain if an event is going to occur or not. The best we can do is attributing the possibility of it occurring. It is usually represented on a scale between 0 and 1.This can be represented mathematically with the following equation:

$$Probability\ of\ an\ Outcome = \frac{Number\ of\ times\ that\ outcome\ was\ observed}{Total\ number\ of\ all\ possible\ outcomes}$$

In decision-making, Probabilities usually are thought of as a guides rather than absolute numbers, because they clearly establish the intrinsic uncertainty. In statistics, Bayes Rule is considered to be in the core of probability theory[153]. In Bayesian terms, probability measures a degree of belief. It links the degree of confidence of an event occurring with and without accounting for a condition. This can be represented mathematically with the following equation:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Where Y is the event, X is the conditioning attribute and "|" is the symbol for given.
In the case of multiple attributes, the product of multiplying multiple probabilities can be used to infer the probability of the event provided that the attributes satisfy one major assumption, class-conditional independence. This assumes that all of the attributes values are independent of one another (There are no relationships between any of the attributes). Unfortunately, in reality this rarely happens. In a biological system, there are so many interconnected and simultaneous interactions that are heavily influenced by earlier events.

To mitigate the effect of class-conditional independence, BBNs were introduced[154, 155]. BBN is a result of the successful partnership between graph theory and Bayesian thermo. BBN decomposes the problem space to multiple smaller subspaces, identifies a joint conditional probability distribution that is relevant to each subspace, and constructs links (Arcs) between these subspaces based on a probabilistic dependency pattern. A BBN has two main components: a DAG and a CPT for each attribute[64, 65]. Having graph they is directed implies the casual relationship between a parent attribute and its children. Preventing the possibility of an attribute to have a cycle or a trail back to itself

augments the understanding of the causal relationship and eliminates ambiguity[67, 125, 155]. There are two basic methods to construct the graph (Network Topology): If the structure is known and relationships are well understood, it can be built by the expert and the data passed through it to identify probabilities; but if the relationships and dependencies are not clear, the graph structure can be deduced from the data[64, 67, 125]. Each attribute will have a CPT that is only based on the dependencies that it has with its antecedents/parents[67, 125, 155]. This way, the attribute will be conditionally independent from all of the other attributes that it has no direct connections with (children or parents)[64, 67, 125].

The main advantages of this approach are that: it acknowledges the existence of some dependencies between connected attributes but yet class conditional independence between subgroups of unconnected attributes; provide a simple, yet elegant, graphical representation of the internal relationships, handles noisy data very well, the model is transparent as all of the model parameters have a clear semantic interpretation "White Box Modeling", multipurpose as it can provide causal or evidential relationships, easily adjustable as it can be improved by expert knowledge, and is augmented by a probability distribution that can be easily interpreted by humans and machines which are central for decision-making[67, 125, 156]. Some of the limitations of this approach are that: it is heavily dependent on the input data quality, it requires discretization of continuous values (most algorithms have a built in function to discretize continuous attributes based on their distribution shape), and it requires that there is no missing data (most algorithms have a built in function to replace missing data).

C.VI SAS and WEKA Experiments Setting and Options

C.VI.i. Logistic regression

```
proc logistic data = dmwin.train plots=roc alpha=0.05 descending;
 class AgeBin(param=ref ref='<60')Frail(param=ref ref='No') Gender (param=ref ref='Female') COPD(param=ref ref='No') NYHA(param=ref ref='Class-1')
 PreopAF (param=ref ref='No') ClustArr(param=ref ref='cluster0') ClustDMDMCont(param=ref ref='cluster0')
 EF30(param=ref ref='>50') IABPwhen2(param=ref ref='None') CVD (param=ref ref='No') ORSTATUS (param=ref ref='Emergenc')
 Turndown(param=ref ref='No') ProcDif(param=ref ref='Single') AS2 (param=ref ref='None') MR2(param=ref ref='None')
 ORIno(param=ref ref='No') TEEIntraOp(param=ref ref='No') AVDIS(param=ref ref='No') BldProWithin48(param=ref ref='No')
 longvent(param=ref ref='No') CVICUHrsBin(param=ref ref='<24hrs');
 model delirium (event='1')= age|frail gender COPD PreopAF ClustArr ClustDMDMCont EF30 EUROII CVD PreopHb Turndown ProcDif
 PreOpEstCrJel AS2 MR2 IABPwhen2 ORIno TEEIntraOp BldProWithin48 CVICUHrsBin|Longvent/ rsq lackfit Selection=Stepwise expb outroc=rocdata;
 score data=dmwin.test5 out=valpred outroc=vroc;
 roc; roccontrast;
 run;
```

Figure C-17: Logistic Regression SAS Code

C.VI.iii. Artificial Neural Network

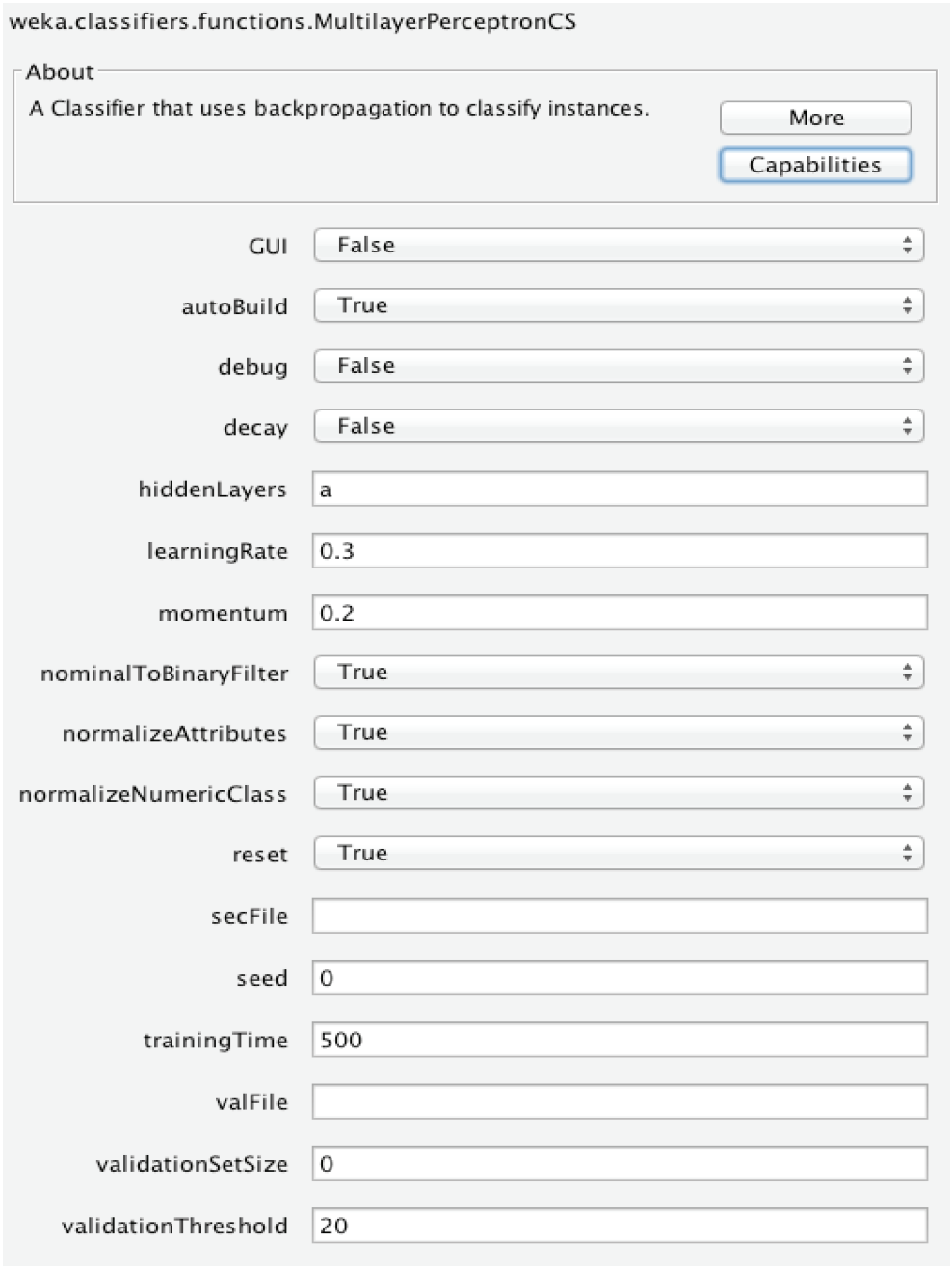

Figure C-18: ANN-Standard Setting in WEKA 3.7

Options:

Debug: If set to true, classifier may output additional info to the console.

Decay: This will cause the learning rate to decrease. This may help to stop the network from diverging from the target output, as well as improve general performance.

Learning Rate: The amount the weights are updated.

Training Time: The number of epochs to train through.

Seed: Seed used to initialize the random number generator.

Validation Threshold: Used to terminate validation testing. The value here dictates how many times in a row the validation set error can get worse before training is terminated.

Auto Build: Adds and connects up hidden layers in the network.

GUI: Brings up a gui interface.

Hidden Layers: This defines the hidden layers of the neural network.

Normalize Numeric Class: This will normalize the class if it's numeric.

Val File: Set the name of a validation file in data source format.

Nominal To Binary Filter: This will preprocess the instances with the filter.

Validation Set Size: The percentage size of the validation set.

Normalize Attributes: This will normalize the attributes.

Sec File: Set the name of a secondary training file in data source format.

Momentum: Momentum applied to the weights during updating.

Reset: This will allow the network to reset with a lower learning rate.

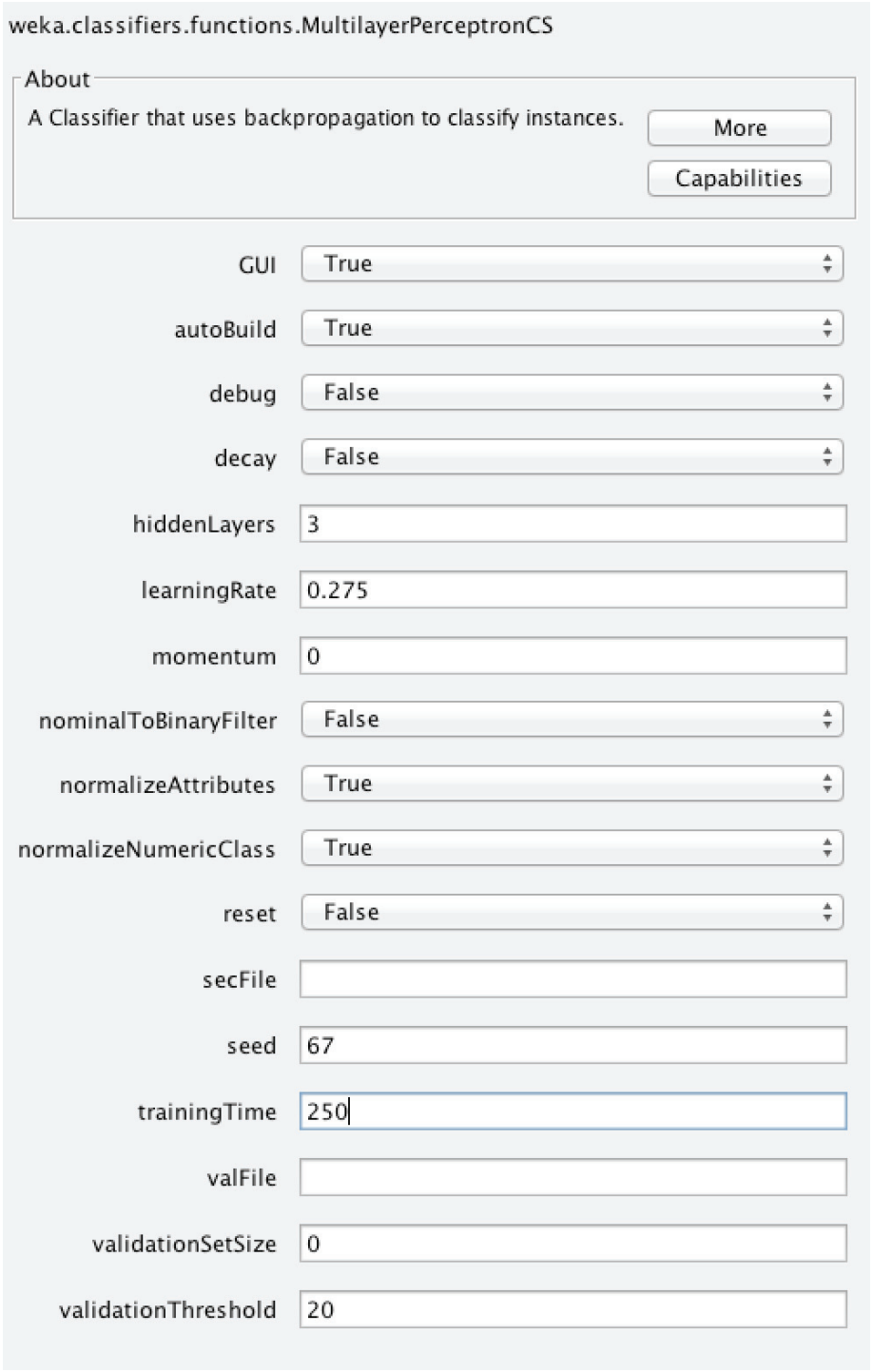

Figure C-19: ANN With 1 Hidden Layer

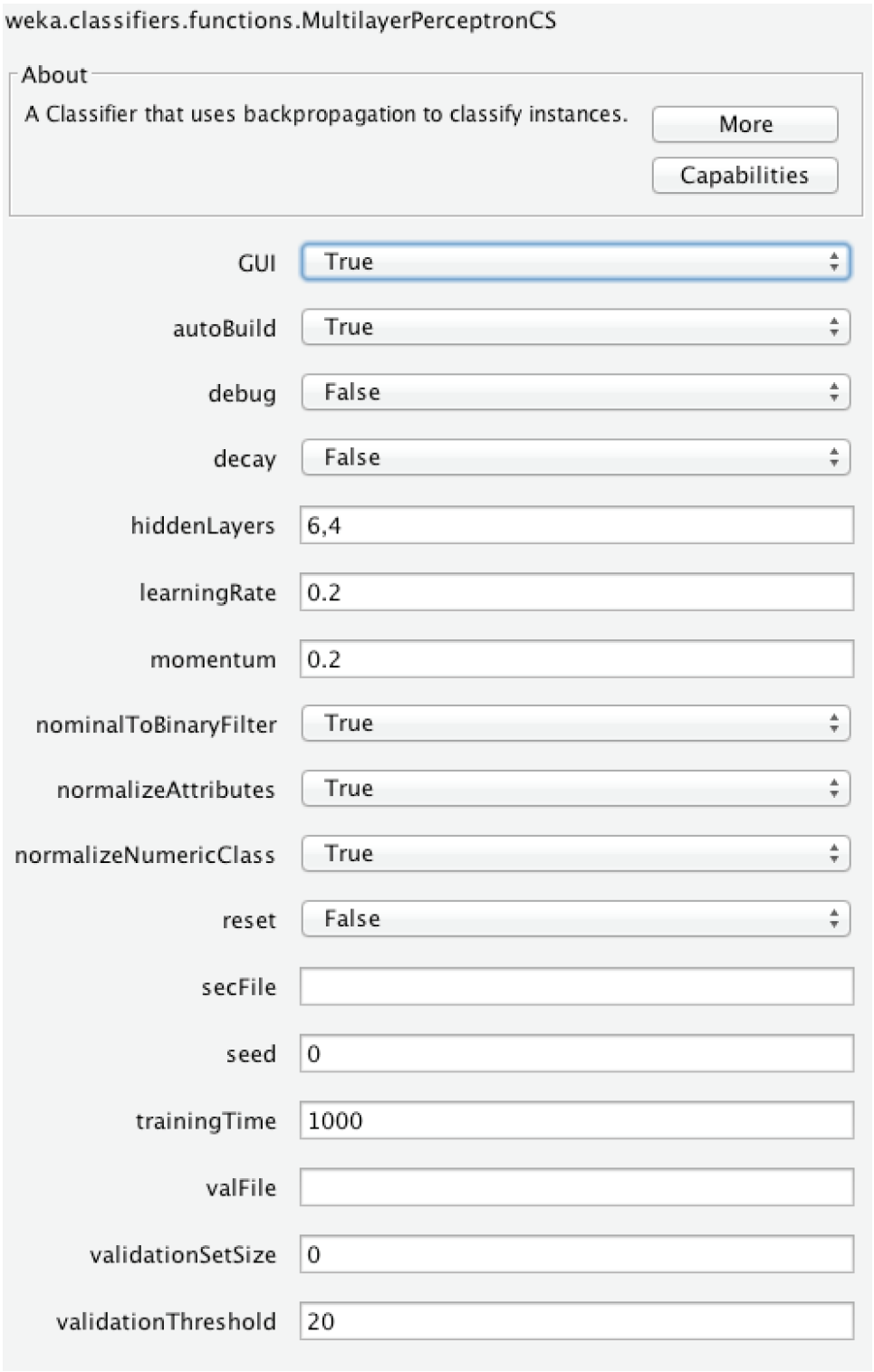

Figure C-20: ANN With 2 Hidden Layers Setting

C.VI.iv. Bayesian Belief Network

WEKA 3.7 has a separate section for Bayesian Models under its Classify tab. The main two options are a Bayes Net model or different types of Naïve Bayesian models. WEKA allows the user to manipulate several parameters when constructing the network (Figure C-21). If the network structure is known, then it can be built and loaded as an "xml" external file [171]; alternatively, it can be determined based on the data using different search algorithms that are provided in WEKA.

The *searchAlgorithm* option provides the option to learn the network structure from the data. Four main categories exist under this tab: Conditional independence tests, fixed, global score, and local score. In the case of local score metrics, learning a network structure can be considered an optimization problem where a quality measure of a network structure given the training data needs to be maximized. In the case of global score metrics, an internal cross validation is applied. The network structure can be learned and then validated using left out observations, thus providing an out-of-sample evaluation method by repeatedly splitting the data in training and validation sets [65, 171].

The default WEKA setting is local scoring metrics, primarily because they are computationally less expensive. The local scoring metrics have several algorithms: K2, Hill Climbing, LAGD Hill Climbing, Simulated Annealing, and others. The *K2 algorithm* (Figure C-22) is a greedy hill climbing approach that adds connections in a fixed order for all of the attributes. It processes each attribute in a cycle and examines the effect of adding connections from other attributes to the current one. In each cycle it chooses the structure that maximizes the network score. When there is no further improvement, it moves to the next attribute [65, 171].

The *estimator* option is the way the CPT is estimated after the structure of the network is learned. The *SimpleEstimator* algorithm calculates the CPT of each node directly based on its connections without accounting for the other none connected nodes. Alpha is a parameter related to setting the initial value of each observation and is used to estimate the probability tables. WEKA uses a correction to prevent zero probabilities[171]. The default setting in WEKA is the *SimpleEstimator* algorithm, with an Alpha=0.5.

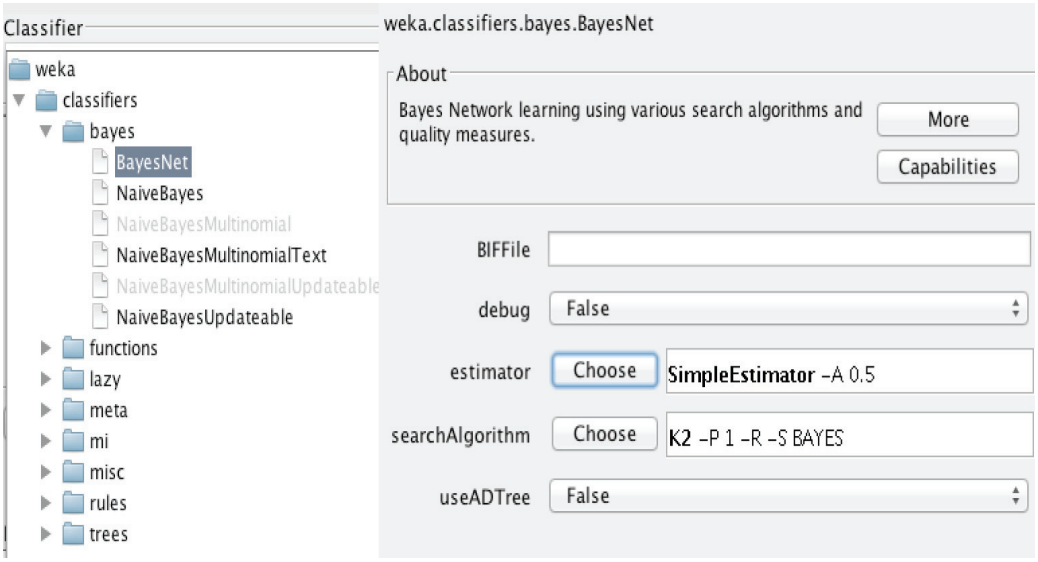

Figure C-21: Default Bayesian Belief Network Setting

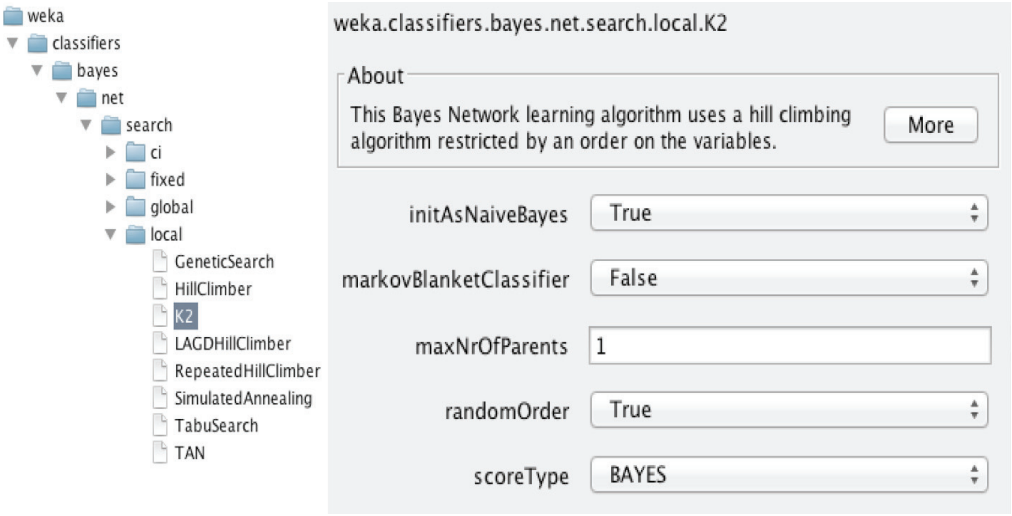

Figure C-22: The K2 Search Algorithm Setting

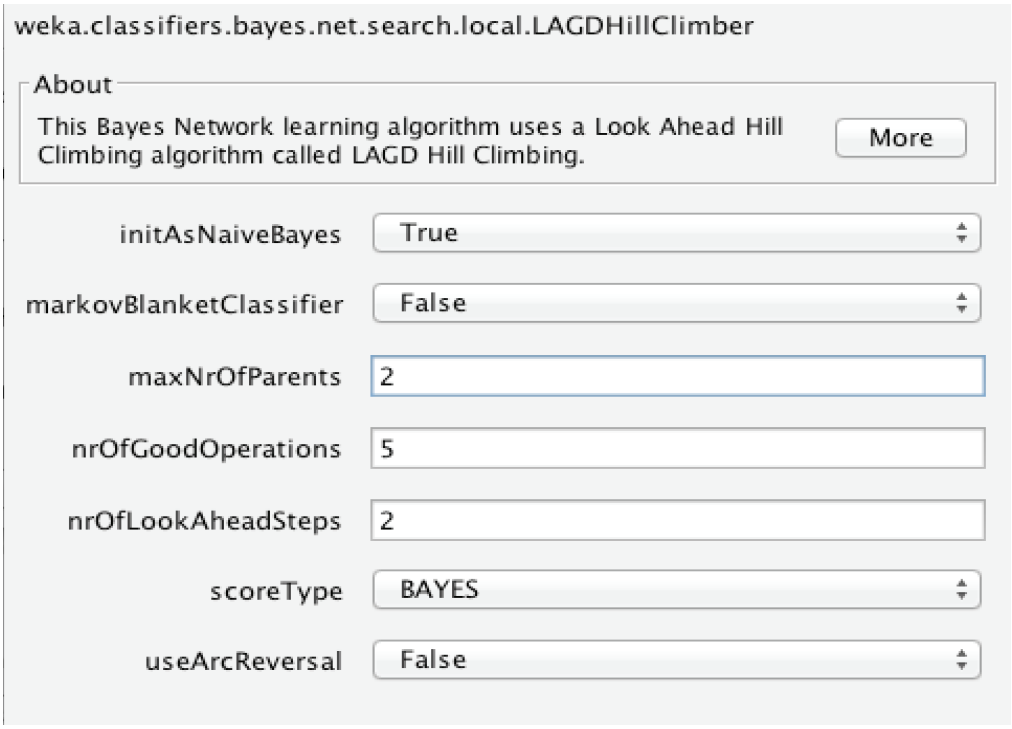

Figure C-23: LAGD Hill Climbing Algorithm Setting

REFERENCES

1 Brothwell, D.R.: 'Digging up bones: the excavation, treatment, and study of human skeletal remains' (Cornell University Press, 1981. 1981)

2 Hill, J.W.a.K., Doris K.: 'Chemistry for changing times' (Pearson: Prentice Hall. Upper saddle river, New Jersey., 10th edition edn)

3 Royston, D., and Cox, F.: 'Anaesthesia: the patient's point of view', The Lancet, 2003, 362, (9396), pp. 1648-1658

4 Deschka, H., Schreier, R., El-Ayoubi, L., Erler, S., Alken, A., and Wimmer-Greinecker, G.: 'Survival, functional capacity, and quality of life after cardiac surgery followed by long-term intensive care stay', Thorac Cardiovasc Surg, 2013, 61, (8), pp. 696-700

5 Abdallah, M.S., Wang, K., Magnuson, E.A., Spertus, J.A., Farkouh, M.E., Fuster, V., and Cohen, D.J.: 'Quality of life after PCI vs CABG among patients with diabetes and multivessel coronary artery disease: a randomized clinical trial', JAMA : the journal of the American Medical Association, 2013, 310, (15), pp. 1581-1590

6 Romero, P.S., de Souza, E.N., Rodrigues, J., and Moraes, M.A.: 'Changes in quality of life associated with surgical risk in elderly patients undergoing cardiac surgery', International journal of nursing practice, 2014

7 Bartels, K., McDonagh, D.L., Newman, M.F., and Mathew, J.P.: 'Neurocognitive outcomes after cardiac surgery', Current opinion in anaesthesiology, 2013, 26, (1), pp. 91-97

8 Koster, S., Oosterveld, F.G., Hensens, A.G., Wijma, A., and van der Palen, J.: 'Delirium after cardiac surgery and predictive validity of a risk checklist', The Annals of thoracic surgery, 2008, 86, (6), pp. 1883-1887

9 Martin, B.-J., Buth, K.J., Arora, R.C., and Baskett, R.: 'Delirium as a predictor of sepsis in post-coronary artery bypass grafting patients: a retrospective cohort study', Crit Care, 2010, 14, (5), pp. R171

10 Gottesman, R.F., Grega, M.A., Bailey, M.M., Pham, L.D., Zeger, S.L., Baumgartner, W.A., Selnes, O.A., and McKhann, G.M.: 'Delirium after coronary artery bypass graft surgery and late mortality', Annals of neurology, 2010, 67, (3), pp. 338-344

11 Bucerius, J., Gummert, J.F., Borger, M.A., Walther, T., Doll, N., Falk, V., Schmitt, D.V., and Mohr, F.W.: 'Predictors of delirium after cardiac surgery delirium: effect of beating-heart (off-pump) surgery', The Journal of thoracic and cardiovascular surgery, 2004, 127, (1), pp. 57-64

12 Cavallazzi, R., Saad, M., and Marik, P.E.: 'Delirium in the ICU: an overview', Annals of intensive care, 2012, 2, (1), pp. 49

13 Edelstein, D.M., Aharonoff, G.B., Karp, A., Capla, E.L., Zuckerman, J.D., and Koval, K.J.: 'Effect of postoperative delirium on outcome after hip fracture', Clinical orthopaedics and related research, 2004, 422, pp. 195-200

14 Martin, B.-J., Buth, K.J., Arora, R.C., and Baskett, R.J.: 'Delirium: a cause for concern beyond the immediate postoperative period', The Annals of thoracic surgery, 2012, 93, (4), pp. 1114-1120

15 Bakker, R.C., Osse, R.J., Tulen, J.H., Kappetein, A.P., and Bogers, A.J.: 'Preoperative and operative predictors of delirium after cardiac surgery in elderly patients', European Journal of Cardio-Thoracic Surgery, 2012, 41, (3), pp. 544-549

16 Boeken, U., Litmathe, J., Feindt, P., and Gams, E.: 'Neurological complications after cardiac surgery: risk factors and correlation to the surgical procedure', The Thoracic and cardiovascular surgeon, 2005, 53, (01), pp. 33-36

17 Afonso, A., Scurlock, C., Reich, D., Raikhelkar, J., Hossain, S., Bodian, C., Krol, M., and Flynn, B.: 'Predictive model for postoperative delirium in cardiac surgical patients', in Editor (Ed.)^(Eds.): 'Book Predictive model for postoperative delirium in cardiac surgical patients' (SAGE Publications, 2010, edn.), pp. 212-217

18 Tan, M.C., Felde, A., Kuskowski, M., Ward, H., Kelly, R.F., Adabag, A.S., and Dysken, M.: 'Incidence and predictors of post-cardiotomy delirium', American Journal of Geriatric Psych, 2008, 16, (7), pp. 575-583

19 Smulter, N., Lingehall, H.C., Gustafson, Y., Olofsson, B., and Engstrom, K.G.: 'Delirium after cardiac surgery: incidence and risk factors', Interactive cardiovascular and thoracic surgery, 2013, 17, (5), pp. 790-796

20 Andrejaitiene, J., and Sirvinskas, E.: 'Early post-cardiac surgery delirium risk factors', Perfusion, 2012, 27, (2), pp. 105-112

21 American Psychiatric Association, A.P.A.D.S.M.T.F.: 'Diagnostic and statistical manual of mental disorders : DSM-5' (2013. 2013)

22 Hudetz, J.A., Iqbal, Z., Gandhi, S.D., Patterson, K.M., Byrne, A.J., and Pagel, P.S.: 'Postoperative delirium and short-term cognitive dysfunction occur more frequently in patients undergoing valve surgery with or without coronary artery bypass graft surgery compared with coronary artery bypass graft surgery alone: results of a pilot study', Journal of cardiothoracic and vascular anesthesia, 2011, 25, (5), pp. 811-816

23 Norkiene, I., Ringaitiene, D., Kuzminskaite, V., and Sipylaite, J.: 'Incidence and risk factors of early delirium after cardiac surgery', BioMed research international, 2013, 2013, pp. 323491

24 Young, J., Murthy, L., Westby, M., Akunne, A., and O'Mahony, R.: 'Diagnosis, prevention, and management of delirium: summary of NICE guidance', BMJ, 2010, 341

25 APA, W.G.O.D.: 'PRACTICE GUIDELINE for the Treatment of Patients With Delirium', in Editor (Ed.)^(Eds.): 'Book PRACTICE GUIDELINE for the Treatment of Patients With Delirium' (American Psyciatric Assosiation, 2010, edn.), pp.

26 Lee, D.H., Buth, K.J., Martin, B.J., Yip, A.M., and Hirsch, G.M.: 'Frail patients are at increased risk for mortality and prolonged institutional care after cardiac surgery', Circulation, 2010, 121, (8), pp. 973-978

27 Spiller, J.A., and Keen, J.C.: 'Hypoactive delirium: assessing the extent of the problem for inpatient specialist palliative care', Palliative Medicine, 2006, 20, (1), pp. 17-23

28 Stransky, M., Schmidt, C., Ganslmeier, P., Grossmann, E., Haneya, A., Moritz, S., Raffer, M., Schmid, C., Graf, B.M., and Trabold, B.: 'Hypoactive delirium after cardiac surgery as an independent risk factor for prolonged mechanical ventilation', Journal of cardiothoracic and vascular anesthesia, 2011, 25, (6), pp. 968-974

29 Tse, L., Schwarz, S.K., Bowering, J.B., Moore, R.L., Burns, K.D., Richford, C.M., Osborn, J.A., and Barr, A.M.: 'Pharmacological Risk Factors for Delirium after Cardiac Surgery: A Review', Current neuropharmacology, 2012, 10, (3), pp. 181

30 Hamerman, D.: 'Toward an understanding of frailty', Annals of internal medicine, 1999, 130, (11), pp. 945-950

31 Hardy, S.E., Dubin, J.A., Holford, T.R., and Gill, T.M.: 'Transitions between states of disability and independence among older persons', American journal of epidemiology, 2005, 161, (6), pp. 575-584

32 Lally, F., and Crome, P.: 'Understanding frailty', Postgraduate medical journal, 2007, 83, (975), pp. 16-20

33 Buth, K.J., Gainer, R.A., Legare, J.F., and Hirsch, G.M.: 'The changing face of cardiac surgery: practice patterns and outcomes 2001-2010', The Canadian journal of cardiology, 2014, 30, (2), pp. 224-230

34 Pierri, M.D., Capestro, F., Zingaro, C., and Torracca, L.: 'The changing face of cardiac surgery patients: an insight into a Mediterranean region', European journal of cardio-thoracic surgery : official journal of the European Association for Cardio-thoracic Surgery, 2010, 38, (4), pp. 407-413

35 Inouye, S.K., Bogardus Jr, S.T., Charpentier, P.A., Leo-Summers, L., Acampora, D., Holford, T.R., and Cooney Jr, L.M.: 'A multicomponent intervention to prevent delirium in hospitalized older patients', New England journal of medicine, 1999, 340, (9), pp. 669-676

36 Kalisvaart, K.J., De Jonghe, J.F., Bogaards, M.J., Vreeswijk, R., Egberts, T.C., Burger, B.J., Eikelenboom, P., and Van Gool, W.A.: 'Haloperidol Prophylaxis for Elderly Hip‐Surgery Patients at Risk for Delirium: A Randomized Placebo‐Controlled Study', Journal of the American Geriatrics Society, 2005, 53, (10), pp. 1658-1666

37 Katznelson, R., Djaiani, G.N., Borger, M.A., Friedman, Z., Abbey, S.E., Fedorko, L., Karski, J., Mitsakakis, N., Carroll, J., and Beattie, W.S.: 'Preoperative use of statins is associated with reduced early delirium rates after cardiac surgery', Anesthesiology, 2009, 110, (1), pp. 67-73

38 Goodwin, S.: 'Data rich, information poor (DRIP) syndrome: is there a treatment?', Radiology management, 1995, 18, (3), pp. 45-49

39 Berwick, D.M., and Hackbarth, A.D.: 'Eliminating waste in US health care', JAMA : the journal of the American Medical Association, 2012, 307, (14), pp. 1513-1516

40 Bradley, P.: 'Predictive analytics can support the ACO model', Healthcare financial management : journal of the Healthcare Financial Management Association, 2012, 66, (4), pp. 102-106

41 Peiris, A.N., and Patel, C.B.: 'Predictive analytics: the fifth clinical element', Southern medical journal, 2013, 106, (4), pp. 290-291

42 Bell, L.M., Grundmeier, R., Localio, R., Zorc, J., Fiks, A.G., Zhang, X., Stephens, T.B., Swietlik, M., and Guevara, J.P.: 'Electronic health record–based decision support to improve asthma care: a cluster-randomized trial', Pediatrics, 2010, 125, (4), pp. e770-e777

43 Chaudhry, B., Wang, J., Wu, S., Maglione, M., Mojica, W., Roth, E., Morton, S.C., and Shekelle, P.G.: 'Systematic review: impact of health information technology on quality, efficiency, and costs of medical care', Annals of internal medicine, 2006, 144, (10), pp. 742-752

44 O'Connor, P.J., Sperl-Hillen, J.M., Rush, W.A., Johnson, P.E., Amundson, G.H., Asche, S.E., Ekstrom, H.L., and Gilmer, T.P.: 'Impact of electronic health record clinical decision support on diabetes care: a randomized trial', The Annals of Family Medicine, 2011, 9, (1), pp. 12-21

45 Stone, A.A., Shiffman, S., Schwartz, J.E., Broderick, J.E., and Hufford, M.R.: 'Patient compliance with paper and electronic diaries', Controlled clinical trials, 2003, 24, (2), pp. 182-199

46 Ueckert, F., Goerz, M., Ataian, M., Tessmann, S., and Prokosch, H.-U.: 'Empowerment of patients and communication with health care professionals through an electronic health record', International journal of medical informatics, 2003, 70, (2), pp. 99-108

47 Knops, A.M., Legemate, D.A., Goossens, A., Bossuyt, P.M., and Ubbink, D.T.: 'Decision aids for patients facing a surgical treatment decision: a systematic review and meta-analysis', Annals of surgery, 2013, 257, (5), pp. 860-866

48 Kawamoto, K., Houlihan, C.A., Balas, E.A., and Lobach, D.F.: 'Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success', Bmj, 2005, 330, (7494), pp. 765

49 McDonald, K.M., Matesic, B., Contopoulos-Ioannidis, D.G., Lonhart, J., Schmidt, E., Pineda, N., and Ioannidis, J.P.: 'Patient Safety Strategies Targeted at Diagnostic ErrorsA Systematic Review', Annals of internal medicine, 2013, 158, (5_Part_2), pp. 381-389

50 McCoy, A., Wright, A., Eysenbach, G., Malin, B., Patterson, E., Xu, H., and Sittig, D.: 'State of the art in clinical informatics: evidence and examples', Yearbook of medical informatics, 2013, 8, pp. 13-19

51 Behrends, M., DePalma, G., Sands, L., and Leung, J.: 'Association between intraoperative blood transfusions and early postoperative delirium in older adults', Journal of the American Geriatrics Society, 2013, 61, (3), pp. 365-370

52 van den Boogaard, M., Schoonhoven, L., van der Hoeven, J.G., van Achterberg, T., and Pickkers, P.: 'Incidence and short-term consequences of delirium in critically ill patients: A prospective observational cohort study', International journal of nursing studies, 2012, 49, (7), pp. 775-783

53 Stoltzfus, J.C.: 'Logistic regression: a brief primer', Academic Emergency Medicine, 2011, 18, (10), pp. 1099-1104

54 Menard, S.: 'Applied logistic regression analysis' (Sage, 2002. 2002)

55 Breiman, L.: 'Statistical modeling: The two cultures (with comments and a rejoinder by the author)', Statistical Science, 2001, 16, (3), pp. 199-231

56 Menardi, G., and Torelli, N.: 'Training and assessing classification rules with imbalanced data', Data Mining and Knowledge Discovery, 2014, 28, (1), pp. 92-122

57 Ganganwar, V.: 'An overview of classification algorithms for imbalanced datasets', Int. J. Emerg. Technol. Adv. Eng, 2012, 2, (4), pp. 42-47

58 Chawla, N.V.: 'Data mining for imbalanced datasets: An overview': 'Data mining and knowledge discovery handbook' (Springer, 2005), pp. 853-867

59 He, H., and Garcia, E.A.: 'Learning from imbalanced data', Knowledge and Data Engineering, IEEE Transactions on, 2009, 21, (9), pp. 1263-1284

60 Parikh, R., Mathai, A., Parikh, S., Sekhar, G.C., and Thomas, R.: 'Understanding and using sensitivity, specificity and predictive values', Indian journal of ophthalmology, 2008, 56, (1), pp. 45

61 Sasaki, Y.: 'The truth of the F-measure', Teaching, Tutorial materials, Version: 26th October, 2007

62 Clark, A.F., and Clark, C.: 'Performance characterization in computer vision a tutorial', in Editor (Ed.)^(Eds.): 'Book Performance characterization in computer vision a tutorial' (Citeseer, 1999, edn.), pp.

63 Lalkhen, A.G., and McCluskey, A.: 'Clinical tests: sensitivity and specificity', Continuing Education in Anaesthesia, Critical Care & Pain, 2008, 8, (6), pp. 221-223

64 Han, J., Kamber, M., and Pei, J.: 'Data Mining: Concepts and Techniques' (Morgan Kaufmann Publishers Inc., 2011. 2011)

65 Ian H. Witten, F.E., Mark A. Hall.: 'Data Mining: Practical Machine Learning Tools and Techniques' (Morgan Kaufmann Publishers, 2011, Third Edition edn. 2011)

66 Tufféry, S.: 'Data Mining And Statistics For Decision Making' (WILEY, 2011. 2011)

67 Koller, D., and Friedman, N.: 'Probabilistic graphical models: principles and techniques' (MIT press, 2009. 2009)

68 Marsland, S.: 'Machine learning: an algorithmic perspective' (CRC Press, 2011. 2011)

69 Tomar, D., and Agarwal, S.: 'A survey on Data Mining approaches for Healthcare', International Journal of Bio-Science & Bio-Technology, 2013, 5, (5)

70 Zadrozny, B., and Elkan, C.: 'Learning and making decisions when costs and probabilities are both unknown', in Editor (Ed.)^(Eds.): 'Book Learning and making decisions when costs and probabilities are both unknown' (ACM, 2001, edn.), pp. 204-213

71 Smith, P.C., Newman, S., Ell, P., Treasure, T., Joseph, P., Schneidau, A., and Harrison, M.G.: 'Cerebral consequences of cardiopulmonary bypass', The Lancet, 1986, 327, (8485), pp. 823-825

72 AIHW: 'Disability in Australia: acquired brain injury', in Editor (Ed.)^(Eds.): 'Book Disability in Australia: acquired brain injury' (Authoritative information and statistics, 2007, edn.), pp.

73 Strömberg, A., and Jaarsma, T.: 'Thoughts about death and perceived health status in elderly patients with heart failure', European journal of heart failure, 2008, 10, (6), pp. 608-613

74 Bekker, A.Y., and Weeks, E.J.: 'Cognitive function after anaesthesia in the elderly', Best Practice & Research Clinical Anaesthesiology, 2003, 17, (2), pp. 259-272

75 Mazzola, P., Bellelli, G., Broggini, V., Anzuini, A., Corsi, M., Berruti, D., De Filippi, F., Zatti, G., and Annoni, G.: 'Postoperative delirium and pre-fracture disability predict 6-month mortality among the oldest old hip fracture patients', Aging clinical and experimental research, 2014, pp. 1-8

76 Reade, M.C., and Finfer, S.: 'Sedation and delirium in the intensive care unit', The New England journal of medicine, 2014, 370, (5), pp. 444-454

77 Ni Chonchubhair, A., Valacio, R., Kelly, J., and O'Keefe, S.: 'Use of the abbreviated mental test to detect postoperative delirium in elderly people', British journal of anaesthesia, 1995, 75, (4), pp. 481-482

78 Inouye, S.K., van Dyck, C.H., Alessi, C.A., Balkin, S., Siegal, A.P., and Horwitz, R.I.: 'Clarifying confusion: the confusion assessment method. A new method for detection of delirium', Ann Intern Med, 1990, 113, (12), pp. 941-948

79 Ely, E.W., Inouye, S.K., Bernard, G.R., Gordon, S., Francis, J., May, L., Truman, B., Speroff, T., Gautam, S., and Margolin, R.: 'Delirium in mechanically ventilated patients: validity and reliability of the confusion assessment method for the intensive care unit (CAM-ICU)', JAMA : the journal of the American Medical Association, 2001, 286, (21), pp. 2703-2710

80 ElBardissi, A.W., Aranki, S.F., Sheng, S., O'Brien, S.M., Greenberg, C.C., and Gammie, J.S.: 'Trends in isolated coronary artery bypass grafting: an analysis of the Society of Thoracic Surgeons adult cardiac surgery database', The Journal of thoracic and cardiovascular surgery, 2012, 143, (2), pp. 273-281

81 Ackoff, R.L.: 'From data to wisdom', Journal of Applied Systems Analysis, 1989, 16, pp. 3-9

82 Snijders, C., Matzat, U., and Reips, U.-D.: '" Big Data": Big Gaps of Knowledge in the Field of Internet Science', International Journal of Internet Science, 2012, 7, (1)

83 Berman, S.: 'Capitalizing on Complexity', IBM Global Business Services, Somers, USA, 2010

84 http://www.nlm.nih.gov/pubs/factsheets/medline.html, accessed 19 December 2013

85 Brian S. Alper, J.A.H., Susan G. Elliott,Scott Kinkade,Michael J. Hauan,Daniel K. Onion,Bernard M. Sklar,: 'How much effort is needed to keep up with the literature relevant for primary care?', Journal of Medical Library Association, 2004, 92, (4), pp. 429-437

86 Salama, G.I., Abdelhalim, M., and Zeid, M.A.-e.: 'Breast cancer diagnosis on three different datasets using multi-classifiers', Breast Cancer (WDBC), 2012, 32, (569), pp. 2

87 Chitra, R., and Seenivasagam, V.: 'REVIEW OF HEART DISEASE PREDICTION SYSTEM USING DATA MINING AND HYBRID INTELLIGENT TECHNIQUES', ISSN: 2229-6956 (ONLINE) ICTACT JOURNAL ON SOFT COMPUTING, 2013, 3, (04)

88 Ramani, R.G., and Sivagami, G.: 'Parkinson Disease Classification using Data Mining Algorithms', International Journal of Computer Applications, 2011, 32, (9)

89 Niaksu, O., Skinulyte, J., and Duhaze, H.G.: 'A Systematic Literature Review of Data Mining Applications in Healthcare', in Editor (Ed.)^(Eds.): 'Book A Systematic Literature Review of Data Mining Applications in Healthcare' (Springer, 2014, edn.), pp. 313-324

90 Izadi, M.T., and Buckeridge, D.L.: 'Decision theoretic analysis of improving epidemic detection', in Editor (Ed.)^(Eds.): 'Book Decision theoretic analysis of improving epidemic detection' (American Medical Informatics Association, 2007, edn.), pp. 354

91 Cooper, G.F., Dash, D.H., Levander, J.D., Wong, W.-K., Hogan, W.R., and Wagner, M.M.: 'Bayesian biosurveillance of disease outbreaks', in Editor (Ed.)^(Eds.): 'Book Bayesian biosurveillance of disease outbreaks' (AUAI Press, 2004, edn.), pp. 94-103

92 Bertsimas, D., Bjarnadóttir, M.V., Kane, M.A., Kryder, J.C., Pandey, R., Vempala, S., and Wang, G.: 'Algorithmic prediction of health-care costs', Operations Research, 2008, 56, (6), pp. 1382-1392

93 http://www.ncbi.nlm.nih.gov/pubmed/?term=medical+data+mining, accessed 31/January/2014 2014

94 Yan, S., Abidi, S.S., and Artes, P.H.: 'Analyzing Sub-Classifications of Glaucoma via SOM Based Clustering of Optic Nerve Images', Studies in health technology and informatics, 2005, 116, pp. 483-488

95 Qiang Ji, J.E., Eric Craine: 'Texture Analysis for Classification of Cervix Lesions', IEEE, 2000, 19, pp. 1144-1149

96 Mao, Y., Chen, Y., Hackmann, G., Chen, M., Lu, C., Kollef, M., and Bailey, T.C.: 'Early Deterioration Warning for Hospitalized Patients by Mining Clinical Data', International Journal of Knowledge Discovery in Bioinformatics (IJKDB), 2011, 2, (3), pp. 1-20

97 Pandey, A.K., Pandey, P., Jaiswal, K., and Sen, A.K.: 'DataMining Clustering Techniques in the Prediction of Heart Disease using Attribute Selection Method', heart disease, 14, pp. 16-17

98 http://www.sts.org/national-database., accessed May 22, 2013 2013

99 Peterson, J.F., Pun, B.T., Dittus, R.S., Thomason, J.W., Jackson, J.C., Shintani, A.K., and Ely, E.W.: 'Delirium and its motoric subtypes: a study of 614 critically ill patients', J Am Geriatr Soc, 2006, 54, (3), pp. 479-484

100 Canadian Institutes of Health Research, N.S., and Engineering Research Council of Canada, and Social Sciences and Humanities Research Council of Canada: 'Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans.', in Editor (Ed.)^(Eds.): 'Book Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans.' (Govermant of Canada, 2010, edn.), pp.

101 Microsoft: 'Microsoft Excel.', in Editor (Ed.)^(Eds.): 'Book Microsoft Excel.' (MIcrosoft, 2013, 2013 edn.), pp.

102 Inc, S.A.S.I.: 'SAS Software, Version 9.3', in Editor (Ed.)^(Eds.): 'Book SAS Software, Version 9.3' (S. A. S. Institute Inc, 2011, edn.), pp.

103 Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I.H.: 'The WEKA data mining software: an update', SIGKDD Explor. Newsl., 2009, 11, (1), pp. 10-18

104 R Devolopment Core Team: 'R: A language and enviroment for statistical computing', in Editor (Ed.)^(Eds.): 'Book R: A language and enviroment for statistical computing' (R Foundation for Statistical Computing, 2013, edn.), pp.

105 Robnik-Šikonja, M., and Kononenko, I.: 'Theoretical and empirical analysis of ReliefF and RReliefF', Machine learning, 2003, 53, (1-2), pp. 23-69

106 Sembiring, R.W., Zain, J.M., and Embong, A.: 'Dimension Reduction of Health Data Clustering', arXiv preprint arXiv:1110.3569, 2011

107 Orlitsky, A.: 'Supervised dimensionality reduction using mixture models', in Editor (Ed.)^(Eds.): 'Book Supervised dimensionality reduction using mixture models' (ACM, 2005, edn.), pp. 768-775

108 Saeys, Y., Inza, I., and Larranaga, P.: 'A review of feature selection techniques in bioinformatics', Bioinformatics, 2007, 23, (19), pp. 2507-2517

109 Kumar, D.A., and Annie, M.: 'Clustering dichotomous data for health care', International Journal of Information Sciences and Techniques (IJIST), 2012, 2, (2)

110 Belciug, S., Gorunescu, F., Salem, A.-B., and Gorunescu, M.: 'Clustering-based approach for detecting breast cancer recurrence', in Editor (Ed.)^(Eds.): 'Book Clustering-based approach for detecting breast cancer recurrence' (IEEE, 2010, edn.), pp. 533-538

111 Escudero, J., Zajicek, J., and Ifeachor, E.: 'Early detection and characterization of Alzheimer's disease in clinical scenarios using Bioprofile concepts and K-means', in Editor (Ed.)^(Eds.): 'Book Early detection and characterization of Alzheimer's disease in clinical scenarios using Bioprofile concepts and K-means' (IEEE, 2011, edn.), pp. 6470-6473

112 Liu, Z., Sokka, T., Maas, K., Olsen, N.J., and Aune, T.M.: 'Prediction of disease severity in patients with early rheumatoid arthritis by gene expression profiling', Human Genomics and Proteomics, 2009, 1, (1), pp. 484351

113 Chipman, H., and Tibshirani, R.: 'Hybrid hierarchical clustering with applications to microarray data', Biostatistics, 2006, 7, (2), pp. 286-301

114 Santhi, P., and Bhaskaran, V.M.: 'Performance of clustering algorithms in healthcare database', International Journal for Advances in Computer Science, 2010, 2, (1), pp. 26-31

115 Do, C.B., and Batzoglou, S.: 'What is the expectation maximization algorithm?', Nature biotechnology, 2008, 26, (8), pp. 897-900

116 Angelova, A.: 'EM algorithm updates for dimensionality reduction using automatic supervision', in Editor (Ed.)^(Eds.): 'Book EM algorithm updates for dimensionality reduction using automatic supervision' (Technical report, 2007, edn.), pp.

117 Raeder, T., Perlich, C., Dalessandro, B., Stitelman, O., and Provost, F.: 'Scalable supervised dimensionality reduction using clustering', in Editor (Ed.)^(Eds.): 'Book Scalable supervised dimensionality reduction using clustering' (ACM, 2013, edn.), pp. 1213-1221

118 Farhangfar, A., Kurgan, L., and Dy, J.: 'Impact of imputation of missing values on classification error for discrete data', Pattern Recognition, 2008, 41, (12), pp. 3692-3705

119 Su, X., Khoshgoftaar, T.M., and Greiner, R.: 'Using imputation techniques to help learn accurate classifiers', in Editor (Ed.)^(Eds.): 'Book Using imputation techniques to help learn accurate classifiers' (IEEE, 2008, edn.), pp. 437-444

120 Su, X., Greiner, R., Khoshgoftaar, T.M., and Napolitano, A.: 'Using classifier-based nominal imputation to improve machine learning': 'Advances in Knowledge Discovery and Data Mining' (Springer, 2011), pp. 124-135

121 Acuna, E., and Rodriguez, C.: 'The treatment of missing values and its effect on classifier accuracy': 'Classification, Clustering, and Data Mining Applications' (Springer, 2004), pp. 639-647

122 Katz, S.: 'Assessing self-maintenance: activities of daily living, mobility, and instrumental activities of daily living', Journal of the American Geriatrics Society, 1983

123 http://www.merriam-webster.com/dictionary/classification, accessed April 2014

124 Bishop, C.M.: 'Pattern recognition and machine learning' (springer New York, 2006. 2006)

125 Murphy, K.P.: 'Machine learning: a probabilistic perspective' (MIT Press, 2012. 2012)

126 Eng, J.: 'Receiver Operating Characteristic Analysis: A Primer1', Academic radiology, 2005, 12, (7), pp. 909-916

127 Fawcett, T.: 'ROC graphs: Notes and practical considerations for researchers', Machine learning, 2004, 31, pp. 1-38

128 http://scott.fortmann-roe.com/docs/MeasuringError.html, accessed 15th May 2014

129 Fawcett, T.: 'An introduction to ROC analysis', Pattern recognition letters, 2006, 27, (8), pp. 861-874

130 Hanley, J.A.: 'Characteristic (ROC) Curvel', Radiology, 1982, 743, pp. 29-36

131 Aslan, O., Yıldız, O.T., and Alpaydın, E.: 'Statistical Comparison of Classifiers Using Area Under the ROC Curve'

132 Hanley, J.A., and McNeil, B.J.: 'A method of comparing the areas under receiver operating characteristic curves derived from the same cases', Radiology, 1983, 148, (3), pp. 839-843

133 Bostanci, B., and Bostanci, E.: 'An Evaluation of Classification Algorithms Using Mc Nemar's Test', in Editor (Ed.)^(Eds.): 'Book An Evaluation of Classification Algorithms Using Mc Nemar's Test' (Springer, 2013, edn.), pp. 15-26

134 Di Eugenio, B., and Glass, M.: 'The kappa statistic: A second look', Computational linguistics, 2004, 30, (1), pp. 95-101

135 Viera, A.J., and Garrett, J.M.: 'Understanding interobserver agreement: the kappa statistic', Fam Med, 2005, 37, (5), pp. 360-363

136 Nelsen, R.B.: 'Proofs without words: Exercises in visual thinking' (MAA, 1993. 1993)

137 Batista, G.E., Prati, R.C., and Monard, M.C.: 'A study of the behavior of several methods for balancing machine learning training data', ACM Sigkdd Explorations Newsletter, 2004, 6, (1), pp. 20-29

138 Altman, D.G., Vergouwe, Y., Royston, P., and Moons, K.G.: 'Prognosis and prognostic research: validating a prognostic model', BMJ: British Medical Journal, 2009, pp. 1432-1435

139 Edelstein, P.: 'Emerging directions in analytics. Predictive analytics will play an indispensable role in healthcare transformation and reform', Health management technology, 2013, 34, (1), pp. 16-17

140 Ginsburg, G.S., Staples, J., and Abernethy, A.P.: 'Academic medical centers: ripe for rapid-learning personalized health care', Science translational medicine, 2011, 3, (101), pp. 101cm127-101cm127

141 Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., and Philip, S.Y.: 'Top 10 algorithms in data mining', Knowledge and Information Systems, 2008, 14, (1), pp. 1-37

142 Koh, H.C., and Tan, G.: 'Data mining applications in healthcare', Journal of Healthcare Information Management—Vol, 2011, 19, (2), pp. 65

143 Li, L., Tang, H., Wu, Z., Gong, J., Gruidl, M., Zou, J., Tockman, M., and Clark, R.A.: 'Data mining techniques for cancer detection using serum proteomic profiling', Artificial intelligence in medicine, 2004, 32, (2), pp. 71-83

144 Ayer, T., Chhatwal, J., Alagoz, O., Kahn, C.E., Woods, R.W., and Burnside, E.S.: 'Comparison of logistic regression and artificial neural network models in breast cancer risk estimation', Radiographics, 2010, 30, (1), pp. 13-22

145 Nookala, G.K.M., Orsu, N., Pottumuthu, B.K., and Mudunuri, S.B.: 'Performance Analysis and Evaluation of Different Data Mining Algorithms used for Cancer Classification', International Journal, 2013

146 Acid, S., de Campos, L.M., Fernández-Luna, J.M., Rodrıguez, S., María Rodríguez, J., and Luis Salcedo, J.: 'A comparison of learning algorithms for Bayesian networks: a case study based on data from an emergency medical service', Artificial intelligence in medicine, 2004, 30, (3), pp. 215-232

147 Watt, E.W., and Bui, A.A.: 'Evaluation of a dynamic bayesian belief network to predict osteoarthritic knee pain using data from the osteoarthritis initiative', in Editor (Ed.)^(Eds.): 'Book Evaluation of a dynamic bayesian belief network to

predict osteoarthritic knee pain using data from the osteoarthritis initiative' (American Medical Informatics Association, 2008, edn.), pp. 788

148 Lisboa, P.J.: 'A review of evidence of health benefit from artificial neural networks in medical intervention', Neural networks, 2002, 15, (1), pp. 11-39

149 Parsaeian, M., Mohammad, K., Mahmoudi, M., and Zeraati, H.: 'Comparison of Logistic Regression and Artificial Neural Network in Low Back Pain Prediction: Second National Health Survey', Iranian J Publ Health, 2012, 41, (6), pp. 86-92

150 Eftekhar, B., Mohammad, K., Ardebili, H.E., Ghodsi, M., and Ketabchi, E.: 'Comparison of artificial neural network and logistic regression models for prediction of mortality in head trauma based on initial clinical data', BMC Medical Informatics and Decision Making, 2005, 5, (1), pp. 3

151 Acid, S., de Campos, L.M., Fernandez-Luna, J.M., Rodriguez, S., Maria Rodriguez, J., and Luis Salcedo, J.: 'A comparison of learning algorithms for Bayesian networks: a case study based on data from an emergency medical service', Artificial intelligence in medicine, 2004, 30, (3), pp. 215-232

152 Carpio, K., and Hermosilla, A.: 'On multicollinearity and artificial neural networks', Compexity International, Monash University, AU.(draft manuscript: http://www. complexity. org. au/ci/draft/draft/hermos01/hermos01s. pdf), 2002

153 Bernardo, J.M., and Smith, A.F.: 'Bayesian theory' (John Wiley & Sons, 2009. 2009)

154 Lauritzen, S.L., and Spiegelhalter, D.J.: 'Local computations with probabilities on graphical structures and their application to expert systems', Journal of the Royal Statistical Society. Series B (Methodological), 1988, pp. 157-224

155 Pearl, J.: 'Probabilistic reasoning in intelligent systems: networks of plausible inference' (Morgan Kaufmann, 1988. 1988)

156 Tillman, R.E.: 'Learning Directed Graphical Models from Nonlinear and Non-Gaussian Data'

157 Zhang, L., and Kim Roddis, W.: 'Machine learning in updating predictive models of planning and scheduling transportation projects', Transportation Research Record: Journal of the Transportation Research Board, 1997, 1588, (1), pp. 86-94

158 Toll, D., Janssen, K., Vergouwe, Y., and Moons, K.: 'Validation, updating and impact of clinical prediction rules: a review', Journal of clinical epidemiology, 2008, 61, (11), pp. 1085-1094

159 Waljee, A.K., Higgins, P.D., and Singal, A.G.: 'A Primer on Predictive Models', Clinical and translational gastroenterology, 2014, 5, (1), pp. e44

160 Moons, K.G., Kengne, A.P., Grobbee, D.E., Royston, P., Vergouwe, Y., Altman, D.G., and Woodward, M.: 'Risk prediction models: II. External validation, model updating, and impact assessment', Heart, 2012, pp. heartjnl-2011-301247

161 http://canworksmart.com/how-often-should-you-update-predictive-models/, accessed June 13 2014

162 http://canworksmart.com/why-update-predictive-models/, accessed June 13 2014

163 Baxt, W.G.: 'Use of an artificial neural network for the diagnosis of myocardial infarction', Annals of Internal Medicine, 1991, 115, (11), pp. 843-848

164 Artis, S.G., Mark, R., and Moody, G.: 'Detection of atrial fibrillation using artificial neural networks', in Editor (Ed.)^(Eds.): 'Book Detection of atrial fibrillation using artificial neural networks' (IEEE, 1991, edn.), pp. 173-176

165 Joshi, S., Shenoy, D., Vibhudendra Simha, G., Rrashmi, P., Venugopal, K., and Patnaik, L.: 'Classification of Alzheimer's disease and Parkinson's disease by using machine learning and neural network methods', in Editor (Ed.)^(Eds.): 'Book Classification of Alzheimer's disease and Parkinson's disease by using machine learning and neural network methods' (IEEE, 2010, edn.), pp. 218-222

166 Abramovici, M., Neubach, M., Fathi, M., and Holland, A.: 'Competing fusion for bayesian applications', in Editor (Ed.)^(Eds.): 'Book Competing fusion for bayesian applications' (2008, edn.), pp. 379

167 Salehi, E., and Gras, R.: 'An empirical comparison of the efficiency of several local search heuristics algorithms for bayesian network structure learning', in Editor (Ed.)^(Eds.): 'Book An empirical comparison of the efficiency of several local search heuristics algorithms for bayesian network structure learning' (2009, edn.), pp.

168 Vandel, J., Mangin, B., and de Givry, S.: 'New local move operators for Bayesian network structure learning', Proceedings of PGM-12, Granada, Spain, 2012

169 Koivisto, M., and Sood, K.: 'Exact Bayesian structure discovery in Bayesian networks', The Journal of Machine Learning Research, 2004, 5, pp. 549-573
170 Silander, T., and Myllymaki, P.: 'A simple approach for finding the globally optimal Bayesian network structure', arXiv preprint arXiv:1206.6875, 2012
171 Bouckaert, R.R.: 'Bayesian network classifiers in weka' (Department of Computer Science, University of Waikato, 2004. 2004)
172 Scholz, F.: 'Maximum likelihood estimation', Encyclopedia of Statistical Sciences, 1985
173 SAS: 'Iterative Algorithms for Model Fitting', 2014, 2014, (April 13)
174 Yang, Q., and Wu, X.: '10 challenging problems in data mining research', International Journal of Information Technology & Decision Making, 2006, 5, (04), pp. 597-604
175 Estabrooks, A., Jo, T., and Japkowicz, N.: 'A multiple resampling method for learning from imbalanced data sets', Computational Intelligence, 2004, 20, (1), pp. 18-36
176 Pyle, D.: 'Data preparation for data mining' (Morgan Kaufmann, 1999. 1999)
177 Williams, D.P., Myers, V., and Silvious, M.S.: 'Mine classification with imbalanced data', Geoscience and Remote Sensing Letters, IEEE, 2009, 6, (3), pp. 528-532
178 Liu, X.-Y., Wu, J., and Zhou, Z.-H.: 'Exploratory undersampling for class-imbalance learning', Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 2009, 39, (2), pp. 539-550
179 Weiss, G.M., and Provost, F.J.: 'Learning when training data are costly: the effect of class distribution on tree induction', J. Artif. Intell. Res.(JAIR), 2003, 19, pp. 315-354
180 , !!! INVALID CITATION !!!
181 Wennberg, D.: 'Systems and methods for predicting healthcare related risk events', in Editor (Ed.)^(Eds.): 'Book Systems and methods for predicting healthcare related risk events' (Google Patents, 2005, edn.), pp.
182 Dhar, V.: 'Data science and prediction', Communications of the ACM, 2013, 56, (12), pp. 64-73

183 RD., M.: 'Simplified Calculation of Body-Surface Area', New England journal of medicine, 1987, 317, (17), pp. 1098-1098

184 Cockcroft, D.W., and Gault, M.H.: 'Prediction of creatinine clearance from serum creatinine', Nephron, 1976, 16, (1), pp. 31-41

185 Marx, G., Blake, G., Galani, E., Steer, C., Harper, S., Adamson, K., Bailey, D., and Harper, P.: 'Evaluation of the Cockroft–Gault, Jelliffe and Wright formulae in estimating renal function in elderly cancer patients', Annals of oncology, 2004, 15, (2), pp. 291-295

186 Liao, W.-k., Liu, Y., and Choudhary, A.: 'A grid-based clustering algorithm using adaptive mesh refinement', in Editor (Ed.)^(Eds.): 'Book A grid-based clustering algorithm using adaptive mesh refinement' (2004, edn.), pp.

187 CEPPELLINI, B.R., Siniscalco, M., and Smith, C.: 'THE ESTIMATION OF GENE FREQUENCIES IN A RANDOM‐MATING POPULATION', Annals of Human Genetics, 1955, 20, (2), pp. 97-115

188 Tufféry, S.: 'Overview of Data Mining': 'Data Mining and Statistics for Decision Making' (John Wiley & Sons, Ltd, 2011), pp. 1-24

189 Smyth, P.: 'Clustering Using Monte Carlo Cross-Validation', in Editor (Ed.)^(Eds.): 'Book Clustering Using Monte Carlo Cross-Validation' (1996, edn.), pp. 126-133

190 Thorndike, R.L.: 'Who belongs in the family?', Psychometrika, 1953, 18, (4), pp. 267-276

191 Schorfheide, F., and Wolpin, K.I.: 'On the use of holdout samples for model selection', The American Economic Review, 2012, 102, (3), pp. 477-481

192 http://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_probability/BS704_Probability.html, accessed May 10 2014

193 Adeney, K.L., Weiss, N.S., and Shoben, A.B.: 'Epidemiology and biostatistics: an introduction to clinical research' (Springer, 2009. 2009)

194 Olsen, J., Christensen, K., Ekbom, A., and Murray, J.: 'An introduction to epidemiology for health professionals' (Springer, 2010. 2010)

195 Schoenbach, V.J., and Rosamond, W.D.: 'Understanding the fundamentals of epidemiology: an evolving text', Chapel Hill: North Carolina, 2000

196 Zou, K.H., O'Malley, A.J., and Mauri, L.: 'Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models', Circulation, 2007, 115, (5), pp. 654-657

197 McNemar, Q.: 'Note on the sampling error of the difference between correlated proportions or percentages', Psychometrika, 1947, 12, (2), pp. 153-157

198 CHAWLA, N.: 'Synthetic Minority Over-sampling Technique. 2002. 37p', Journal of Artificial Intelligence research

199 Morik, K., Brockhausen, P., and Joachims, T.: 'Combining statistical learning with a knowledge-based approach: a case study in intensive care monitoring', in Editor (Ed.)^(Eds.): 'Book Combining statistical learning with a knowledge-based approach: a case study in intensive care monitoring' (Technical Report, SFB 475: Komplexitätsreduktion in Multivariaten Datenstrukturen, Universität Dortmund, 1999, edn.), pp.