NOTE TO USERS

This reproduction is the best copy available.

UMI[®]

An Integrated Decision Tree-Artificial Neural Network

Hybrid to Estimate clinical outcomes:

ICU mortality and pre-term birth

by

Nicole Yu, B.Sc. (Joint Advanced Major in Chemistry and Biology),

Saint Francis Xavier University (Antigonish, Nova Scotia)

A thesis submitted to The Faculty of Graduate Studies and Research

In partial fulfillment of the requirements for the degree of

Master of Applied Science in Biomedical Engineering

Ottawa -Carleton Institute for Biomedical Engineering (OCIBME)

Department of Systems and Computer Engineering

Carleton University

Ottawa, Ontario, Canada, K1S 5B6

December 2009

©2009, Nicole Yu

Library and Archives
Canada

Published Heritage
Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file *Votre référence*
ISBN: 978-0-494-63825-5
Our file *Notre référence*
ISBN: 978-0-494-63825-5

NOTICE:

The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protègent cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

Abstract

Engineering and designing an artificial intelligence tool for pattern classification has promising use in the field of medicine. The data mining approach integrates Decision Trees (DTs); Artificial Neural Networks (ANNs), specifically a Classification-based Multi-Layer Perceptron (MLP); and a MLP ANN with risk stratification. This tool predicted mortality in an adult intensive care unit: sensitivity of 75.0%, specificity of 89.54% and an area under the curve (AUC) of 0.9417 for postoperative cases; non-postoperative cases had a sensitivity of 90.90%, specificity 75.16% and an AUC of 0.8333. Prediction of high-risk preterm birth had a sensitivity of 65.13%, specificity of 84.07% and an AUC of 0.8195 for Parous cases and a sensitivity of 61.08%, specificity of 71.14%, and an AUC of 0.7195 for Nulliparous cases. The trained integrated tool reduced the complexity while yielding prediction accuracies that exceed current results in the literature.

Fear will only hinder the route to endless possibilities

Acknowledgments

To all my professors, especially my thesis supervisor, Dr. Monique Frize: your constant support, encouragement and motivation are inspiring and I am very grateful to have had the opportunity to work with you.

To all my colleagues: Your support, friendly advice, and life experience has been a sanction in this unknown territory. I am appreciative of your willingness to take time out of your own busy schedule to share and pass your knowledge onto me.

To my family, especially my two loving parents, Raymundo and Marilyn Yu: You gave me choices, encouraged me to persevere, think positive thoughts and to climb those mountains. You have been my rock throughout my educational career and taught me to see the light during its most challenging moments. For this, I am forever appreciative. Also, my sincere gratitude to my sister, Patricia: Your love for food kept me fed and functioning.

To Jeremy Bayuga: Your adventurous nature gave me a chance to explore Ottawa outside the classroom walls. Your belief in my skills was motivating. Thank you for your support, encouragement, and expectations to know of other important things in life, such as pop culture and snowboarding.

Finally, to all my dearest friends: Your lasting confidence in me will always be remembered. Though times have changed and being together is short and far apart; I will always look forward and value our moments together.

Contents

1	Introduction, Medical Environment and Problem Statement	1
1.1	Motivation	1
1.1.1	Health Care Perspective:	1
1.1.2	Technology Perspective	2
1.2	Problem Statement	2
1.3	Introduction of the Medical Environment	3
1.4	The Intensive Care Unit (ICU)	4
1.4.1	Cost of ICU	4
1.4.2	The Need for Prediction Models in an Adult ICU	5
1.5	The Importance of Delivering to Term	6
1.5.1	Rate of Preterm Birth (PTB)	6
1.5.2	Health of Preterm Infants	7
1.5.3	Cost of PTB	8
1.6	Thesis Objectives	8
1.6.1	First Objective	8
1.6.2	Second Objective	9
1.7	Thesis Outline	10
2	Literature Review:	
	Outcome Estimation Systems and Technology	11
2.1	Prognostic Models in the Adult ICU	11
2.2	Critical Analysis for the Intensive Care Unit	11
2.2.1	The Accuracy of Standard Scoring Systems for an Adult ICU	11
2.2.2	Limitations	14
2.3	Types of Preterm Birth	15
2.4	Factors Contributing to Preterm Birth	16
2.4.1	Current Knowledge	18
2.5	Critical Analysis for Preterm Birth	19
2.5.1	Screening Tests for Preterm Birth	19
2.5.2	Heterogeneity	20
2.5.3	Accuracy Prediction of Cervical Length (CL)	20
2.5.4	Accuracy Prediction fetal fibronectin (fFN)	22
2.5.5	Risk Factor Scoring	25
2.6	Current Prediction Models for Preterm Birth	26
2.6.1	Preterm Birth Prediction	26
2.6.2	Uterine Electromyography ANN prediction	26
2.6.3	Early Prediction of Preterm Birth	27
2.6.4	Limitations to Current PTB Prediction Methods	27

2.7	Pattern Classification Methods	28
2.7.1	Hybrid Artificial Neural Network	28
2.7.2	Tree-based Neural Network (TBNN)	29
2.7.3	Pruning-based Neural Networks (PBNN)	29
2.8	Pattern Classification in MIRC	30
2.8.1	ANN Research Framework for Medical Outcome	30
2.8.2	Skewed <i>A Priori</i> Outcome	30
2.8.3	Techniques to Improve Performance	31
2.8.4	Weight-elimination	32
2.8.5	Probability Estimation of Classification Outcomes	32
2.8.6	Predicting High-Risk Preterm Birth —Initial Findings	33
2.8.7	Critical Analysis:	35
2.9	Contributions to Support Systems	35
3	Methodology	37
3.1	Restating the Problem	37
3.2	Ethical Clearance	38
3.3	The Adult ICU Database	39
3.4	The PRAMS Database	39
3.5	Performance Measures of Data Mining Methods	40
3.5.1	Confusion Matrix (Contingency Table)	40
3.5.2	Correct Classification Rate (CCR)	41
3.5.3	Error Rate	41
3.5.4	Sensitivity	41
3.5.5	Specificity	42
3.5.6	Prevalence	42
3.5.7	Positive Predictive Value (PPV)	42
3.5.8	Negative Predictive Value (NPV)	43
3.5.9	Receiver Operating Characteristic (ROC) Curve	43
3.5.10	Area Under the Curve (AUC)	44
3.6	Normalization/Standardization	45
3.7	Using Hybrid Models	45
3.7.1	Feature Selection	46
3.7.2	Decision Tree Ensemble Classifiers	46
3.7.3	Attribute Interval	47
3.7.4	Logarithmic Sensitivity Index	49
3.7.5	Performance Measures for ANNs and DTs	49
3.7.6	Developing the DT-ANN Hybrid models	50
3.8	First Objective	51
3.8.1	Accomplishing Objective #1 - Methodology	51
3.8.2	Data Processing	52
3.8.3	Creating the DT Hybrid Input Variables (DThy)	53
3.8.4	ANN Classification	54
3.8.5	Reproducing the classifiers	56

3.9	Second Objective	56
3.9.1	Accomplishing Objective #2 - Methodology	58
3.9.2	Establishing the Risk Model	58
3.9.3	Data Processing	59
3.9.4	Creating the DT Hybrid Input Variables (DThy)	60
3.9.5	ANN classification	62
3.10	Methodology Conclusion	63
4	Results and Discussion	64
4.1	Creating the Attribute Intervals using Decision Trees	64
4.1.1	POSTOP Decision Tree Results	64
4.1.2	noPOSTOP Decision Tree Results	64
4.2	POSTOP ANN Results	66
4.2.1	POSTOP DT-ANN Hybrid Results	66
4.2.2	POSTOP ANN with weight-elimination	68
4.2.3	Comparing POSTOP DT-ANN Hybrid with we_ANN	70
4.3	noPOSTOP ANN Results	71
4.3.1	noPOSTOP DT-ANN Hybrid Results	71
4.3.2	noPOSTOP ANN with Weight-elimination	73
4.3.3	Comparing noPOSTOP DT-ANN Hybrid and we_ANN	75
4.3.4	Hybrid Classifier Results and Number of Nodes	76
4.4	Assessing PPV and NPV for Predicting Mortality	78
4.4.1	POSTOP Hybrid PPV and NPV	78
4.4.2	POSTOP we_ANN PPV and NPV	78
4.4.3	noPOSTOP Hybrid PPV and NPV	79
4.4.4	noPOSTOP we_ANN PPV and NPV	80
4.5	Prediction of High-Risk Preterm Birth	81
4.5.1	Decision Tree Results	81
4.5.2	HR-PTB Hybrid Classifier Results and Number of Nodes	84
4.5.3	Hybrid Classification ANN Results	86
4.5.4	Risk Stratification Results	88
4.6	PPV and NPV for HR-PTB	90
4.7	Discussion - Mortality Prediction	91
4.7.1	Predicting mortality in an Adult ICU	91
4.7.2	Comparing the DT-ANN Hybrid and we_ANN Method	93
4.8	Discussion - HR-PTB Prediction	96
5	Concluding Remarks, Contributing Knowledge and Future Work	99
5.1	Conclusion	99
5.1.1	Accomplished Preliminary Objectives	99
5.2	Contribution to Knowledge	103
5.3	Future Work	105
	Bibliography	107

Appendix	119
A Tables	120
B Data Formatting	127
B.1 Data Pre-processing	127
B.1.1 Missing Values	127
B.2 PRAMS Data Processing	128
C Modeling Cost Functions and Deriving a Training Algorithm	132
D Mortality Results	134
D.1 POSTOP Final Classifier Performance Results	134
D.2 noPOSTOP Final Classifier Performance Results	139
D.3 POSTOP and noPOSTOP—Hybrid Classifier Biases and Weights	143
E High-Risk Preterm Birth Results	145
E.1 HR-PTB Final Classifier Performance Results	145
E.2 HR-PTB —Hybrid Risk Stratification Biases and Weights	148
F Artificial Neural Networks (ANNs)	156
F.1 Artificial Neural Networks (ANNs)	156
F.1.1 ANN Application	156
F.1.2 The Multi-Layer Perceptron (MLP)	157
F.1.3 Activation Function	159
F.1.4 Feed-forward and Learning	159
F.1.5 Backpropagation	160
F.1.6 Weight Decay and Weight Elimination	162
F.1.7 Steps	162
F.2 Ann Research Framework (ANN-RFW)	163
F.2.1 Parameters	163
F.2.2 Early Stopping Criterion	164
G Decision Trees	165
G.1 Decision Trees (DTs)	165
G.1.1 What is a Decision Tree?	165
G.1.2 First Step —Constructing a Decision Tree	165
G.1.3 Second Step —Pruning Options	167
G.1.4 Third Step —Rule Induction Based on the Pruned Tree	168
G.1.5 Fuzzy Thresholds	168
G.1.6 Prior frequency and Cost	169
G.1.7 Advantages of a DT	169
G.1.8 Disadvantages of a DT	170

H	Additional Model Estimation Techniques	171
H.1	Model Estimation	171
H.2	Bootstrap Methods	171
H.3	Ensembles for Feature Selection	173
H.4	Average Squared Error	173
H.5	Hosmer-Lemeshow	174
J	Atlee Application	175

List of Tables

2.1	The Four Generations of the ICU Severity Prognosis Models	12
2.2	Accuracy of fFN Studies for the Prediction of PTB	24
3.1	Confusion Matrix	41
3.2	AUC index	45
3.3	Division of <i>POSTOP</i> and <i>noPOSTOP</i> for DT classifier	54
3.4	Division of <i>POSTOP</i> and <i>noPOSTOP</i> database for ANN	55
3.5	Division of <i>Parous</i> and <i>Nulliparous</i> for DT classifier	61
3.6	Division of <i>POSTOP</i> and <i>noPOSTOP</i> database for ANN	63
4.1	<i>POSTOP</i> Decision Tree Results	65
4.2	<i>noPOSTOP</i> Decision Tree Results	65
4.3	<i>POSTOP</i> Attribute Interval Splitting Values	66
4.4	<i>noPOSTOP</i> Attribute Interval Splitting Value	66
4.5	<i>POSTOP</i> Best Performing Hybrid ANN without weight-elimination	67
4.6	<i>POSTOP</i> Best Performing ANN with weight elimination	70
4.7	<i>POSTOP</i> Comparing the classifiers	71
4.8	<i>noPOSTOP</i> Best Performing Hybrid ANN without weight-elimination	73
4.9	<i>noPOSTOP</i> Best Performing ANN with weight elimination	75
4.10	<i>noPOSTOP</i> Comparing the classifiers	76
4.11	<i>POSTOP</i> Hybrid PPV and NPV	78
4.12	<i>POSTOP</i> we_ANN PPV and NPV	79
4.13	<i>noPOSTOP</i> Hybrid PPV and NPV	80
4.14	<i>noPOSTOP</i> we_ANN PPV and NPV	81
4.15	<i>PAROUS</i> Decision Tree Results	83
4.16	<i>NULLIPAROUS</i> Decision Tree Results	84
4.17	<i>PAROUS</i> Best Performing Hybrid Classification ANN with WE	87
4.18	<i>NULLIPAROUS</i> Best Performing Hybrid Classification ANN with WE	87
4.19	<i>PAROUS</i> Best Performing Hybrid Risk Stratification ANN with WE	89
4.20	<i>NULLIPAROUS</i> Best Performing Hybrid Risk Stratification ANN with WE	89
4.21	<i>PAROUS</i> Comparing Classification ANN with Risk Stratification ANN	90
4.22	<i>NULLIPAROUS</i> Comparing Classification ANN with Risk Stratification ANN	90
4.23	HR-PTB PPV and NPV	91
A.1	Original Adult ICU Variables	120
A.2	<i>POSTOP</i> and <i>noPOSTOP</i> mean and standard deviation	121
A.3	ICU ANN starting paramaters	121
A.4	PRAMS Input Variable List for Decision Tree	122

A.5	PAROUS DT Splitting Values	124
A.6	NULLIPAROUS DT Splitting Values	125
A.7	Parous ANN starting paramaters	126
A.8	PAROUS Final Hybrid Network Parameters	126
A.9	NULLIPAROUS Final Hybrid Network Parameters	126
D.1	POSTOP Final Classifier Performance Results	134
D.2	noPOSTOP Final Classifier Performance Results	139
E.1	HR-PTB Final Classifier Performance Results	145
H.1	Methods for Creating and Testing a Classifier	172

List of Figures

2.1	Predicted Probability of CL and Time of Measurement	22
3.1	An ROC Curve	44
3.2	Attribute Intervals	48
4.1	POSTOP Hybrid Tr1 logsens Plot	77
4.2	noPOSTOP Hybrid Tr2 logsens Plot	77
4.3	Parous Hybrid logsens Plot	85
4.4	Nulliparous Hybrid logsens Plot	85
F.1	Two-Layer Neural Network Types	157
F.2	Three-layered Neural Network	158
F.3	Neural Network Structure	158
F.4	Activation Function	159
F.5	Learning Rates of an ANN	161
G.1	Diagram of a Decision Tree	166

List of Acronyms

ANN	Artificial Neural Network
APGAR	Activity, Pulse, Grimace, Appearance and Respiration
APACHE	Acute Physiology and Chronic Health Evaluation
ASE	Average Squared Error
AUC	Area Under the Curve
CCR	Correct Classification Rate
CDSS	Clinical Decision Support System
CHEO	Children’s Hospital of Eastern Ontario
CL	Cervical Length
DT	Decision Tree
fFN	Fetal Fibronectin
FN	False Negative
FP	False Positive
GDM	Gestational Diabetes Mellitus
HR-PTB	High-risk Preterm Birth
ICU	Intensive Care Unit
LBW	Low Birthweight
logsens	Logarithmic Sensitivity Index
MIRG	Medical Information-technology Research Group
MPM	Mortality Probability Models
MSE	Mean Squared Error
MV	Mechanical Ventilation
NICU	Neonatal Intensive Care Unit
NPV	Negative Predictive Value

PBNN	Pruning Based Neural Network
PPESO	Perinatal Partnership Program of Eastern and Southern Ontario
PPV	Positive Predictive Value
PTB	Preterm Birth
PRAMS	Pregnancy Risk Assessment Monitoring System
ROC	Receiver Operating Characteristic
SAPS	Simplified Acute Physiology Score
TBNN	Tree Based Neural Network
TN	True Negative
TP	True Positive

Introduction, Medical Environment and Problem Statement

1.1 Motivation

1.1.1 Health Care Perspective:

The field of medicine is rapidly evolving, giving rise to sophisticated technology - from high resolution diagnostic imaging [1] to machines that temporarily replace dysfunctional organs [2]. The possibilities seem endless and though the technology may exist - resources and money are limited [3]. Many systems are in place to provide an efficient management of health care resources [4, 5], a field called Clinical Decision Support System (CDSS). A CDSS is put in place to aid the physician's decision making process and complement his/her own experience and knowledge [6]. One area targeted in CDSS is the successful identification of future medical conditions. Accurate prediction of medical outcomes would allow health care systems efficient allocation of their resources (i.e. proper equipment, needed physicians and other health care professionals) and earlier planning of interventions. Unfortunately, this area remains a challenge as many standard clinical methods have had moderate success [7, 8, 9]. Many standard methods use physiological and biochemical indicators as predictors for medical conditions. Though the indicators of interest may increase the risk and are highly associated with the medical outcome they are usually not directly related; the presence of the indicator does not necessarily mean that the patient will succumb to the disease. The moderate success of disease prediction this field indicates that the relationship between these variables are not as simple as the standard clinical prediction methods that use these indicators [10, 11, 12].¹ By linking together medical ideas and advanced mathematical concepts, a sound tool may be designed to improve the

¹evidence of this can be found in Section 2.2.2 and Section 2.5

prediction of complicated outcomes found in the medical field.

1.1.2 Technology Perspective

Medical Information-technology Research Group (MIRG) has been working in this area, applying advanced mathematical application tools to solve medical problems, a field called artificial intelligence which includes data mining approaches. Data mining approaches (also known as knowledge discovery) is the process of extracting information from data for the use of prediction, description, explanation or exploration of a phenomenon using pattern classification approaches [13]. MIRG has identified Artificial Neural Networks (ANNs) as an ideal pattern classification approach [14, 15, 16] as it applies advanced, non-parametric techniques to model complicated medical data. ANNs use a mathematical model of numerous neurons connected in a network to model the behavioural characteristics of biological nervous systems [17].² Another approach is the use of Decision Trees (DTs). DTs classify by splitting the dataset according to feature values that best discriminate between outcomes.³ Complementing the power of ANNs (ability to model non-linear relationships, its robustness, and learning implementation [17]) with DTs (ability to handle nominal data, missing values and its quick output) [18]) can create an improved prediction system [19, 20]. An improved system can increase the efficiency and credibility of a CDSS.

1.2 Problem Statement

The purpose of prediction models is to successfully predict an outcome of choice. In this thesis work an integrated system was developed to predict dichotomous medical outcomes - the presence (positive) or absence (negative) of a condition. Though many medical prognosis systems created attempt to identify positive cases - successful identification of

²For a detailed explanation of ANNs refer to Appendix F

³For a detailed explanation of DTs, refer to Appendix G

positive cases proves to be difficult. The observation of positive outcomes is often a rare occurrence compared to the presence of negative outcomes - thus many systems are unable to classify positive cases. Additionally, irrelevant variables add complexity and noise of the problem reducing the ability for the prognosis system to identify positive cases. This is an issue as successful identification could mean earlier treatment, intervention, improved allocation of resources and a reduced rate of the outcome of interest. This work explored methods of fine-tuning the positive outcome prediction (sensitivity) while maintaining satisfactory specificity.⁴

The primary objective was to combine the advantages of DTs and ANN to create a system that reduced the complexity of the problem and improved the sensitivity while maintaining specificity. This approach used DTs to select the features, allowing the ANN to focus on the important ones in an attempt to create a highly accurate prediction model. This integrated hybrid system was designed and tested on an adult intensive care unit database for the prediction of death or survival of a patient. Once a fundamental design had been implemented, the second objective was to apply the integrated hybrid approach to a large, complex dataset pertaining to a multi-factorial problem found in obstetrical care. This second objective is to improve on preliminary results of a pattern classification approach to predict premature birth in obstetrical care. Due to the complexity of both problems—an in-depth description can be found in Section 3.1.

1.3 Introduction of the Medical Environment

As mentioned, an integrated approach was designed and tested on one medical database (the adult intensive care unit) and then applied to another (obstetrical care). Therefore, there are two parts to this chapter. Both parts provide a deeper exploration of our

⁴**Sensitivity:** given that the actual value is positive, the fraction of value that is actually predicted as positive. **Specificity:** given that the actual value is negative, the fraction of value that is actually predicted as negative.

motivation, the risks and consequences of mortality in an adult ICU and preterm birth in the obstetrical environment, and what prediction systems are in place for these conditions. Though both problems pertain to the medical field, their environments are completely different: the first section (Section 1.4), deals with adults, requires quick predictions and fast medical interventions; in the second section (Section 1.5), interventions may not be as urgent and physicians have more time when making decisions.

1.4 The Intensive Care Unit (ICU)

The typical ICU admits patients from the emergency room, an operating room or from another part of the hospital [21]. Here, the medical staff constantly monitors the most critically ill of patients in the hospital. The medical staff experiences a high level of stress, as quick decisions have to be made for the life or death outcome of a patient [22]. The staff is faced with choosing interventions that result in a life or death outcome and constant ethical deliberations. Due to the quick-paced environment and critically ill condition of the patient - there is a higher rate of mortality in the ICU compared to other areas of the hospital [22]. There are many reasons attributed to a higher risk of mortality in the ICU: (1) risks of iatrogenic or inadequate therapy [23], (2) the severity of illness of the patient, underlying disease, organ function, age, and physiological conditions affecting the outcome of a patient [23].⁵

1.4.1 Cost of ICU

With the advent of new and sophisticated technology, more and more patients are able to survive. However, this technology is accompanied with a large cost related to usage, maintenance and replacement of the equipment. Additionally to survive in the ICU does not necessarily ensure a good quality of life. Money, as the sole measurement of “cost” is a difficult subject to acknowledge in the ICU because without these expensive measures,

⁵**Iatrogenic:** an inevitable or unforeseen condition caused by a treatment [24].

the patient will die or be left severely impaired [25]. “Cost” is a highly deliberated topic, as it concerns ethical decisions. To explain this, one should know that the primary goal of an ICU is to reduce short-term mortality. In this unit, the lives of patients are unstable and depend on the existence and proper use of the life-sustaining equipment [26]. The ethical decisions involve balancing the high financial cost with quality of life, life expectancy and long-term mortality expectations of a patient [27]. Though the number of ICU beds in the US accounts for <10% of the hospital’s bed usage, the cost is estimated to be a third of the in-patient cost. Intrigued by this high cost, Dasta *et al* (2005) looked into the daily cost of an ICU patient. They report that the first day in the ICU is about 4 times more expensive than a non-ICU day while the subsequent ICU days are about 2.5 times more expensive. These combined factors have placed pressure on the ICUs to become more efficient, limit practice variation and contain costs [5]. Prognostic models have been a widely used strategy for the purpose of assessing a patient’s outcome and providing risk management, benchmarking performance and further interventions [28, 29].

1.4.2 The Need for Prediction Models in an Adult ICU

Difficult decisions are made in the ICU and to constantly decide the life or death outcome of a patient is an emotionally exhausting task. It is ideal to have accurate prognosis systems available as the value of correct prediction outcome is of immediate value. To correctly identify patients as a non-survivor could mean earlier development and execution of a treatment plan. This can be in the form of a more aggressive intervention, more medical attention allocated to these patient, or if the case - assign palliative care instead of restorative care for terminally ill patients [27].⁶ The quality of care is not the only reason for morbidity and mortality, but as well as well the patient’s underlying health status, physiologic reserve, and admitting diagnosis [28]. It is a difficult topic to acknowledge, but due to the high cost of resource intensive technology in the ICU, the cost versus benefits should

⁶**Terminally Ill:** patients who will survive mortality in the ICU but will not survive long-term [27].

be considered [25]. To accomplish this, the medical staff needs a sound prognostic system to support their decision, especially in situations of high cost and low-health benefits [25].

1.5 The Importance of Delivering to Term

In 2008, over 360, 000 infants were born in Canada and over 4 million in the US. Pregnancy, though an exciting time for many expectant families, is often also a time for caution and utmost care. There is a demand for pregnancy health, as many seek prenatal care to ensure their pregnancy term progresses as planned. An ideal term is about 280 days (40 weeks) counted from the last menstrual period [30]. During this term, the developing fetus(es) is carried inside the mother's uterus, a natural environment for survival and growth. It is important for the fetus(es) to remain inside the uterus until completion of its term, as each day, cells, tissues and organs are undergoing development and maturation for survival in an external environment. Unfortunately, for a small percentage of pregnancies, the infant is delivered before completion of the term. Such events are classified as preterm births (PTBs) and are defined as the "birth of a baby before 37 weeks (259 days)" [31]. Thus, the implication of PTB can be divided into three groups [32]: Due to past discrep-

Subgroups	Gestational Age	Birthweight (grams)
Extremely PTB/LBW	24 weeks to 27 weeks 6 days	<1000
Very PTB/LBW	<31 weeks 6 days	<1500
PTB/LBW	<36 weeks 6 days	<2250

ancies of gestational age, some obstetric and pediatrics statistics also use low birthweight (LBW) as an indicator of the severity of preterm birth.⁷

1.5.1 Rate of Preterm Birth (PTB)

Over 10 years ago, a journal sought determinants of preterm birth rates in Canada due to the rising trend of preterm birth (6.3 % of live births in 1981 to 6.8% in 1994). The rise was attributed to changes in the frequency of births, obstetrical intervention, and

⁷**Gestational Age:** The period in which the fertilized egg develops into an infant ready to be delivered

in vitro fertilization methods (leading to multiple births) and the use of ultrasound-based estimates of gestational age [33].⁸ Since the time this report was published, numerous studies have been dedicated to improving our knowledge of preterm birth. Unfortunately, even with the advancement of prenatal care technology as well as the implementation of educational programs, the rate continues to rise. In Canada, PTB has increased by >19% to a rate of 8.1% (2006-2007) [34]. Also of great concern, the US rate has always been substantially higher than Canada (9.4% in 1981) and has increased to >36 % in 2006 [35]. Between 1996 and 2006, the rate of preterm births in the US has increased more than 16%, rising from 11.1% to 12.8 % of all live births

1.5.2 Health of Preterm Infants

The preterm (PT) neonate's first experience in an external environment is met with required technology and resources. Incubators, oxygen support and pharmaceuticals are essential for its survival. Depending on the severity of PTB, the magnitude of initial health complications can vary. Breathing problems, internal bleeding, poor blood circulation and temperature control, liver failure, seizures, and feeding problems are only a few of the problems to mention [36]. Though complications of late preterm are not expected to be as severe as early PTB, the incidence of immediate newborn morbidities is higher than with term infants [37]. It is only logical that the increased severity of PTB places the neonates at an ever higher risk of mortality and morbidity for internal bleeding, breathing problems and infections [38]. In addition to these conditions, severe disability is a common incident (over 49%) among children classified as early preterm birth (less than 25 weeks of gestational age) [39]. Such situations may put an emotional and financial strain on families.

⁸**In Vitro:** Fertilization of the egg outside of a living body and then implantation into the uterus.

1.5.3 Cost of PTB

In 2005, the average cost in the US in inpatient and outpatient care for a preterm infant in the first year was over \$32,325. This is considerably higher than the cost for a term infant (\$3,325). With about 542,893 PTB in the US, over 18 billion dollars are spent per year [40]. These values are the reported average cost; statistics of in-hospital cost of a Canadian hospital shows the magnitude of this problem.⁹ The cost of neonatal intensive care per survivor in 1998-1999 was a little over \$100,000 for a neonate born at 23 to <25 weeks, while a neonate born at 29-31 weeks is just over \$20,000 [41]. According to the March of Dimes, in 2005 the societal economic cost (medical, educational and loss of productivity) associated with preterm birth was at least \$26.2 billion (about \$51,600 for every infant born preterm) [40]. According to Gilbert (2006), initial costs of preterm newborns are even higher, with over \$200,000 (US) allocated to essential equipment and expertise needed for hospitalization. Additional economic costs associated with premature birth include health care for chronic diseases and resources needed to support and develop neurological and developmental disabilities which may result from premature birth.

1.6 Thesis Objectives

1.6.1 First Objective

The first objective was to create an improved classifier system used to optimize sensitivity and specificity of predictions. This work investigated the advantages of combining DTs and ANNs to improve the sensitivity while maintaining high specificity. The new integrated approach used DTs to select an optimal feature set allowing the ANN to focus and refine the prediction results. Past research work by the Medical Information-technology Research Group (MIRG) found weight-elimination ANNs to be a powerful tool in im-

⁹IWK Health Center in Halifax, Nova Scotia

proving the accuracy of predictions [42]. The DT-ANN hybrid and weight-elimination approaches were used for the creation of a classifier to predict mortality in an adult intensive care unit. Both approaches were tested on a small homogeneous dataset (mortality in an adult intensive care environment: all are post-operative patients) and in a small heterogeneous environment (mortality in an adult medical intensive care, where the type of patients is very diverse). The results of the integrated DT-ANN hybrid system and weight-elimination ANN were to be compared using sensitivity, specificity, and the Area Under the Curve (AUC) as performance indices.

1.6.2 Second Objective

The second objective of this thesis work was to improve preliminary results for predicting high-risk preterm birth as established by Christina Catley [15]. Preliminary results established by Catley used a classification based ANN and a hybrid integrated screening system using risk-stratification ANN, processing the test data through three passes: one to maximize sensitivity, one to maximize specificity and the third to classify ambiguous cases using a decision-tree voting algorithm. This 48-variable model was developed and tested on a US database for women who had previous children (Parous).¹⁰ This resulted in a sensitivity of 64% and specificity of 84% [15]. Once a fundamental DT-ANN approach was established and tested on small database, to complete the first objective - the integrated hybrid approach was extended onto this large, complex, heterogeneous dataset pertaining to a multi-factorial problem - the prediction of high-risk preterm birth in obstetrical care. This work attempts to improve the prediction accuracy of past work by (a) reducing the complexity of the classifier by minimizing the number of input variables, (b) maintaining or exceeding current sensitivity, specificity and AUC, (c) using input variables that are not US specific while maintaining or improving the accuracy of the estimations, (d) obtaining preliminary results for women who have not had previous children (Nulliparous), creating

¹⁰The US database was the Pregnancy Risk Assessment Monitoring System (PRAMS)

a system that can be extended to a general population. The sensitivity, specificity and the area under the curve (AUC) were selected as the performance indices.

1.7 Thesis Outline

Chapter 1 provides the motivation, objective, problem statement and contributions of this thesis work. It establishes the medical environment for the problem to be solved: the adult intensive care unit and obstetrical care.

Chapter 2 provides a detailed review of current literature. The first part of the literature review acknowledges the accuracy of standard clinical prediction methods used in the intensive care unit and in obstetrical care. The second part of the literature review explains past methods used to improve medical data mining results.

Chapter 3 explains the methodology used to create the models.

Chapter 4 presents the results and discussion for predicting ICU mortality and preterm birth.

Chapter 5 presents the concluding remarks, contributing knowledge and future work discovered from this thesis.

Literature Review: Outcome Estimation Systems and Technology

2.1 Prognostic Models in the Adult ICU

Assessment of the severity of illness of a patient and quality of care given by the medical staff have been designed in the form of prognostic models. These have been created to identify patients who are at a higher risk of morbidity and death so that the appropriate intervention can be applied [10]. Scoring systems have been used for quality control, benchmarking, risk stratification, and patient classification [28]. The three most popular prognostic models are the Acute Physiology and Chronic Health Evaluation (APACHE), Simplified Acute Physiology Score (SAPS) and the Mortality Probability Models (MPM) [28, 4]. Each of these models has gone through changes for the purpose of improving the performance and can be divided into four generations, as shown in Table 2.1.

2.2 Critical Analysis for the Intensive Care Unit

2.2.1 The Accuracy of Standard Scoring Systems for an Adult ICU

Since there are many prognostic model used, the focus is on the APACHE II, APACHE III, SAPS II, MPM II for older generations (since they have been well-tested and clinically used) and the SAPS III, MPM III and APACHE IV for the 4th generation model [29]. Each ICU prognostic model generates a probability of mortality. Since logistic regression is used to determine the probability, many studies validate the prognostic models using the Hosmer-Lemeshow statistic for the ‘goodness of fit’ and the AUC for discrimination effectiveness [4, 28].¹¹

¹¹**Goodness of Fit:** assesses how well the probability outputted by the model represents a true probability of mortality based on the samples. A value >0.05 indicates that the model is able to approximate the probabilities for a mortality outcome [43]. See Appendix H.5 for more information.

Table 2.1: The four generations of the severity prognostic models in the Adult Intensive Care Units; APACHE, Acute Physiology and Chronic Health Evaluation; SAPS, Simplified Acute Physiology Score; MPM, Mortality Probability Model. (A) generation of the ICU prognostic model; (B) year that the model was published; (C) name of the system; (D) number of input variables; (E) time that the data is collected [29, 28]

A	B	C	D	E
First	1981	APACHE-I	34	
	1983	SAPS-I	13	<24hrs
Second	1985	APACHE-II	12	<24hrs
	1985	MPM-I ₀	7	upon admission
	1985	MPM-I ₂₄	7	in 24hrs
	1991	APACHE-III	17	<24hrs
	1993	MPM-II ₀	15	upon admission
Third		MPM-II ₂₄	13	within 24 hrs
	1993	SAPS-II	17	<24hrs
	1994	MPM-II ₄₈		24-48hrs
		MPM-II ₇₂		48-72hrs
	1998	APACHE-III-I		
	2001	APACHE-III-J		
Fourth	2005	SAPS-III	20	≤1hr
	2006	APACHE-IV	142	
	2007	MPM-III ₀	16	≤1hr

A 1998 cohort study on 1393 patient reported the SAPS II accuracy with an AUC of 0.80 and a poor goodness of fit ($P < 0.001$) [44]. In a later study (2000), the researchers reported a sensitivity of 39.4% and a specificity of 95.6% for SAPS II and the APACHE II with a sensitivity and specificity of 31.6% and 97.2%. The AUC for SAPS II was 0.816 and the APACHE II 0.805 [45]. A 1995 study by Castella compared the AUC of the APACHE II, APACHE III, MPM₀-II, MPM₂₄-II, and SAPS-II: as 0.863, 0.848, 0.805, 0.833, 0.847, respectively. Though these discrimination values were good, all P-values were unacceptable except for SAPS II ($P = 0.1019$) [46]. Among the different models, the SAPS II reported the highest discrimination value and P-value, however, subsequent studies reported that the model must be customized for good calibration [29].¹² Validation on an external database

¹²**Customization** involves improvement of models require modifying logistic equation coefficient or co-efficient of all variables in the model to improve calibration to the ICU databases. This is to due to changing

showed acceptable discrimination capacity: APACHE II of 0.818 ± 0.005 , SAPS II of 0.831 ± 0.005 , MPM-II₀ of $.796 \pm 0.005$ and MPM-II₂₄ of 0.822 ± 0.005 [49]. There was a poor “goodness of fit” which was attributed to the sample size and grouping value. This study also supported a slight superiority to the SAPS II model while the MPM-II₀ had the poorest performance. Unfortunately, the sensitivity and specificity were not available. It is unfortunate that many research studies have failed to recognize the need to report sensitivity and specificity, as this provides the portion of correctly predicted positive and negative cases.¹³

Due to changes in ICU case-mix and resulting deterioration of results of older models, recent model application utilized customization (creation of new equation, changing weights of constituent variables, or the addition of new predictor variables) [52, 29]. Changes in the ICU case-mix had also given rise to completely new models such as the MPM-III, SAPS-III and APACHE IV. Tests have determined acceptable discrimination for all the APACHE IV models, although the calibration of the model varies. In a 2006 study, a comparison of the APACHE-III model found that the APACHE IV had an AUC of 0.880 and calibrated well with a $P=0.08$ [53]. In a later study (2008), the discrimination of the APACHE IV (0.892) was found to be superior to the MPM-III₀ (0.809) model and SAPS II model (0.873); in contrast to the 2006 study, only the MPM-III₀ and SAPS II were found to be well calibrated [54]. SAPS III was found to have an AUC of 0.848 and $P=0.16$, also making it an acceptable model [55]. While these models have shown to have acceptable discrimination and some are well calibrated; this generation has yet to be validated on an external independent set [29]. In addition, though the APACHE IV was found to discrimi-

baseline characteristics of the population (age, diagnosis) since the scoring systems were first developed [47, 48].

¹³The majority of studies in this area only report the AUC and forego the sensitivity and specificity. Such reports are misleading. A high AUC indicates acceptable discrimination (Section 3.5.1); however, due to different curve shapes (Section 3.5.9) that may arise when comparing two different tests - the sensitivity and specificity values can vary widely [50, 51].

nate well - the number of variables needed to be collected (142) is a huge downside to this model.

Another area examined is the comparison between the model's prognosis and the physician's. In a systematic review, 12 studies compared the APACHE II, SAPS II and MPM systems with the physician's diagnosis. The best report for the prognosis system was Kruse *et al* with a sensitivity of 67.0% and specificity of 87.0% (n: 366); however, this was a computer model of the APACHE II (not the original system) and organ failure variable. All other prognosis systems reported a sensitivity range of 43.0% to 58.0% and specificity range of 75.0% to 99.0%. The physician's estimation review reported a sensitivity range of 37% to 82% and 69.0% to 98.0% for specificity. It was determined through a series of statistical tests that the physician's prognosis was actually more accurate than the scoring system; however both systems have a moderate ability to discriminate between survival and non-survival [8]. The 4th generation has yet to be tested. Caution is advised when using scoring systems to predict mortality - these systems should be tested among different ICU environments on a global scale [53].

2.2.2 Limitations

The first obvious limitation is the poor prediction accuracy of the current prognosis systems. Although ICU prognostic models have been in place for over 25 years, the ability to predict survival and non-survival is poor and reported results vary widely (due to different patient diagnostic, timing, changes in the effectiveness of therapy, quality of care) [8, 28]. Most of the systems have less than a 60% sensitivity, or if there was a high sensitivity, it was followed with a decreased specificity. Additionally, these reviews were on a patient population of <600 patients [8]. Other limitations are the high cost and extra time needed for data collection. In one study, the MPM-II₀ had a mean time of 11.1 minutes, the SAPS II was 19.6 minutes and the APACHE IV was 37.3 minutes [54] and was labelled as

“a burden”. This collection time may be feasible for hospitals involved in research and development of efficient systems; but for independent ICUs, collection of variables may not be part of the clinical routine. Time is essence in an ICU and in many units; the medical staff are already working in a high pace, under-staffed, excessive workload, environment [22]. This stressful environment may affect the hospital’s outlook on data collection and may cause errors in the reported data, resulting in a miscalculated prognosis. The common goal among the studies was to strive for a better system for the prediction of death or survival in an ICU [28, 8]. The creation of newer generation models (many of which are derived from older models), also indicates the need for a superior prognosis system in the ICU.

2.3 Types of Preterm Birth

Obstetrical interventions, applied when the pregnancy may be harmful to the mother or infant, have contributed to the rise of PTB. Medically indicated births are performed by caesarean section or induction to reduce adverse risk to both parties.¹⁴ It is reported that about 20% of preterm births are indicated. Situations where PTB is indicated include intrauterine growth restriction, preeclampsia, placenta previa, and non-reassuring fetal testing [30, 56]. Other types of preterm birth are spontaneous (unexplained) preterm labour, intra-amniotic infection and preterm pre-labour rupture of the membranes (PROM) [30]. Pregnancies are divided into low-risk and high-risk cases.¹⁵ High-risk patients are then divided into symptomatic and asymptomatic cases. Symptomatic women have intact membranes but are experiencing symptoms of preterm labour (PTL) - contractions, pelvic heaviness and low back pain - well before complete pregnancy gestation. They are usually admitted and assessed in emergency obstetric care. On the other hand, asymptomatic women do not

¹⁴**Indicated births:** pregnancies marked for artificial births; **Caesarean section:** a surgical operation for delivering an infant through the abdominal wall; **Induction:** labour is started using artificial means such as drugs that stimulate contraction or surgical rupture of the membrane

¹⁵**Low-risk:** mothers who do not have medical problems or symptoms that may cause obstetrical problems; **High-risk:** Mothers who have medical problems or symptoms that may cause obstetrical problems

experience symptoms of PTL but their lifestyle or pregnancy history suggests an increased risk for PTB.

2.4 Factors Contributing to Preterm Birth

PTB is a multi-factorial problem resulting from different causes: medical and mechanical reasons (multiple gestation, placental abruption) and social and demographic factors (stress due to poverty and violence) [11, 30]. PTB remains a difficult problem because though many predictor variables have been found to be highly associated with PTB, their use in predicting PTB has had limited success. The next section presents a brief summary of some of these factors.

Demographic The most popular demographic variables examined in studies are age and race [57]. Age is a contributing risk factor for LBW and PTBs. Teenagers are more likely to give birth under these conditions [58]. Due to the higher risk of pre-existing and the development of adverse health conditions, women over 40 are also more likely for adverse birth outcomes. Many North American models present race as a participating factor in an increased incidence of preterm birth. Higher incidents are reported for African Americans compared to other races.

Behavioral Characteristics Women who continued drinking but stopped prior to their second trimester showed an increased rate of PTB [59]. Although the study showed a low association between PTB and low intake of alcohol, there was a small increased risk associated with binge drinking. Likewise, women who smoke expose themselves and their developing infant to an abundance of toxic chemicals. These chemicals are strongly correlated to morbidity and diseases [60].

Stress A study released by the March of Dimes found that the average cost of having a baby in the United States in 2004 was more than \$8,000 (this includes insurance paid by large employer private plans). Though the majority of mothers were covered under

insurance, about 40% relied on publicly financed programs [40, 61]. Insufficient funds, poverty and low socioeconomic status are already stressful situations which can escalate for an expectant family. Furthermore, divorce, death, drugs or alcohol problems, emotional and physical abuse, partner violence, and experiences of racism, are all strong stress factors that may contribute to PTB [11, 30, 32].

Maternal Health Both an increased and decreased maternal BMI has a negative impact on neonatal outcome. Increased BMI was associated with adverse birth outcomes such as hypertensive complications and caesarean section, while decreased BMI was associated with preterm birth, lower birthweight and increased mortality on the neonatal ward [62]. Inadequate pregnancy weight gain and pregnancy intervals of less than 18 months increased the risk factor. Both pre-existing hypertension and developed hypertension are conditions that are monitored closely during pregnancy. Pre-eclampsia is hypertension developed after 20 weeks of gestation or in the early postpartum period and returns to normal within 3 months postpartum [30] and is widely associated with primiparity.¹⁶ The risk is 4.1% for the first pregnancy and increases for the subsequent pregnancy given that previous history of pre-eclampsia existed. For example, the risk was 31.9% for the third pregnancy given that they had already developed preeclampsia in their previous two pregnancies [63]. Diabetes or similar medical conditions such as glucose intolerance, hypoglycemia and gestational diabetes mellitus (GDM) have been associated with adverse perinatal outcomes. Hedderston *et al* concluded the presence of these conditions (isolated from other adverse condition such as pre-eclampsia) have an increased risk of spontaneous PTB (sPTB). It is difficult to completely isolate the data from all known risk factor such as pre-pregnancy weight and parity [64]. Increased rates of sPTB have been reported to range from 24% to 62%. In a later study, Melamed *et al* (2008) found that pregnancies complicated by pre-gestational GDM had a higher risk of sPTB and indicated PTB. The majority of PTB were

¹⁶**Postpartum:** the period immediately following the birth of the baby; **Primiparity:** first pregnancy.

determined to be indicated PTB [65]. The presence of vaginal bleeding is a concern during pregnancy. The severity and length of bleeding affects the risk of PTB. Minor bleeding has a virtually weak association with PTB, but bleeding in both trimesters, heavy bleeding, recurrent bleeding and bleeding over multiple days is greatly associated with PTB [66]. Mechanical problems arise when there are cervical anomalies (surgery, trauma or congenital).¹⁷ Used to reduce the possibility of preterm cervical dilation and rupture of the membrane, a cervical cerclage is performed.¹⁸

Pregnancy history Women who have had previous preterm births are at a greater risk of preterm delivery. Those with prior sPTB have are at a risk of 15% (95% CI: 5.0-38.0) to 3.0% (95% CI: 3.0-3.0)[68]. For recurrent preterm birth, there is a higher risk of delivering the next infant as preterm.¹⁹ The risk is inversely related to the gestational age of prior preterm births; the lower the gestational age of prior PTBs, the higher the risk of PTB for the present pregnancy.

Current Fetal Conditions The presence of multiple fetuses inside the uterus leads to less room for fetal development and growth. The increased hormonal and mechanical effects of carrying multiple fetuses could lead to pregnancy related complications. The average gestational age for multiple birth decreases with an increased number: twins, triplets and quadruplet is 35 weeks, 33 weeks and 29 weeks respectively. It is common for birthweights to be decreased as well [69].

2.4.1 Current Knowledge

The non-linearity, heterogeneous, complexity nature of the problem is evident. In addition to the many different variables that have been found to increase the risk of PTB; it has also been well established that different ranges variables values also contributed

¹⁷The cervix is a firm, cylindrical, muscular structure which forms the canal length between the uterus and the vagina. It is used to retain the fetus and also allows for its passage and delivery.

¹⁸This procedure involves inserting a suture around the cervix (transabdominal or transvaginal) in order to prevent preterm delivery [67].

¹⁹**Recurrent PTB** is defined as having two or more previous birth delivered before 37 weeks [68]

to an increase or decrease of PTB risk. The ability to successfully identify and predict PTB can be beneficial as it could define a population for higher risk and provide a better understanding of the pathways leading to PTB [70]. Also, the physician can initiate an earlier intervention, treatment, and initiate proper allocation of necessary resources. This may then lead to a decreased rate of PTB. Though past studies have discovered valuable information, the successful prediction of PTB still remains a continuing problem that has yet to be solved.

2.5 Critical Analysis for Preterm Birth

2.5.1 Screening Tests for Preterm Birth

To date, cervical length (CL) measurement and the fetal fibronectin (fFN) test have been advocated as the gold standard for the prediction of PTB. CL screening measures the changes in the cervical length using transvaginal sonography (TVS). A shortened cervix indicates a higher risk of spontaneous preterm birth. The most useful time for CL measurement is between 16 and 32 weeks of gestation. After 32 weeks, the cervix naturally shortens for women who deliver to term [71]. Fetal fibronectin is a protein found between the maternal-fetal interface and is normally found in low levels in cervico-vaginal secretions [72]. A swab of vaginal secretion is taken and tested for its level of fibronectin. The presence of this protein is normally seen in vaginal secretions up to 22 weeks of pregnancy, and not again until one to three weeks before delivery. Levels ≥ 50 ng/mL at/or after 22 weeks are marked as abnormal and associated with an increased risk of spontaneous PTB. A test taken at less than 22 weeks of gestation is not useful for predicting premature birth [73] and after 34 weeks, as fFN levels normally increase after this time. Therefore this test is only applicable between the gestational interval of 22 to 32 weeks. For the last 20 years, numerous ongoing studies have been dedicated to determining the prediction accuracy of these tests.

2.5.2 Heterogeneity

PTB is a multi-factorial problem that affects many sub-populations of pregnant women. Many women who have been prescribed a CL or fFN test have already shown one or more identifiable risk factors and/or are experiencing symptoms of PTL. It is difficult to compare results of the prediction accuracy of fFN and CL due to the heterogeneous nature of studies conducted. Such studies have been focused on a wide range of obstetrical population: asymptomatic patients in low and high-risk populations of singleton births, multiple gestation and symptomatic women with either preterm labor or preterm premature rupture of the membranes [74, 71]. Also, the cut-off gestational ages for PTB differ among studies as some report preterm at <34 [75], <35 [76, 12] to <37 weeks [68]. Other studies even provide cut-off values as low as < 32 weeks [77, 72]. For CL screenings, other factors include timing of the test and CL cut-off values assigned have a different associated risk for preterm birth [78, 79, 74, 77]. Such heterogeneity makes it difficult to make a direct comparison between prediction accuracies of various studies.

2.5.3 Accuracy Prediction of Cervical Length (CL)

Identifying the risk for PTB is not simple. The main factor to consider is the gestation age at which CL is used as it affects the predictive capability for PTB. Between 10–14 weeks of gestation, the CL measured for women who delivered prematurely showed no significant difference from those who delivered at term. It was later in the pregnancy, between 20-24 weeks of gestation that the CL measurements were significantly shorter for women who delivered prematurely (28.4 mm for PTB $\leq$ 35 weeks) compared to those who delivered to term (37.8 mm) [78]. As mentioned earlier in section 2.5.1, the clinical standard practices measure the CL between 16-32 weeks of gestation; therefore concluding from this study at 16 weeks, diagnosis should be taken with caution. The mean CL at a certain gestational age is greater for those who will deliver closer to term. Patients who delivered

between 32 and 36 weeks of gestation had a significantly greater mean CL ($36. \pm 7.4$ mm) compared to those who delivered at 32 weeks (32.9 ± 9.9 mm) [80]. As indicated by Figure 2.1, the probability of PTB decreases with a greater CL value; yet this probability is also a factor of gestational age at the time of CL measurement [74, 80]. Hassen (2000) concluded that the prediction accuracy for PTB ≤ 32 weeks with a CL of ≤ 15 mm had a PPV of 47.6%, a NPV of 96.7%, sensitivity of 8.2%, and specificity of 99.7% ($n=6877$). The sensitivity is incredibly low, and only increased to 14.7% for CL measurements ≤ 25 mm [80]. Another study reports the accuracy prediction of preterm delivery (cut-off times of ≤ 32 weeks, ≤ 35 weeks, ≤ 37 weeks) with a sensitivity between 13–44%, specificity between 90–99%, PPV between 15–47% and NPV between 80–98% ($n=760$) [77]. The wide range is attributed to the different gestational age and CL thresholds. As noted, the sensitivity range does not exceed 50%, indicating that preterm births were not successfully identified. In a 2005 study, a sensitivity of 46%, specificity of 94%, PPV of 67% and NPV of 87% was discovered ($n=384$ and prevalence = 21.4%) [79]. The risk for preterm birth is inversely proportional to the CL, however results vary widely. One study reports a CL of ≤ 25 mm has a prediction sensitivity of 47% and PPV of 37% for birth before 35 weeks [81]. Another study found different results: a length of ≤ 25 mm had a sensitivity of 7%, a specificity of 100% and PPV of 15% for preterm delivery before 35 weeks. In this case the sensitivity is too low to be credible. For delivery < 37 weeks, the sensitivity decreased to 6% (95CI: 2,13), specificity 100% and PPV of 39% (95CI: 14,68) [79]. With an extremely low sensitivity is it questionable as to why this method is a clinical standard. Attention is focused on the very high specificity rate for the prediction of PTB (reported between 90% to 100%) and to the low sensitivity rate (which ranges from 0% to 50%). The predictive power lies in the negative test, indicating which cases will not result in preterm birth; whereas a low sensitivity indicates a high false negative rate and preterm births are not well predicted. When the CL prediction is marked as positive for PTB, additional prenatal testing is usually prescribed [71]. Though this method is a clinical standard, the literature

review suggests that stronger methods of prediction are needed for preterm births.

Figure 2.1: Predicted probability of delivery before week 35, by cervical length (mm) and time of measurement (week of pregnancy). Figure is from [74].

2.5.4 Accuracy Prediction fetal fibronectin (fFN)

fFN is deemed to be one of the most effective markers for preterm delivery with higher predictive power in symptomatic women. Clinical guidelines deter the use of fFN for asymptomatic patients and reassurance that preterm labour will not occur. Therefore many studies focus on symptomatic or high-risk populations when using this test. It has been noted that sensitivity rates are higher with symptomatic women than asymptomatic women while specificity rates are slightly lower [82]. Taking recent fFN studies that were comparable, a visual table was created (Table 2.2) of the fFN prediction accuracy. In order to report consistent results, only studies that defined preterm birth as births ≤ 37 weeks were reported and a non-specialized sub-population (i.e. only births with multifetal reduction). Many studies for fFN tests were performed on symptomatic patients and for those who are asymptomatic, studies were performed on high-risk cases. The five studies in-

cluded are as follows: (1) 2009 - Fetal fibronectin as a predictor of spontaneous preterm labour in asymptomatic women with a cervical cerclage [83], (2) 2007 - Performance of cervicovaginal fetal fibronectin in a community hospital setting [84], (3) 2004 - Prospective randomized controlled trial of fetal fibronectin on preterm labor management in a tertiary care center [85], (4) 1995 Fetal fibronectin improves the accuracy of diagnosis of preterm labor [86], (5) 2000 - Fetal fibronectin detection as a predictor of preterm birth in actual clinical practice [87].

The visual table shows the prominent heterogeneity between studies conducted: the inclusion and exclusion criteria, population sample and gestational age during the testing period are a few differences to mention. It is interesting to note that Study 1, used to assess the accuracy prediction between women with cerclage and no cerclage, had the highest number of cases (n=751) with the lowest reported positive cases (11.0%). This study showed the lowest sensitivity value of 26.8% [95CI:20–34.5] and the highest specificity value 92.8% [95CI: 90.4–94.7]. The wide range of accuracy of results shown in Table 2.2 for the prediction of preterm birth is affected by different variables between studies. This emphasizes Leitich and Kaider’s statement in their 2003 meta-analysis, “due to the heterogeneous results in different studies, caution should be taken when they are applied to a specific population”. In their study, the best prediction results for fFN had a sensitivity of 52% [95CI:42 –63] and a specificity of 85% (95CI: 82–87) [82]. Sensitivity and specificity results from a more recent study, Study no.2 (shown in Table 2.2), fall within these confidence interval: 45.2 [C95 I:34.3–55.3] and 78.8 [95CI:71.8–85.2] respectively [84]. These results indicate that not much improvement has been made to the prognostic capability of fFN. The predictive power of fFN lies within the high specificity and negative predictive of this test. A positive fFN test does not necessarily indicate preterm birth as most of these studies report a low sensitivity <50%, indicating that the majority of PTB cases are not identified. At a low sensitivity for a cost of \$100 Canadian, plus additional

Table 2.2: Accuracy of fFN studies for the prediction of PTB. (A) Purpose of Study; (B) Inclusion Criteria; (C) Exclusion Criteria; (D) Gestational Age; (E) Cervical Dilation; (F) Symptomatic; (G) Total Number of Cases; (H) Number of Positive cases; (I) Percentage of PTB; Sens = sensitivity; Spec = specificity.

	1	2	3	4	5
A	Accuracy of labour in asymptomatic high-risk women with a cervical cerclage ^a	Prediction of fFN in clinical setting	Prediction capability of rapid fFN on the length of hospital stay and the use of preterm labor	Determine whether fFN expression is more useful for predicting preterm delivery in clinical practice than it has appeared to be in prospective blinded studies	Assess the utility of cervicovaginal expression of fFN in the diagnosis of preterm labor
B	Singleton pregnancy with history of previous PTB or preterm prelabour rupture of membranes (PPROM) <37 weeks' gestation, a previous late miscarriage (16 ⁺⁰ – 23 ⁺⁶ weeks' gestation), uterine anomalies or previous cervical surgery	no major fetal anomaly or placenta previa	maternal age is ≥ 16 years of age	a singleton pregnancy, intact membrane	
C	History of iatrogenic PTB	multifetal gestations, cerclage, preterm PROM, vaginal bleeding			
D	23 ⁺⁰ and 27 ⁺⁶ weeks	<36 weeks	23 and 34 weeks	24 to 35 weeks	24 and 34 weeks
E	<2 cm	<2 cm	≤ 3 cm (primigravid) ≤ 4 cm (multiparous)	≤ 3 cm	≤ 3 cm
F	No	Yes	Yes	Yes	Yes
G	751	230	46	85	192
H	84	33	11	20	45
I	11.19	14.3	23.9	23.5	23.4
Sens	26.8 (20-34.5)	45.2 (34.3-55.3)	38.5 (13.9-68.5)	35	44
Spec	92.8 (90.4-94.7)	78.8 (71.8-85.2)	81.8 (64.5-93.0)	92	86
PPV	48.8	57.6 (43.6-70.3)	45.5 (16.8-76.6)	85	60
NPV	83.2	69.3 (63.2-75.0)	77.1 (59.9-89.6)	52	76

^aOnly used results for no cerclage

laboratory requirements [88], the need for a better prognostic system is evident.

2.5.5 Risk Factor Scoring

There are several modifiable and non-modifiable risk factors associated with preterm birth. Risk factors such as smoking, drinking and drug use can be reduced with lifestyle changes; while other factors such as race, parity, sex and maternal age cannot be modified. Over the years, risk scoring systems have been developed, not only to predict PTB, but also to identify factors that are highly associated with PTB [89]. A higher score indicates more risks are present during the pregnancy. A 2004 review of risk scoring systems developed over the years showed a need for a better prediction system. It has been almost 30 years since the first reported scoring system included in the review conducted in 2004 (by Kaminski in 1973); yet the poor prediction accuracy remains as a challenge. The only scoring system tested on sub-populations (black women, nulliparous, multiparous, 1st trimester, 2nd trimester) of obstetrical women is the Creasy scoring system. Even with sub-population testing, the specificity is high (remaining around 80-95%), the sensitivity does not exceed past 50% [90]. Other scoring systems report an even lower sensitivity, <30% [89]. Therefore, the inability for the scoring systems to correctly classify the preterm outcomes - which results in a low sensitivity - has deterred researchers from using this method [30]. Poor prediction accuracy of these systems may be attributed to different association values between the factor and outcome. For example, in a risk assessment system study, a low body mass index (<19.8) and an increasing Bishops score were found to be significantly associated with spontaneous PTB [89]. The scoring systems are important because they highlight fundamental risk factors for PTB, but obviously requiring more complex pattern recognition systems that scoring systems have not been able to offer.

2.6 Current Prediction Models for Preterm Birth

2.6.1 Preterm Birth Prediction

A project in 1992–1993 created a model for preterm birth prediction (≤ 37 weeks) using Learning from Examples using Rough Sets (LERS). A holdout method was applied to two large prenatal databases - half of the data was used to train the model and the other half to test.²⁰ The first dataset produced 350 rules and the second produced 170 rules using biophysical and symptom variables. There were a high number of variables used to create the model - 51 for Dataset 1 and 77 for Dataset 2. The reported accuracy of this Dataset 1 was 88.83%, FP of 6.34%, FN of 7.09% and 0.44% were unclassified. Dataset 2 had an accuracy of 59.28%, FP of 12.15%, FN of 25.29% and 3.28% were unclassified.²¹ Results may be biased as the cases were not from a broad, general population - the study mentioned that the majority of the 18,890 subjects were Caucasian, married and well-educated women in their late 20s. At the same time, the data was collected from home uterine monitoring vendors where patients had already experienced PTB or were at high-risk. It is difficult to calculate the sensitivity and specificity since its true category (fullterm or preterm births) was not given.

2.6.2 Uterine Electromyography ANN prediction

This method combines the use of biomedical instrumentation and a pattern classification approach. Studies using uterine electromyography (uEMG) have shown promising results for monitoring patients during pregnancy.²² ANN, in conjunction with uEMG has been shown to be an effective method of classifying preterm and term birth. The uEMG values were calculated for a number of uEMG parameters and these values were used as

²⁰Refer to Appendix H.1 for more information

²¹**FP (False Negative):** positive cases predicted as negative; **FP (False Positive):** negative cases predicted as positive

²²**uEMG:** Uterine contractions produce electrical activity that can be monitored using electromyography. The method consists of placing electrodes on the abdominal surface to monitor uterine contractions.

inputs to an ANN. The patient cases were divided into women who delivered to term [term laboring (TL): $n = 75$; term non-laboring (TN): $n = 59$] and women who delivered preterm [preterm laboring (PTL): $n = 13$; preterm non-laboring (PTN): $n = 38$]. Labouring and non-labouring refer to clinical diagnosis of labour by a physician. This method of prediction resulted in high classification rates of TL (79.0%), PTL (92.0%), TN (86.0%) and PTN (71%) [91].²³ Although this study presents promising results, the sample size was small ($n: 192$) and the sensitivity and specificity were not reported. This makes it impossible to determine how many preterm and term cases were correctly identified.

2.6.3 Early Prediction of Preterm Birth

Multivariate logistic regression was used to create a model for the prediction of preterm birth (defined as ≤ 32 weeks). This restrictive outcome prediction was used because neonatal mortality and morbidity is higher in 20-31 weeks of gestation than for later preterm birth. Using the United States birth registration data ($n: 2,988,844$), Tan *et al* used a multivariate logistic regression model to predict preterm birth for singleton, twins and triplets. The sensitivity, specificity, PPV, and NPV for singletons were 24.58%, 93.54%, 5.91%, and 98.69%, respectively; for twins 64.66%, 57.04%, 16.29% and 92.59%, respectively; and for triplets were 63.57%, 53.58%, 42.96%, and 72.78%, respectively [92]. Though the specificity was high for singleton births (93.54%), the sensitivity and PPV results were very low. As for twins and triplets, the sensitivity increased, but the specificity was too low to be clinically useful.

2.6.4 Limitations to Current PTB Prediction Methods

The current standard clinical prediction modalities (fFN and CL measurement) can be costly for follow-up tests, pertain to high-risk asymptomatic or symptomatic patients, and are effective in the middle of the gestational period. In addition, these methods are

²³**Classification rate:** the number of correctly predicted scores by the classifier.

invasive which may pose additional stressors at a very emotional time. At the same time, these methods have a large range of sensitivity values, all of which are under the 50% value. This means that less than 50% of the preterm birth cases are correctly identified. The prediction model using LERS was biased as the majority of the 18,890 subjects were Caucasian, married and well-educated women in their late 20s. For the uEMG study, the classification rate was reported but the sensitivity remained unknown, and sample size was limited. For the use of multivariate regression, the sensitivity (24.58%) and PPV (5.91%) were too low to be clinically useful for singleton pregnancies; while when the sensitivity increased (for twins and triplets), it resulted in a decreased specificity. It is evident that this field lacks sufficient prediction capacity needed to fulfill the screening protocol criteria proposed by Iams (2003): *efficient, high sensitivity, high negative predictive value, and low cost* [93]. As emphasized, PTB is a multi-factorial problem. With the low sensitivity of current methods, it is evident that a complex relationship exists between the predictor variables being used and the outcome of interest. This thesis will explore the application of non-linear modeling techniques, such as ANN and decision trees, to estimate this complex problem.

2.7 Pattern Classification Methods

2.7.1 Hybrid Artificial Neural Network

The idea behind hybrid methods is to combine the advantages of different pattern classification approaches while limiting its shortcomings. An ANN is a “black box” classifier, with its knowledge found within the weights, biases and connection links [94]. By combining an ANN with a knowledge based network - the generalization capability of the ANN may improve [17]. An ANN would benefit with a pattern classification approach that extracts crude domain knowledge from the dataset such as a DT. The following section explains methods of how decision trees and neural networks have been combined.

2.7.2 Tree-based Neural Network (TBNN)

A method of directly mapping the DT into a three-layer network was proposed by Ivanova and Kubat [95]. The algorithm proposed involves creating a DT to extract knowledge that defines the architecture of the ANN, feature selection and initialization of the weights. A decision tree is created, translated into an ANN and trained using a backpropagation algorithm. This translation from a DT to an ANN requires that the feature intervals created by the DT splitting value to be transformed into individual input variables for the ANN. Boolean variables are created giving the input variable a value of 1 if it falls within the feature interval. This network contains an interval (input layer) layer, AND-layer and OR-layer. The AND-layer models the conjunctions providing negated weights if the value does not fall within the interval. It corresponds to a path from the root of the tree to the leaf. The OR-layer is the output neuron corresponding to the number of classes [95]. Translating of a DT to an ANN follows the procedure: (i) Build a DT using a subset of the training data, (ii) Create interval layer, (iii) Provide additional links to the regular links to ensure that the net was fully forward-connected, and (iv) Train the network using back-propagation [96]. When developing an ANN model, an ANN is not able to reduce redundant, noisy data and focus on significant features before a classifier is created [96]; thus by designing the network to contain the DT interval layer, the ANN can focus on features that contribute the most to the outcome prediction and reduce computational cost [96, 97]. In addition to reducing computational costs, this method was observed to improve the classification accuracy of sleep stage scoring for babies. The TBNN improved a MLP classification accuracy from 60.0 ± 2.8 to 85.5 ± 3.6 .

2.7.3 Pruning-based Neural Networks (PBNN)

Similar to TBNN, the PBNN generates attribute intervals and decision rules from the decision tree given in the following steps: (i) Attribute intervals are derived from deci-

sion tree. Decision rules are derived from decision trees and attribute intervals, (ii) Decision rules and attribute intervals generate binary training patterns. The difference between TBNN and PBNN lies in the method of generating an ANN; variables and rules are not directly mapped into the ANN, as with previous methods. Binary training patterns are generated from the attribute intervals and decision rules and the ANN is trained using binary patterns. The input variables are trained on a one hidden layer MLP where the number of nodes is arbitrary. The ANN is then pruned using the Neural Network Pruning with Penalty Function (N2P2F) algorithm to obtain optimal topology. The PBNN contains fewer connections and hidden units than the TBNN network structure, thus reducing unnecessary complexity [97]. This reduces computational time while improving accuracy.

2.8 Pattern Classification in MIRG

2.8.1 ANN Research Framework for Medical Outcome

ANNs have been successfully used to predict medical outcomes and has been one of the main pattern classification methods investigated by MIRG. Stemming from their first backpropagation neural network structure, over the years MIRG has used and refined their ANNs (see Appendix F.2). Continued research studies by MIRG have used the backpropagation neural network structure as a base, and have developed it into more sophisticated software. This includes the addition of a verification tool, weight-elimination function, an exhaustive automated structure search and the performance measurements outputs. MIRG has made many important contributions in the field of medical informatics.

2.8.2 Skewed *A Priori* Outcome

It is a common problem in the field of medical informatics: many times, the positive outcome is severely under-represented leaving a highly skewed database. For many conditions, the positive (disease) state is exceedingly lower than the negative outcome, making it

difficult to develop models that will recognize positive cases. Training an ANN with a rare outcome proves to be difficult and results in the poor performance of correctly classifying the true positive; thus resulting in a low sensitivity value. The early work of Ennett and Frize (2000) with ANNs looked at various techniques in improving performance for highly skewed outcomes. In this study, five artificial training sets, with the same sample size, were created. In each of the training sets, the prevalence of non-survivors (post-surgical mortality of coronary artery bypass graft (CABG) cases) were artificially raised from 3.7% to a percentage of 10%, 20%, 30%, 40% and 50%. This change to the training set resulted in a higher sensitivity at the expense of specificity, CCR and AUC. The sensitivity for 3.7% 10%, 20%, 30%, 40% and 50% were 10.85%, 18.18%, 43.55%, 59.09%, 70.45% and 72.73%, respectively. However, the specificity decreased to 98.26%, 96.38%, 93.83%, 82.37%, 74.67%, and 68.47%, respectively. The increase in sensitivity was significant as this meant that the model's correct classification of the rare outcomes increased [98]. The training set that was raised to 20% was found to be the best compromise between sensitivity and specificity in this database. In addition, Ennett and Frize's work (2000) showed an improvement to the sensitivity with the use of weight-elimination.

2.8.3 Techniques to Improve Performance

A study by Frize *et al* examined various techniques to improve the ANN prediction of ventilation hours in an adult ICU. The medical database (from the ICU at the Doctor Everett Chalmers Hospital (DECH) in Fredericton, NB, Canada) contained over 3000 cases with 98 variables. Pre-processing of the data and paring down of variables lead to a subset of the original database with 1491 cases and 51 input variables [99]. A back-propagation three layer network with 'weight-elimination' was used to develop a model. The researchers examined the impact of four techniques to improve the performance results: weight-elimination cost function, constant predictor, 'high-low' node approach for input variables and reduction of network complexity. Using performance measurements—CCR

and AUC—Frize *et al* assessed the impact of these techniques. They concluded that all but the ‘high-low’ node approach made a positive improvement in their results. Such methods may be valuable tools in our research [99, 100].

2.8.4 Weight-elimination

Rybchynski made corrections to a Garson-Goh weight-elimination equation used to calculate the relative weights of the input variables after the ANN classification. By using this equation, the input variables could then be ranked according to their weights. Variables with the lowest weighting were eliminated [14]. A disadvantage of this method was that feature reduction occurs after ANN processing. Subsequently, this is a time-consuming effort to improve the ANN performance—especially for large databases. In contrast, the DT-ANN Hybrid methods chooses the variables before ANN processing occurs [95, 97], which may be a more efficient feature selection method.

2.8.5 Probability Estimation of Classification Outcomes

Zhou and Frize (2006) used the Canadian Neonatal Network (CNN) database to develop a risk level estimation model. In this work, a three-layer ANN was created not only to classify distinct outcomes (i.e. death or survival) in the neonatal intensive care unit (NICU), but also to provide a direct estimate of the “*a posterior* probability” of mortality. In this study, a maximum likelihood estimation models the cost function (instead of the mean squared error function) of the neural network and the weight formula is updated through **gradient ascent**.²⁴ Two approaches were examined as methods to improve sensitivity without changing *a priori* distribution: adjustment of the neural network cost function and adjustment of the classification cut-off point. Though sensitivity of the model improved using both methods; the ‘goodness of fit’ (using the Hosmer-Lemeshow test) of the cost function adjustment was unsatisfactory. Changing the cut-off threshold within the

²⁴The cost function and training algorithm is shown in Appendix C

range of predicted probability proved to be of more value for clinical prediction [43].²⁵ Once the model was developed, the cut-off point was adjusted to improve sensitivity. In this case, decreasing the cut-off threshold to 0.15 improved the sensitivity from 27.5% to 61.9%. The specificity decreased from 99.4% to 95.8% [16]. This compromise, between a large increase in sensitivity in exchange for a small decrease in specificity, was a valuable tool in improving the classification outcome. This research project was also able to predict the severity of the illness (*a posterior* probability) to each case, thus providing a numerical comparison between classification outcomes [16]. This tool could be valuable to this thesis as it provides an adjustment to the cut-off value; thus fine-tuning the sensitivity and specificity. Catley (2007) applied this tool to a more complex medical problem—preterm birth.

2.8.6 Predicting High-Risk Preterm Birth —Initial Findings

Catley’s thesis research, titled “An Integrated Hybrid Data Mining System for Preterm Birth Risk Assessment Based on a ‘Semantic Web Services for Healthcare’ Framework”, established a prediction model for preterm birth and high-risk preterm birth (HR-PTB). Catley’s work presents an in-depth analysis of two perinatal databases, the PPESO and PRAMS database.²⁶ In this research, the model examined the use of obstetrical and socio-demographic information to predict the birth outcome early in gestation (before 23 weeks) and for the general population of expecting mothers.²⁷ Catley *et al* (2005) reported preliminary results for the prediction of PTB. Here she used an 8-variable model derived from the PPESO database. The PPESO database contained 37 medical variables, but since

²⁵The **cut-off threshold** refers to probability values greater than the cut-off threshold (originally set at 0.5) would be classified as death; while a probability value less than the cut-off threshold would be classified as survival.

²⁶**PPESO:** Perinatal Partnership Program of Eastern and Southern Ontario; **PRAMS:** Perinatal Risk Assessment Monitoring System, Atlanta

²⁷The “gold standard” methods (fFN and cervical length measurements) are clinically useful in the middle of gestation (after 23 weeks) and are used in the screening of symptomatic and high-risk patients.

this database was targeted for perinatal use, the number of useful variables was limited.²⁸ Only 8 variables pertained to obstetrical observations available before 23 weeks of gestation: maternal age, number of babies this pregnancy, number of previous preterm babies, number of previous term babies, parity (total, number of previous children), baby's gender, whether mother has intention to breastfeed, and maternal smoking after 20 weeks gestation. Using back-propagation ANN with weight-elimination, Catley *et al* observed 36.64% sensitivity and 93.89% specificity for the prediction of PTB (defined as births <37 weeks of gestation) [101]. In later work on the PRAMS database, Catley risk model evolved to HR-PTB. She explains this as a refinement of the output, as neonates born at 36 weeks and 6 days of gestation may have the same outcome as an infant born at 37 weeks. The network may not be able to sufficiently discriminate cases around the 37 week threshold. Cases <33 weeks were labelled as very PTB and have a higher incidence of medical complications. Therefore, a new risk model was created: HR-PTB was defined as a delivery <33 weeks or <37 weeks accompanied with one of the following factors: (1) Birthweight <1500g, (2) neonate was transferred to intensive care unit, (3) APGAR score < 7, (4) neonate required resuscitation.²⁹ Cases between the 33-36 weeks with these criteria would have lead to greater medical complications. This refinement of the output definition lead to a slight increase of the sensitivity 35.6 ± 0.062 to 37.3 ± 0.057 from PTB to HR-PTB respectively. Meanwhile, the increase in specificity was substantial, 88.6 ± 0.0058 (PTB) to 92.2 ± 0.0053 (HR-PTB) [103]. The PRAMS database was able to offer many more variables that contributed to the outcome prediction. Catley was able to create a 48-variable model that outperformed the PPESO model. One PRAMS **Classification-based** model's verification results were 65.0 ± 0.040 and a specificity of 80.2 ± 0.011 while another model had

²⁸**Perinatal** is the period 5 months before and 1 month after birth

²⁹The **APGAR** score represents the status of the newborn and the response to resuscitation. A score of 0, 1 and 2 is assessed with for each of the five components - Appearance (color); Pulse (heartbeat); Grimace (reflex irritability); Activity (muscle tone) and Respiration (breathing). The score reported at different time intervals indicates certain health status of the newborn where a score of 7 or over indicates a good health status [102] (The APGAR Score, 2006)

a sensitivity of 62.3 ± 0.0273 and specificity of 82.3 ± 0.009 . Catley proceeded to develop a hybrid screening system using **Risk-Stratification** and processing the test data through 3 passes: one to maximize sensitivity, one to maximize specificity and the third to classify ambiguous cases using a decision-tree voting algorithm. The highest clinical performance criteria were able to achieve a sensitivity of 65.0% and specificity of 84.0% [15].

2.8.7 Critical Analysis:

Preliminary results for the PRAMS database were encouraging, as they met clinical expectations set by Catley's clinical advisor Dr. Robin Walker³⁰. Catley's thesis provided the basis of our current research goals: to improve results (higher sensitivity and specificity); to reduce the computational complexity of the 48-variables model (by reducing the number of variables); to create a model that both the US and Canada may be able to use (create a Generalized North American Model) and to identify a set of variables that best predict the outcome.

2.9 Contributions to Support Systems

Effective Computer Decision Support Systems (CDSS) are designed to provide physicians with clinical outcomes which complement the physician's own experience and knowledge. In order for a CDSS to be an ideal system, it must cater to the needs of the users - to be a credible and efficient system. For credibility we need to maximize the clinical prediction accuracy (maintain or exceed current standards); while optimizing the physician's time (minimizing the number of predictor variables needed to input) as well as the time of interaction with the CDSS. In the past, MIRG's contribution to CDSSs focused on the use of ANNs as a pattern classification approach and has resulted in valuable contributions used to improve the sensitivity and specificity of medical outcome estimation in a num-

³⁰"Based on discussions with Dr. C. R. Walker, it was decided that a specificity approaching 85% would indicate a good true negative rate, to be considered acceptable the specificity must be greater than 80%" [15]

ber of environments: ventilation and mortality in a neonatal intensive care unit (NICU), post-surgical mortality of coronary artery bypass graft and high-risk preterm birth in an obstetrical environment [104, 16, 98, 100, 15]. In order to contribute to past research to creating an efficient CDSS, a DT-ANN Hybrid method is proposed. Hybrids, the combination of pattern recognition systems, have resulted in improved classification results [95] or high performance results [105, 20]. This has been attributed to applying the DT's advantages (ability to handle nominal data and missing values, quick classifier generation, and easy interpretation [106, 96, 18]) to an ANN's ability to model non-linear patterns and learn by reducing the error [18]. DTs are a comprehensible method for identifying contributing risk factors and are also used as a feature selection method. By using the DT to select relevant features, the ANN can focus on important variables and further refine the performance results. By combining the advantages of DTs and ANNs for a more efficient CDSS - this thesis work continued the efforts established by MIRG. Here an integrated DT-ANN Hybrid was used to improve the medical outcome estimation of ICU mortality and preterm birth.

3.1 Restating the Problem

Inherently, there are two problems of interest in this thesis work for the estimation of ICU mortality and high-risk preterm birth: (i) the poor ability to correctly identify positive outcomes; and (ii) complexity that exists between predictor variables being used. Rare prevalence of medical conditions and inadequate use of predictor variables play a substantial role with the moderate success of identifying positive outcomes (Sections 2.2.1; 2.5.3; 2.5.4). These are evidence that the relationship between variables have not been successfully modeled by current methods. This remains a problem as there is a high cost associated with wrong predictions (Sections 1.4.1; 1.5.3). The primary objective was to develop an estimation approach by combining the advantages of DTs and ANNs. This integrated hybrid approach was used to minimize input variables (ultimately reducing the complexity of the problem for improved generalization) and optimize the sensitivity while maintaining sufficient specificity.

The first objective was to develop an integrated hybrid system (DT-ANN) to first eliminate variables which do not affect the prediction, and compare this system to the method previously used by MIRG, that is ANN with weight-elimination. This was applied to a small database to estimate ICU mortality; as the main issues for current ICU prognosis models are the data collection burden [54] and the low sensitivity established by many prognosis systems [45, 8]. The second objective was to extend the integrated hybrid approach to a more complex application found in obstetrical care, where the same issues are found—poor sensitivity of the ‘gold standard’ fFN and CL prediction methods (Section 2.5.3 and 2.5.4), biased or poor estimation of past pattern classification approaches on the

databases used (Section 2.6.4) and a high number of predictor variables found to contribute to a higher risk (Section 2.4). The preliminary models designed by Catley to estimate this outcome focused on parous cases and used US specific variables.³¹ This model also used a hybrid method —network parameters were extracted from a classification-based ANN and then applied to a risk-stratification ANN.³² The system processed the test data through three passes: one to maximize sensitivity, one to maximize specificity and the third to classify ambiguous cases using a simple decision-tree voting algorithm. The final test results had a 64% sensitivity, an 84% specificity and an AUC of 0.7795; unfortunately, the model required 48 variables to maintain this performance [15]. While these results met expectations set by the Catley's clinical advisor Dr. Robin Walker³³, there are a few enhancements to be performed: (1) The number of input variables need to be reduced for a more effective and improved generalization of the model; (2) A General North American Model is needed—some variables used in the classifier were only applicable to the US; and (3) In order to produce a system that is applicable to a general obstetrical environment, it must be applicable to parous and nulliparous women. Such modifications should maintain past performance results or preferably, exceed them.

3.2 Ethical Clearance

This thesis work was approved by the Carleton University Research Ethics Committee for the duration of October 2008-August 2010. The databases used in this thesis work: (1) Adult Intensive Care Unit database collected from the Doctor Everett Chalmers Hospital (DECH) in Fredericton, NB, Canada from 1991-1993; and (2) Pregnancy Risk Assessment Monitoring System (2002-2004) was obtained by past MIRG researchers for

³¹**Parous:** women who have had previous children

³²For a more detailed explanation of Catley's thesis work - see Section 2.8.6

³³"Based on discussions with Dr. C. R. Walker, it was decided that a specificity approaching 85% would indicate a good true negative rate, to be considered acceptable the specificity must be greater than 80%" [15]

research work.³⁴ A contract was signed protecting the intellectual property of the data before accessing both databases for this thesis.

3.3 The Adult ICU Database

The medical database was collected from the ICU at the Doctor Everett Chalmers Hospital (DECH) in Fredericton, NB, Canada between 1991-1993 by Heather McGowan-Trigg for her thesis work [99]. This database has had careful analysis of the data to ensure proper values were in place. This database contained 1491 viable records with 26 variables pertaining to the adult ICU —administrative, admission, diagnosis, demographic and APACHE II parameters.³⁵ Four outcomes were available: mortality, ventilation hours, number of days in the ICU and duration of concentrated nursing care (CNC) stay. Seventeen of these variables were chosen as variables that contribute to mortality (the outcome of interest), as shown in Table A.1. The database was split between post-operative (surgery prior to admission to ICU) a ‘homogeneous’ subset of the adult ICU database. The ‘homogeneity’ can be attributed to the fact that admission to the ICU of post-operative patients after surgery was a planned event. It may be apparent during surgery that a greater level of support was needed by the patient than previously planned and ICU care was therefore required [107]. Non-postoperative patients (medical related admissions) are of a ‘heterogeneous’ case-mix containing trauma cases where surgery was not a planned event and a greater level of medical support was necessary.

3.4 The PRAMS Database

The PRAMS database is substantially larger and more complex than the adult ICU database. Our access to the PRAMS database contained over 113,000 cases with over 300 variables relating to demographics, maternal and infant behaviour and medical condi-

³⁴In the past, the Adult ICU data has been used to predict ventilation hours in the ICU [100], while the PRAMS database has been used to predict preterm birth and high-risk preterm birth [15]

³⁵This was pared down in from the original 51 input variables in Heather Trigg’s thesis [99]

tions between the years of 2002-2004. The PRAMS data was collected by the Centers for Disease Control and Prevention (CDC) in Atlanta, USA. PRAMS is a national database from 32 participating states of America. The PRAMS database was established by the CDC in 1987 and using state-specific, population based sampling—data is collected on maternal and infant behaviour and experiences of prenatal, birth and post-partum [108]. The PRAMS database, designed to supplement data from vital records, is used to monitor changes in maternal and child health indicators. The questionnaire contains questions pertaining to pre-pregnancy, pregnancy and post-partum behaviour used as a supplement to original birth certificate variables.³⁶ The PRAMS database was a combination of birth certificate variables (taken at the time of birth) as well as variables collected from the PRAMS questionnaire (collected a few months after delivery).

3.5 Performance Measures of Data Mining Methods

Once a model has been created, its predictive power is evaluated. It is important to measure the performance of classification models to determine how well the models will perform with new cases.

3.5.1 Confusion Matrix (Contingency Table)

The confusion matrix is a matrix that displays the number of scores that are correctly predicted and incorrectly predicted by the model. The figure below displays the confusion matrix for a two-class problem although a confusion matrix can be extended to an n -class problem [109].

³⁶The PRAMS data collection methodology begins with a mail-in questionnaire sent 2 to 4 months after delivery of the child. Non responders are followed up by telephone.

Table 3.1: Confusion Matrix
Actual Value

Prediction Outcome	Actual Value	
	Positive	Negative
Positive	True Positive (TP)	False Positive(FP)
Negative	False Negative (FN)	True Negative (TN)

3.5.2 Correct Classification Rate (CCR)

The number of correctly predicted scores by the classifier. It is also known as the accuracy of the classifier:

$$\text{Accuracy} = \frac{\text{number of correct predictions}}{\text{total number of predictions}} = \frac{\mathbf{TP + TN}}{\mathbf{TP + TN + FP + FN}} \quad (3.1)$$

3.5.3 Error Rate

The number of incorrect prediction. The prediction model for the training data gives the apparent error which is usually lower than that of the true error found using the validation and test data.

$$\text{Error rate} = \frac{\text{number of wrong predictions}}{\text{total number of predictions}} = \frac{\mathbf{FP + FN}}{\mathbf{FP + FN + TN + TP}} \quad (3.2)$$

3.5.4 Sensitivity

The true positive rate (TPR) of the classifier —given that the actual value is positive, the fraction of value that is actually classified as true positive. In medical informatics such rates would be the positive result when the outcome is abnormal [1].

$$\mathbf{TPR} = \frac{\mathbf{TP}}{\mathbf{TP + FN}} \quad (3.3)$$

3.5.5 Specificity

The True Negative Rate (TNR) of the classification model —given that the actual value is negative, the fraction value classified as true negative. In medical informatics, such rates would be the negative result when the outcome is normal [1].

$$\text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (3.4)$$

$$\text{Sp} = 1 - \text{FP} \quad (3.5)$$

3.5.6 Prevalence

Prevalence is the prior probability that the subject has the disease before the test is applied; while the posterior probability is the probability of the disease after a test has been applied. Rare outcomes have low prior probability. In order to assess the usefulness of the test, a comparison between the prior and posterior probability can be made. Predictive values are used for medical diagnostics evaluation to determine the probability that the diagnosis is correct. This is highly dependent on the patient population that the test is used for [110].

3.5.7 Positive Predictive Value (PPV)

The PPV is the probability of the actual disease (positive outcome) given that the prediction is positive. PPV is proportional to prevalence therefore if prevalence of disease is low, PPV will not be close to 1 (even with high specificity and sensitivity). A disease with low prevalence would have a sharp decrease in the PPV. The PPV can be calculated using the following equations [111].

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (3.6)$$

$$\text{PPV} = \frac{\text{sensitivity} \times \text{prevalence}}{\text{sensitivity} \times \text{prevalence} + (1 - \text{spec}) \times (1 - \text{prev})} \quad (3.7)$$

3.5.8 Negative Predictive Value (NPV)

The NPV is the probability of a non-disease or negative outcome given that the prediction is negative. A high NPV indicates a greater chance that a negative test has no abnormality. A rare outcome (low prevalence) would indicate a greater assurance that a patient with a negative test is correctly diagnosed [110]. The NPV can be calculated using the following equations [111]:

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}} \quad (3.8)$$

$$\text{NPV} = \frac{\text{specificity} \times (1 - \text{prevalence})}{\text{specificity} \times (1 - \text{prevalence}) + (1 - \text{sensitivity}) \times \text{prevalence}} \quad (3.9)$$

3.5.9 Receiver Operating Characteristic (ROC) Curve

ROC curves provide a visual model that displays the trade-off between sensitivity and specificity. The ROC curve is produced by graphing the false positive rate (FPR) which is the same as “1-Specificity” against the true positive rate (TPR). The visual model, Figure 3.1 presents points that are of importance as noted by Fawcett (2006).

1. The coordinate (0,0) is located where no positive classification is given therefore no false positive classification exists. The opposite point is (1,1) where the classifier assigns a positive class label to all points, therefore the FPR is at a maximum.
2. An ideal point, *a* is where all the points are correctly classified.
3. At point *b*, the TPR is higher than *a*, but the FPR has increased as well.
4. Point *c* has a low FPR; however, its true positive rate is almost at 60%.
5. The diagonal line, *d*, indicates random guessing. A point on the linear mark indicates that TPR and FPR are equal and the classifier performance is random. This classifier does not contain useful information.

Figure 3.1: An ROC curve and the different points of significance (created from [112, 106]).

6. The point *e* is below the random line and is worse than random guessing. Having a classifier performance in this area may indicate that useful information is contained in the classifier, but the application is incorrect.

Various operating points under different threshold values of the model provides the ROC curve. Though the curve assesses the performance of the model at each individual point, the area under the curve (AUC) presents an assessment of the whole model. The AUC is a scalar value of the ROC curve and is used to compare the performance of different models.

3.5.10 Area Under the Curve (AUC)

The AUC of the diagonal line has a value of 0.5 while the AUC containing point *a*, has a value closer to 1. Values less than 0.5 are worse than random guessing. The AUC is used to compare the performance between classifiers and will always have a value between 0 and 1. The AUC contains statistical property as it is the probability that the “the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance [112].” The AUC is a measurement of the model’s effectiveness in discriminating

outcomes as shown in Table 3.2 [113].

Table 3.2: AUC index and its effectiveness for discrimination

Min	Max	Effectiveness
	≤ 0.500	No discrimination (Random)
0.500	< 0.700	Poor discrimination
0.700	< 0.800	Acceptable discrimination
0.800	< 0.900	Excellent discrimination
0.900	1.00	Outstanding discrimination

3.6 Normalization/Standardization

The ANNs requires the rescaling of input variables, called normalization or standardization, to eliminate the effects of range differences. If normalization is not used, the ANN may calculate the larger values as more important [106, 18, 13].³⁷ In order to scale the values approximately between -1 and 1, the MIRG ANN Guide [114] suggests a modified Z-score transformation equation:

$$\hat{x}_i^n = \frac{x_i^n - \mu_i}{3\sigma_i} \quad (3.10)$$

3.7 Using Hybrid Models

Although the TBNN improves the classification accuracy (see Section 2.7.2), its implementation on the ANN Research-Framework (ANN-RFW) is difficult.³⁸ The ANN-RFW has one hidden layer; therefore the direct mapping algorithm proposed would be computationally expensive. In addition, TBNN was found to be more complex than necessary [97]. The PBNN provides concepts that are of more benefit to MIRG's ANN research studies. Instead of direct mapping of variables and rules to the ANN, the DT is used to

³⁷Variable found in medical databases can range in magnitude as well as meaning. For example partial pressure of oxygen (PO2) ranges from 30-100mm Hg while nerve potentials have a much smaller range (0.01-3mV) [2].

³⁸For more information about ANN-RFW see Appendix F.2.

establish binary patterns which are then inputted into the ANN [97].

3.7.1 Feature Selection

There is a balance between having too many input variables and not enough. Increasing the number of features may improve the performance of the classifier as more information is provided. Too many features can increase the dimensionality of the dataset. This is accompanied with redundancy, computational cost and noise.³⁹ Feature selection methods are needed to determine which features are most valuable for the classifier. Paring down the number of features will reduce the problems that are associated with high dimensionality databases [106]. In turn, this reduces the complexity of the problem and improves the generalization of the classifier [18]. *See5* was the decision tree software used for feature selection. The growing and pruning algorithms used to create a DT, tests for significant contributing variables for the selected outcome (Appendix G). The DT algorithm first classifies each class according to conditions, followed by the pruning steps to global optimum. The pruning portion creates a tree that is able to generalize well and deletes variables with low significance [18]. The features appearing in the pruned DT become the selected subset of features to be used for ANN classification [115].

3.7.2 Decision Tree Ensemble Classifiers

Single trees are known to be fairly unstable especially with smaller databases; designing a tree depends greatly on the dataset and variables used for training and testing. Consequently, relying on the creation of only one tree is simply not enough. In order to maintain robustness in the performance of a DT classifier, an ensemble method was used.⁴⁰ This required the creation of numerous decision trees, a process called ensemble classifiers. The ensemble classifiers were made by bootstrapping the original training sets and

³⁹This is called the “Curse of Dimensionality”, which greatly affects the creation of a classifier.

⁴⁰For a detailed explanation of ensemble methods see Appendix H.3.

allocating these into bootstrapped training subsets (bTRAIN).⁴¹ The DT classifier created a tree from each training subset and tested each tree on the test set.

A typical tree is created using the best split among all variables; however, another issue with DTs is that they choose variables according to local optima. This does not necessarily guarantee a global optimum [17]. Once a tree is created in *See5*, the variables are ranked according to usage where some variables are not used at all. In order to optimize the DT classification once an initial tree was created, a variable used in the split was removed. A new tree was created from the remaining variables. If the performance of the new tree deteriorated after the removal of the variable, the variable was re-introduced, the next variable removed and a new tree created. These steps continued until the best performing tree was discovered. The best performing tree was used as one of the trees of the ensemble classifier. This process was performed for each bootstrapped training subset.

3.7.3 Attribute Interval

The hybrid method was derived from the Pruning-Based Neural Network method (Section 2.7.3) which used decision trees to obtain an input interval layer. Decision trees generated by *See5* contained the variable splits to be used for the attribute intervals. Division of significant features into attribute intervals were based on the threshold value given in the decision nodes of the tree. For continuous values or categorical variables, a feature value falling within the interval range was given a value of 1. If this feature value fell outside the interval range, it was given a value of 0 [97]. The DT-ANN Hybrid method used in this work assigned a value of -1, instead of 0. New features were created for each attribute interval, a modified dataset that carried the information forward to the ANN.

⁴¹For a detailed explanation of bootstrapping see Appendix H.2.

Figure 3.2: Attribute intervals of features using decision tree. Feature 1 (f1) has been split into three groups by two threshold values by the DT. Attribute Interval A, B and C are then created as a result of this split (figure modified from [97])

Single splitting threshold For features with only one variable splitting threshold, cases (found in the original dataset) with feature values greater than the threshold value were given a value of 1. Value less than the splitting criterion were given a value of -1.

Multiple splitting thresholds For features with more than one variable splitting value, cases (found in the original dataset) with feature values less than the minimum value were given a value of -1. If the values fell within the corresponding splitting interval range, they were assigned a value of 1. If the values were greater than the maximum splitting values, they were given a value of 1. All other feature values were given a value of -1.

When using an ANN, the values should fall within the range of -1 to 1, which requires data normalization (Section 3.6). Since continuous and categorical values were already in a binary format; these variables were set for ANN classification. Variables already in this format (two categorical variables such as NO, YES or MALE, FEMALE [-1,1] respectively)—were kept in their original format for ANN classification.⁴²

⁴²Data Analysis and Processing Software - SPSS 17 (from SPSS Inc, a predictive analysis tool) and Excel (from 2009 Microsoft Corporation) were two valuable software used to clean the database, compute new variables, reveal trends, and calculate simple statistical outputs (frequency outputs, distribution, mean, standard deviation).

3.7.4 Logarithmic Sensitivity Index

Optimal performance (early stopping criterion-see Appendix F.2.2) of the network can be determined using the CCR and Average Squared Error (ASE) values. Unfortunately a drawback of these stopping criteria is the poor representation of highly skewed data.⁴³ Another early stopping criterion is the logarithmic-sensitivity *logsens* equation which calculates the sensitivity and specificity balance with the following equation:

$$\text{logsens} = -\text{sensitivity}^n \times \log(1 - \text{sensitivity} \times \text{specificity}) \quad (3.11)$$

The *logsens* value, helps to balance the rarity of some outcome while improving the correct classification rate of the rare outcome. The *logsens* tends to slightly favour a higher sensitivity [116, 117].⁴⁴ Since medical outcomes are usually highly skewed and the abnormal outcome (positive) is usually a rare occurrence, it is difficult to train an ANN. The *logsens* index was created to give more weight to the sensitivity rather than the specificity, thus improving the prediction of the rare outcome [14]. As the sensitivity and specificity approach the value of 1, the value of the *logsens* index tends to approach infinity. To avoid a result of infinity, a constraint was coded: a log function equal to zero will have a log sensitivity of five. The greater the *logsens* value, the better the performance.

3.7.5 Performance Measures for ANNs and DTs

The AUC index should be maximized —a higher AUC index value indicates better discrimination. A high sensitivity and specificity is needed, although in highly skewed datasets (with rare outcomes), such classification results may not be possible. According to Scales [117], MIRG has chosen *logsens* as the ideal performance indicator and is set as the default criterion for the ANN-RFW (see Appendix F.2). This early stopping criterion

⁴³For a detailed explanation of ASE see Appendix H.4

⁴⁴Modification of the n value (a factor used to increase the results toward predicting a higher sensitivity) value can put heavier weighting upon the sensitivity ($(0 \leq n \leq 1)$ [116, 117].

has been shown to be better than the minimum ASE or the CCR in databases with highly skewed outcomes [116, 117]. In conclusion, the CCR, sensitivity and specificity were used to evaluate the DT results; while the sensitivity, specificity, *logsens* and AUC were used to evaluate the performance of the ANN models. The PPV and NPV were also reported.⁴⁵

3.7.6 Developing the DT-ANN Hybrid models

The dataset was divided into the following sets - training, testing and validation sets (to be used for verification of the model).⁴⁶ The training portion of the ANN used the train set and the test set. Both sets are needed to create an ANN structure and balance between the classification of rare outcome and memorization of the data. The test set assisted with validating the model. Training stopped when the minimum of the test set is reached. A single presentation of all patterns to the system is called an epoch. In order to give time for the network to learn, results were not recorded until 25 epochs have passed up to a maximum of 2000 epochs. Slow climbing results will continue indefinitely unless this maximum constraint is in place. The second constraint for a stopping point is when the best performance has been reached and is maintained for the last 500 epochs [14]. The validation portion was divided into 10 different datasets of previously unseen data, labelled as the *verification sets*. Using the best point of parameter trials, a single epoch is passed through the ANN with the corresponding weights, biases and network structure. The ANN performance was evaluated on the 10 verification sets, using the verification tool created by Dawn Rybchynski [14]. A satisfactory performance on the verification sets is indicative of a classifier that generalizes well on new cases. Poor performance of the verification set may be suggestive of over-training the network. The verification results were also used to determine the mean and confidence interval of each performance indicator. Five performance measurements were calculated for each epoch: sensitivity, specificity, CR, ASE, and the AUC of ROC.

⁴⁵The PPV and NPV is useful for physicians; but one has to keep in mind that these values are also greatly affected by skewed outcomes and depend on the patient population to which the test was applied.

⁴⁶The dataset was divided to a suggested portion of 45%, 22% and 33%, respectively [14].

Corresponding optimal weights and biases were saved to be used for the final prediction models.

3.8 First Objective

The first objective was to develop an improved classifier system using an integrated DT-ANN hybrid system that would improve the classifier performance compared to weight-elimination ANNs. As mentioned earlier, an issue in classifier systems used for predicting medical outcomes is the low sensitivity prediction for outcomes with rare prevalence. With a highly skewed prevalence of the positive outcomes, it is difficult for the system to classify positive cases. The first objective was to create a data mining approach that would generate a classifier that improves recognition of both positive cases (sensitivity) and negative cases (specificity). The DT was used to select variables the best contributing variables (using an ensemble feature selection method). The knowledge from the DT was then passed onto the ANN-RFW, an automated program used to create an ANN classifier. By allowing the DT to select an optimal variable set, the number of predictor variables was minimized—thus reducing the complexity of the ANN classifier. While the complexity of the ANN was to be reduced—the prediction performance was expected to be of comparable performance to the weight-elimination ANN classifier. The weight-elimination ANN approach is programmed to effectively reduce the connection weights to zero essentially pruning the network for improved performance. Results of both the weight-elimination ANN and DT-ANN hybrid classifier developed were compared using sensitivity, specificity, *logsens* and AUC as the performance indices.

3.8.1 Accomplishing Objective #1 - Methodology

Step 1 Clean the data using appropriate data-processing methods. Establish the risk model.

Step 2 Split the ICU database into *POSTOP* and *noPOSTOP*, creating two separate databases.

Step 3 Determine results of the DT-ANN Hybrid

- Split each database into a train set (TRAIN) and test set (TEST). Bootstrap the TRAIN set to create five bootstrapped datasets b_TRAIN .⁴⁷
- Classify *POSTOP* and *noPOSTOP* using DTs. Determine significant variables for each of the five b_TRAIN .
- Create attribute interval using splitting values of attributes given by DTs.
- Divide the attribute interval dataset into a TRAIN, TEST and VERIFICATION set. Input attribute interval into the ANN-RFW. Run the ANN-RFW automated search for the optimal number of hidden nodes (3 to 11 nodes) - weight-elimination OFF.⁴⁸

Step 4 Determine results using ANN with weight-elimination (we_ANN)

- Normalize *POSTOP* and *noPOSTOP*
- Divide the normalized dataset into the TRAIN, TEST and VERIFICATION sets.
- Input variables into ANN-RFW. Run the ANN-RFW automated search from 3 to 11 hidden nodes - weight-elimination ON

Step 5 Reproduce DT-ANN Hybrid and we_ANN using a second trial (Trial #2).

- Recombine TRAIN and TEST. Redistribute TRAIN and TEST into new sets with the same prevalence. Run a second trial of the ANN using the same parameters as Step 2 and Step 3.

3.8.2 Data Processing

The outcome of interest was ‘mortality’ for the ICU database. Survival cases were labelled as (-1) and mortality cases as (+1). Initial pre-processing selected predictor vari-

⁴⁷It was observed that within the five bootstrapped training subsets the contributing variables, sensitivity, specificity and correct classification rate remained consistent. Since improvements were not seen - it was decided that five bootstrapped training subsets was sufficient in this thesis work.

⁴⁸The DT was used to create the attribute interval to be inputted into the ANN-RFW. Both POSTOP and noPOSTOP had five variables. The ANN-RFW automated search for optimal nodes was terminated at 11 due to the Kolmogorov superposition theory applied to the context of ANN. The best performing nodes is contained in $2*n+1$ nodes; where n is the number of input variables. The maximum number of hidden nodes is set at $2*n+1$ where n is the number of input variables [118].

ables that contributed to the outcome of interest. Using only diagnosis and APACHE II parameters (mentioned in Section 3.3), the database contained 17 input variables (as described in Table A.1). The Adult ICU database was divided into two parts: cases that have had postoperative surgery (*POSTOP*) and cases without postoperative surgery (*noPOSTOP*). *POSTOP* contained 883 cases while *noPOSTOP* had 608 cases. *POSTOP* had a mortality prevalence of 3.2% (28 cases) and *noPOSTOP* had 12.5% (76 cases). Since decision trees have the ability to deal with nominal, missing and continuous values, the values were kept in their original format.

3.8.3 Creating the DT Hybrid Input Variables (DThy)

Decision Tree Ensemble Classifiers

In order to create a DT classifier for feature selection, the dataset was divided into two sets: a training (TRAIN) and testing (TEST) set. The ensemble classifiers were made by bootstrapping the original training sets five times and allocating these into training subsets (*b_train*). This is presented in Table 3.3. A classifier was created for each of *b_train* using the process outlined in Section 3.7.2.

Attribute intervals

All five trees created by the *POSTOP* and *noPOSTOP* datasets were examined and the trees with the highest sensitivity and specificity were chosen to make the attribute intervals. The attribute intervals were produced by using the splitting variables utilized by the DTs (see Section 3.7.3). The original *POSTOP* and *noPOSTOP* datasets were used to make the hybrid input dataset (*h_POSTOP* and *h_noPOSTOP*, respectively).

3.8.4 ANN Classification

Normalization of the data

ANN classification requires the normalization of the data to ensure that values of higher magnitude are not given greater weighing. Both *POSTOP* and *noPOSTOP* were normalized to create *POSTOP_N* and *noPOSTOP_N*, respectively.

Table 3.3: Division of *POSTOP* and *noPOSTOP* dataset used to create the each of the DT classifier. (A) Number of Cases, (B) Percentage out of Total Cases (n=883 for *POSTOP* and n=608 for *noPOSTOP*), (C) Number of Positive cases (mortality) (D) Prevalence of Positive Cases (%), (E) Process used

		A	B	C	D	E
POSTOP	TRAIN_P	526	59.6	18	3.4	POSTOP train set
	b1_train_P			17	3.4	
	b2_train_P			18	3.6	
	b3_train_P	495	NA	16	3.2	bootstrapped training subsets (about 95% of training cases)
	b4_train_P			18	3.6	
	b5_train_P			18	3.6	
	TEST_P	357	40.4	10	2.8	POSTOP test set
noPOSTOP	TRAIN_P	371	61.0	46	12.3	noPOSTOP train set
	b1_train_N			44	12.5	
	b2_train_N			42	12.0	
	b3_train_N	350	NA	44	12.5	bootstrapped training subsets (about 95% of training cases)
	b4_train_N			44	12.5	
	b5_train_N			42	12.0	
	TEST_N	237	38.9	30	12.6	noPOSTOP test set

Division of the Dataset - Train, Test, Verification

The hybrid and normalized datasets were split into the training (TRAIN), test (TEST) and validation (VALIDATION) sets as shown in Table 3.4. The validation set was not used in creating the ANN classifier, but was used to determine how well the classifier performed on new cases. Therefore, these sample sets do not have to be exactly the same size as they mimic ‘independent cases’. To reiterate, there is a limited amount of positive outcomes (mortality) found in the adult ICU database. This requires the training sets to be

artificially re-sampled to a greater prevalence in order for the ANN to successfully identify positive cases. In order to create the validation confidence interval, the validation set is usually divided into 10 verification sets. Again, the limited amount of positive cases requires bootstrapping the validation set to create the 10 verification sets. In the *POSTOP* cases (*h_POSTOP* and *POSTOP_N*), the validation set has 221 cases. Distribution of positive cases among the ten verification sets is virtually impossible; therefore bootstrapping was used to produce 10 verification sets with approximately 200 cases each (about 90%). In the *noPOSTOP* cases (*h_noPOSTOP* and *noPOSTOP_N*), the validation set contained 146 cases with 19 positive cases; bootstrapping was used to produced 10 verification sets with approximately 130 cases each. The number of positive cases varied slightly among each verification set.

Table 3.4: Division of *POSTOP* and *noPOSTOP* databases for ANN using attribute intervals and normalized datasets. (A) Number of Cases, (B) Percentage out of Total Cases (n=883 for *POSTOP* and n=608 for *noPOSTOP*), (C) Number of Positive cases (mortality) (D) Prevalence of Positive Cases (%)

		A	B	C	D
h_POSTOP	TRAIN [†]	415	46.9	14	3.37
and	TEST*	247	27.9	8	3.23
POSTOP_N	VALIDATION [◊]	221	25.0	6	2.71
h_noPOSTOP	TRAIN*	280	46.0	35	12.50
and	TEST*	175	28.8	22	12.57
noPOSTOP_N	VALIDATION [◊]	153	25.1	21	13.72

[†]Oversample of positive cases and under-sample of negative cases to 5.06%. *No change [◊]Bootstrap using about 90% of data to create 10 Verification sets.

ANN classification parameters

Table A.3 gives the starting parameters for both the DT-ANN hybrid and ANN with weight elimination (we_ANN).⁴⁹ Since the DT-ANN Hybrid has determined the significant variables for classification —the weight-elimination function was de-activated. This is

⁴⁹See Appendix F.1.6 and Appendix F.2.1 for detailed information about weight-elimination and the starting parameters.

compared to the regular ANN where the weight-elimination function was activated.

3.8.5 Reproducing the classifiers

Since there was a very limited amount of positive cases found in both database (due to the rare outcome), it is understandable that different cases allocated into the training and test set may vary in the final results. Large deviations in the final performance indices are indicative of poor classifier validation. In order to validate the ANN classified and check for robustness, a second ANN run was performed. This was performed by recombining the training and test set; while excluding the verification set. The recombined data was redistributed for the second trial (Trial#2) using the same train and test rate. A second set of ANN classifiers was created using the new train and test cases. These results were compared to the first ANN trial (Trial #1) in order to examine which number of hidden nodes produced the best results and to determine if the results for each hidden node were reproducible in the second trial (Trial#2).

3.9 Second Objective

The second objective was to create a classifier system that improved or matched preliminary prediction results of the integrated hybrid system developed by Catley [15]. Once a fundamental design was established and implemented on a small database (first objective), the integrated DT-ANN hybrid approach was extended onto a large, complex, heterogeneous dataset pertaining to a multi-factorial problem - the prediction of HR-PTB. Detailed descriptions of the main criteria to fulfill in this estimation model were as follows:

Extend the model to include Parous and Nulliparous cases: Preliminary results for Nulliparous outcome estimation are to be established. Catley's work emphasized the importance of the variable previous preterm birth (PREV_PRE) as a predictor of preterm birth [15]. In her data processing steps, she focused on 32,022 cases with the deletion of over 72.0% of the database. This deletion was due to incomplete

data for the variable PREV_PRE. A closer look at the original PRAMS database (see Appendix B.2) showed missing values for PREV_PRE because the event did not occur. For these cases, the current pregnancy was the first child born to the mother (nulliparous). Therefore, Catley's research concentrated on parous mothers and did not include mothers who did not have previous live births; thus limiting the generalized application of the model. A model should be applicable to both parous and nulliparous women.

Reducing the number of input variables: HR-PTB is a multi-factorial problem and many predictor variables have been associated with this outcome. In Catley's thesis, she needed 48 variables to meet clinical standards. In order to optimize the physician's time and reduce the complexity of the problem, this work attempts to reduce the number of input variables. This is performed by using DTs to select significant variables prior to analyzing with ANNs.

General North American Model: MIRG's research collaboration extends beyond Canadian borders - with access to provincial databases (PPESO) and US databases (PRAMS). In this research, variables specific to USA (Medicaid and Women Infants Children Program) were eliminated while attempting to improve past results. However, variables pertaining to maternal race such as HAWAII, ALASKA, PACIFIC are specific to the US. Since maternal race is applicable to all countries, contribution of these variables may be important to classification and related to demographic factors—these variables were not omitted. These variables may be comparable to a Canadian database where provinces may have similar contribution to the classification of HR-PTB.

These enhancements were to be applied in an attempt to maintain or exceed past performance of sensitivity of 65.0%, specificity 84.0% and AUC 0.7795. By improving these aspects of the prediction model; the complexity of the problem is reduced thus advancing

the generalization of the model to a general obstetrical population.

3.9.1 Accomplishing Objective #2 - Methodology

Step 1 Establish the risk model and outcome of interest

Step 2 Clean the data using appropriate data-processing methods

Step 3 Divide the PRAMS database into *Parous* and *Nulliparous* datasets

Step 4 Create the DT- ANN HYBRID • Split each database into a train set (TRAIN) and test set (TEST). Bootstrap the TRAIN set to create three bootstrapped datasets *b_TRAIN*.⁵⁰

- Classify *Parous* and *Nulliparous* using DTs—determine significant variables for each of the three *b_TRAIN*.
- Create attribute interval using splitting values of attributes given by DTs.
- Divide the attribute interval dataset into a TRAIN, TEST and VERIFICATION set.
- Input into the ANN-RFW and create a Hybrid Classification based ANN (weight-elimination ON).
- Choose the best performing Classification based ANN and input into Risk Stratification ANN.

3.9.2 Establishing the Risk Model

Cases are classified as PTB if the gestational age was less than 37 complete weeks. In order to better discriminate between cases of high-risk infants (which have increased morbidity and mortality rates), the high-risk preterm birth (HR-PTB) risk model was used. Each case was then classified into high-risk (+1) or low-risk births (-1) as defined by Catley

⁵⁰It was observed that significant contributing variables did not remain consistent among each DT. Some variables consistently appeared in each bootstrapped DT; while other variables differed among each DT. Although contributing variables changed with each *b_TRAIN*, the accuracy results had small fluctuations. Therefore only three bootstrapped sets were used as the contributing variable inconsistency was expected to continue if more bootstrapped subsets were used (Section 4.5.1)

et al (2006). HR-PTB was defined as births that met the following criteria:

- (1) Gestational Age < 33 weeks**
- (2) Gestational Age <37 and birthweight < 1500 g**
- (3) Gestational Age < 37 and admitted to the neonatal Intensive Care Unit**

3.9.3 Data Processing

Cleaning the Data

Catley's preliminary research had identified risk factors for HR-PTB outcomes and also indicated the need for early prediction (before 23 weeks of gestation) [15]. Earlier outcome prediction allows more time for the physician to implement a treatment plan. Consequently, the input variables pertain to risk factors that should be available early in the pregnancy (< 23 weeks). The PRAMS database was examined to deal with missing values, outliers and anomalies. Missing values were investigated to determine if values may be inputted with logical reasoning. Cases with outliers were removed and anomalies with the data were repaired or removed. A detailed outline of the processing steps can be found in Appendix B.2. Preliminary data processing ensures that nominal variables with values of NO and YES always have a value of -1 and 1, respectively. Since *See5* has the ability to deal with missing values as well as continuous variables the datasets were not normalized.

Parous and Nulliparous Cases

In Catley's thesis, she describes PREV_PRE as missing in over 50% of the cases. In her processing steps, variables missing over 20% of cases were deleted; however PREV_PRE was found to be an important indicator for PTB. Careful examination of the PRAMS database revealed that the cases with missing values for PREV_PRE was due to the fact that the current case was the first delivered child (nulliparous). Analysis showed that when another variable (number of previous children—PRE_LB) had a value of 0, over 97.9%

cases were missing a value for PREV_PRE. The PREV_PRE variable is irrelevant to nulliparous women and resulted in a missing value. Therefore, the PRAMS database was then split into two birth categories: (i) *Parous* —Infants born to mothers with previous children, and (ii) *Nulliparous* —Infants born to mothers with no previous children. This separation is due to the fact that certain variables found in *Parous* are irrelevant for *Nulliparous*: (i) parity (number of previous children), (ii) previous preterm birth, (iii) previous low birth-weight, and (iv) previous children. These variables have been found to be important for the prediction of high-risk PTB [15]; however, they are not applicable to nulliparous women. Therefore, a different set of variables was required to create a classifier for *Parous* and *Nulliparous* cases.

3.9.4 Creating the DT Hybrid Input Variables (DThy)

Input Variables for DT

The initial variable list (Table A.4) pertains to variables to be inputted into See5 for feature selection. This includes some of the 48 variables identified as risk factors in Catley's thesis [15]. Additional predictor variables discovered during initial data processing steps and suspected to improve classification were also included. One of PRAMS objectives is to supplement the vital record, and therefore, the database contains some confounding variables (smoking, drinking, and morbidity variables). Some of these variables originated from the core birth certificate information while some came from the PRAMS questionnaire. Unfortunately, in some cases, the confounding variables have been found to contain conflicting information [108]. Confounding variables found in the PRAMS questionnaire and the birth certificate were included as *See5* would be used to select the best feature among the list. Table A.4 is a list of all the variables, an explanation of the variables and indicates which variables originated from (1) PRAMS questionnaire, (2) birth certificate, or (3) was created from data processing steps in this thesis work.

Dividing the Dataset

The whole dataset was split into a training and test set: 60% of the data was allocated for the training (TRAIN) portion and tree generation and 40% of the data was used for the testing portion (TEST) of the tree classifier. The training set was bootstrapped three times (using about 90%), creating the bootstrapped training subsets (bTRAIN).

Table 3.5: Division of *Parous* and *Nulliparous* dataset used to create the each of the DT classifier. (A) Number of Cases, (B) Percentage out of Total Cases (n=40297 for *Parous* and n=29925 for *Nulliparous*), (C) Number of Positive cases (HR-PTB), (D) Prevalence of Positive Cases (%), (E) Process used

		A	B	C	D	E
Parous	TRAIN_P	24096	59.7	3860	16.0	Parous train set
	b1_train_P			3690	16.1	bootstrapped training subsets (about 95% of training cases)
	b2_train_P	22900	NA	3651	15.9	
	b3_train_P			3684	16.1	
	TEST_P	16201	40.2	2601	16.1	Parous test set
Nulliparous	TRAIN_P	17915	59.8	3204	17.9	Nulliparous train set
	b1_train_N			3037	17.9	bootstrapped training subsets (about 95% of training cases)
	b2_train_N	17000	NA	3038	17.9	
	b3_train_N			3028	17.8	
	TEST_P	12010	40.1	2518	17.9	Nulliparous test set

Ensemble Decision Tree

Redundant or similar data that came from the core birth certificate variables and PRAMS collected questionnaire were used for the initial DT classifier. All the variables that have been identified as risk factors for HR-PTB were inputted into the DT. There were 51 input variables for *Parous* and 48 input variables for *Nulliparous*. All these variables were used to generate the first DT. *See5* displays which attributes have the most contribution to the classifier in terms of percentage. The steps used in the creation of each of the three DT were as follows:

1. Create a DT using all the predictor variables found in Table A.4 including confounding variables.

2. Attributes that contribute to less than 1% to the classifier are not displayed in the attribute usage list. We ignore these variables when creating the next tree. For confounding variables (i.e. MM_BLEED and MORB_BLD), ignore the variable with the least contribution according to the attribute usage.
 3. Re-generate a new tree using the modified variable subset list.
 4. Remove each variable iteratively before regenerating a tree until the tree has optimal sensitivity, specificity and minimal error.
- . The DT with the best compromise between sensitivity and specificity was used as the classifier for the production of the attribute intervals to create the hybrid database *h_Parous* and *h_Nulliparous*. Attribute intervals were created as outline in Section 3.7.3.

3.9.5 ANN classification

Division of the dataset

The datasets *h_Parous* and *h_Nulliparous* were randomly sampled at 35% to create the ANN classifier. The original *h_Parous* set contained 40,297 cases. 14,124 cases were used for the ANN classification with 2,277 positive cases. Since the training of ANNs are known to be a computational expensive and a time consuming process, this method was utilized to decrease the time needed to classify the patterns. Once 35% of the data was obtained, the sampled data was divided into three different sets: training (TRAIN), testing (TEST) and validation (VALID) as shown in Table 3.6.⁵¹ The validation set was then split into 10 different verification sets (VERIF). The original distribution of the datasets was preserved. ANN classification for HR-PTB used the starting parameter values as shown in Table A.7.

⁵¹The proportion of each sets were approximately 4/9, 2/9 and 3/9 respectively

Table 3.6: Division of *h_PAROUS* and *h_NULLIPAROUS* databases for ANN using attribute intervals and normalized data. (A) Number of Cases, (B) Percentage out of Total Cases (n=14124 for *h_PAROUS* and n=10554 for *h_NULLIPAROUS*), (C) Number of Positive cases (HR-PTB), (D) Prevalence of Positive Cases (%).

		A	B	C	D
<i>h_Parous</i>	TRAIN [◊]	6260	44.3	1010	16.1
	TEST [◊]	3120	22.1	502	16.1
	VALID [†]	4744	33.5	765	16.1
<i>h_Nulliparous</i>	TRAIN [◊]	4648	44.0	830	17.9
	TEST [◊]	2333	22.1	406	17.4
	VALID [†]	3573	33.8	633	17.7

[◊]No change. [†]Divided into 10 evenly distributed validation sets

3.10 Methodology Conclusion

The methodology described in this chapter provides a technique to investigate the applicability of an integrated DT-ANN hybrid system on two different outcome estimation: ICU mortality and high-risk preterm birth. An assessment determined if a difference existed between the DT-ANN Hybrids' performances versus that of an ANN with weight-elimination (we_ANN). Expectations of the DT-ANN Hybrid and we_ANN were to function in a similar manner—as the design of both systems creates a classifier that effectively uses the most significant variables. The DT-ANN Hybrid performs this at the embedded DT level, through the growing and pruning of the DT; whereas we_ANN effectively reduces connections weights to zero during ANN training. The results and corresponding discussions are given in the next chapter.

Results and Discussion

4.1 Creating the Attribute Intervals using Decision Trees

4.1.1 POSTOP Decision Tree Results

The Decision Tree (DT) results (Table 4.1) showed consistent accuracy among the five bootstrapped datasets in terms of attribute usage, sensitivity (which were all at 90%) and specificity (a range from 83.9% to 84.1%). The variables PH, EMER_SRG (ES) and PO2 were used to create all five DTs. The variables that differed among the five trees were TEMP and AGE. TEMP was used to create DT 1 and 2; while AGE was used to create DT 4 and DT 5. The root variable, PH, remained the same among the five classifiers and EMER_SRG remained the second variable in four of the five classifiers. Since the attribute usage of both AGE and TEMP were high (found at 87% and 88%) in each of the classifier - both variables were used to create the attribute intervals. The attribute intervals were created according to the DT splitting values found in Table 4.15, as outlined in Section 3.7.3.

4.1.2 noPOSTOP Decision Tree Results

The decision tree classifiers consistently used GCS, FI02, MAP, and AGE (Table 4.2) to classify the noPOSTOP data, except for DT 2, which excluded GCS. Instead, DT 2 used HEMA and NA to classify the data. Since NA was the lowest ranking attribute for this classifier, it was not used to make the attribute interval. The noPOSTOP test sensitivity and specificity were lower than the POSTOP accuracy results; noPOSTOP had a sensitivity range from 73.3 to 83.3% and specificity range from 63.8 to 67.6%. It was also observed, that even though the same variables were found in the attribute usage list used to classify

the five bootstrapped training sets, the root variable, FIO2 and GCS, differed with each DT. MAP was always listed as the second attribute. The splitting values of the DTs were used to make the attribute intervals as given in Table 4.4.

Table 4.1: POSTOP decision tree results and list of attribute usage

DT #		1	2	3	4	5
Train	Sensitivity	1.0	1.0	87.5	94.4	94.4
	Specificity	84.9	85.5	86.4	85.5	85.5
	CCR	85.5	86.0	86.5	85.6	85.9
Test	Sensitivity	90.0	90.0	90.0	90.0	90.0
	Specificity	83.9	83.9	85.4	84.1	84.1
	CCR	84.0	84.0	85.4	84.3	84.3
Attribute Usage		100% PH	100% PH	100% PH	100% PH	100% PH
		97% ES	97% ES	97% PO2	97% ES	97% ES
		88% TEMP	88% TEMP	93% ES	87% AGE	88% AGE
		86% PO2	86% PO2		86% PO2	86% PO2

Table 4.2: noPOSTOP Decision tree results and list of attribute usage

DT #		1	2	3	4	5
Train	Sensitivity	100.0	100.0	100.0	100.00	100.0
	Specificity	69.3	70.5	67.3	65.7	69.8
	CCR	73.1	74.0	71.4	70.0	73.4
Test	Sensitivity	83.3	73.3	83.3	83.3	83.3
	Specificity	67.6	72.9	63.8	63.8	67.6
	CCR	69.6	72.9	66.2	66.2	69.6
Attribute Usage		100% GCS	100% FIO2	100% FIO2	100% GCS	100% GCS
		67% MAP	71% MAP	68% MAP	67% MAP	68% MAP
		64% FIO2	40% HEMA	40% GCS	40% GCS	63% FIO2
		13% AGE	18% AGE	22% AGE	23% AGE	12% AGE
			11%NA			

Table 4.3: POSTOP Attribute Interval Splitting Values

Variable	Threshold
PH	≤ 7.19 7.19
EMER_SRG	-1 1
PO2	≤ 61 61
TEMP	≤ 34 34
AGE	≤ 86 86

Table 4.4: noPOSTOP Attribute Interval Splitting Value

Variable	Threshold
GCS	≤ 14 14
FI02	0.45 0.45
MAP	≤ 89.7 89.7
AGE	≤ 60 60
HEMA	≤ 40 40

4.2 POSTOP ANN Results

4.2.1 POSTOP DT-ANN Hybrid Results

Results of the best performing ANN classifiers were Node 6 and 11, given in Table 4.5.⁵² The *POSTOP* DT-ANN Hybrid results were similar between Trial#1 (Tr1) and Trial#2 (Tr2). Test sensitivity for both Node 6 (Tr1 and Tr2) and Node 11 (Tr1) remained at 75.00%. The specificity was 89.54% for both nodes found in the Tr1. Node 6 (Tr2) had the highest test specificity at 92.05%. Node 11 (Tr2) had the highest sensitivity of 100.00%, accompanied with the lowest test specificity (82.85%). The AUC scores showed outstanding discrimination (Table 3.2- Page45), as the test AUCs were above 0.900 - except for Node 6 (Tr2) which had the lowest test AUC at 0.8394. The verification results of Tr1 showed a slight increase to its sensitivity (83.38 ± 2.09), specificity (90.75 ± 0.56) and AUC (0.9586 ± 0.0026) compared to its test results. In Tr2, the verification in Node 6 had the same verification sensitivity (83.38 ± 2.09) and specificity (90.75 ± 0.56); however the AUC increased to 0.9600 ± 0.0026 . In Node 11, the verification sensitivity was the highest (100.00 ± 0) among all nodes found in the two trials, subsequently resulting in a slight decrease in verification specificity and AUC. It had the lowest specificity (86.53 ± 0.85) and AUC (0.9374 ± 0.0078) among all classifiers found in the two trials. All of the

⁵²Nodes refer to number of **hidden** nodes found in the hidden layer of the MLP. For all POSTOP classifier results (DT-ANN Hybrid and we_ANN) see Appendix D

verification *logsens* values indicated an ideal balance between sensitivity and specificity as they were all at 0.500 or higher. The mortality prevalence in the post-operative cases was low, at 3.24%. This needed to be re-sampled to 5.06% in order to achieve the results given.

Table 4.5: POSTOP Best Performing Hybrid ANN without weight-elimination (Node 6 and 11). Verification (VERIF) is the mean value and standard deviation of the 10 verification sets. TRAIN prevalence is 5.06% and TEST prevalence is 3.24%. Verification rate is 3.04 ± 0.38 .

Nodes		6	11
Trial#1 (Tr1)			
TRAIN	Sensitivity	71.43	71.43
	Specificity	92.64	92.64
	logsens	0.3362	0.3362
	AUC	0.9321	0.9321
TEST	Sensitivity	75.00	75.00
	Specificity	89.54	89.54
	logsens	0.3596	0.3627
	AUC	0.9417	0.9467
VERIF	Sensitivity	83.38 ± 2.09	83.38 ± 2.09
	Specificity	90.75 ± 0.56	90.75 ± 0.56
	logsens	0.5134 ± 0.0406	0.5134 ± 0.0406
	AUC	0.9586 ± 0.0026	0.9586 ± 0.0026
Trial#2 (Tr2)			
TRAIN	Sensitivity	66.67	90.48
	Specificity	90.86	85.08
	logsens	0.2695	0.5763
	AUC	0.8649	0.9003
TEST	Sensitivity	75.00	100.00
	Specificity	92.05	82.85
	logsens	0.3819	0.7656
	AUC	0.8394	0.9098
VERIF	Sensitivity	83.38 ± 2.09	100.00 ± 0
	Specificity	90.75 ± 0.56	86.53 ± 0.85
	logsens	0.5134 ± 0.0406	0.8716 ± 0.0277
	AUC	0.9600 ± 0.0026	0.9374 ± 0.0078

4.2.2 POSTOP ANN with weight-elimination

The best performing we_ANN models (by test results) of Tr1 were determined to be Node 5, 6, and 11 (Table 4.6). Among the classifiers in Tr1, Node 5 had the highest sensitivity of 50.00% but the lowest specificity of 79.50%. Node 6 and Node 11 had the same sensitivity (37.50%), but Node 6 had a higher specificity (98.75%) compared to Node 11 (80.75%). Node 11 had the poorest AUC score of 0.4686 indicating that this classifier was not better than random (Table 3.2 - Page 45). Although Node 5 was higher at an AUC of 0.6017; and Node 6 had the highest AUC at 0.6645, these were all found to be poor discriminators. Surprisingly, the classifiers verified well using an independent set unseen by the classifier. There was a substantial difference between the test results and verification results. A noticeable increase was found in the AUC which became of acceptable discrimination (Node 11, AUC of 0.7995 ± 0.0272), to excellent discrimination (Node 5, AUC of 0.8021 ± 0.0424 and Node 6, AUC of 0.8913 ± 0.0306). Another was found in the sensitivity and specificity. Node 5, 6 and 11 had a value of sensitivity to specificity as follows: 83.33 ± 6.09 to 72.61 ± 0.84 , 48.00 ± 7.89 to 97.57 ± 0.24 and 66.67 ± 7.70 to 78.81 ± 1.05 , correspondingly. The *logsens* values were highest for Node 5 (0.3416 ± 0.0787) and lowest for Node 6 (0.1383 ± 0.0553).

According to Tr1, Node 5 was the best ANN model; however, it was evident in Tr2 that the second trial was inadequate in reproducing the results found in Node 5 (Tr1). According to the test results in Tr2, Node 5 was the best classifier with a sensitivity of 62.50% and specificity of 97.07%. These values were better than the Node 5 (Tr1). The AUC also increased by 0.1551 (from Tr1) towards acceptable discrimination (0.7568). Unfortunately in the verification set, Node 5 failed to reproduce its test sensitivity. The classifier's inability to predict some of the verification cases accounted for the skewed sensitivity to 18.67 ± 1.72 , an indication of over-fitting. This was not a useful classifier as it is unable to classify

new cases. In Tr2, Node 6 and 11 both had an increase to its test sensitivity from 35.00% (Tr1) to 50.00% (Tr2). The test specificity remained the same, less than a percent difference. Node 6 and 11 (Tr2) also verified well as its values did not deviate far from the test values of Tr2. The best classifier for we_ANN was chosen to be Node 11 due to its verification sensitivity and specificity, its acceptable discrimination ($AUC = 0.7712 \pm 0.0321$) and its ability to reproduce the verification results from the first trial. Both verification sets found in Tr1 and Tr2 showed the same sensitivity (66.67 ± 7.70 for both trials); while its specificity was 78.81 ± 1.05 (Tr1) and 75.40 ± 0.93 (Tr2) respectively. Tr1 indicated an AUC of 0.7995 ± 0.0272 ; while Tr2 had an AUC of 0.7712 ± 0.0321 . Although there was good verification for this node, the test AUC from both trials, indicated poor or random discrimination.

Table 4.6: POSTOP Best Performing ANN with weight elimination (Node 5, 6 and 11). Verification (VERIF) is the mean value and standard deviation of the 10 verification sets. TRAIN prevalence is 5.06% and TEST prevalence is 3.24%. Verification rate is 2.71 ± 0.23 .

Nodes		5	6	11
Trial#1 (Tr1)				
TRAIN	Sensitivity	61.91	76.19	61.91
	Specificity	75.64	100.00	83.50
	logsens	0.1698	0.4749	0.1956
	AUC	0.7597	0.9236	0.8158
TEST	Sensitivity	50.00	37.50	37.50
	Specificity	79.50	98.75	80.75
	logsens	0.1100	0.0753	0.0587
	AUC	0.6017	0.6645	0.4686
VERIF	Sensitivity	83.33 ± 6.09	48.00 ± 7.89	66.67 ± 7.70
	Specificity	72.61 ± 0.84	97.57 ± 0.24	78.81 ± 1.05
	logsens	0.3416 ± 0.0787	0.1383 ± 0.0553	0.2214 ± 0.0653
	AUC	0.8021 ± 0.0424	0.8913 ± 0.0306	0.7995 ± 0.0272
Trial#2 (Tr2)				
TRAIN	Sensitivity	14.29	85.71	57.14
	Specificity	97.72	100.00	79.44
	logsens	0.0093	0.7244	0.1502
	AUC	0.7015	0.9266	0.7072
TEST	Sensitivity	62.50	50.00	50.00
	Specificity	97.07	99.16	79.08
	logsens	0.2533	0.1487	0.1093
	AUC	0.7568	0.7390	0.5704
VERIF	Sensitivity	18.67 ± 1.72	50.00 ± 8.16	66.67 ± 7.70
	Specificity	94.83 ± 0.47	98.09 ± 0.24	75.40 ± 0.93
	logsens	0.0159 ± 0.0030	0.1540 ± 0.0595	0.2075 ± 0.0615
	AUC	0.8695 ± 0.0159	0.8213 ± 0.0429	0.7712 ± 0.0321

4.2.3 Comparing POSTOP DT-ANN Hybrid with we_ANN

Node 11 was chosen as the best performing model to compare between the DT-ANN Hybrid and we_ANN. In this comparison, as shown in Table 4.7, the DT-ANN Hybrid showed a substantially higher sensitivity, specificity and AUC. In terms of sensitivity, the DT-ANN Hybrid had a test sensitivity of 75.00% which was 25.00% greater than the we_ANN (50.00%). The DT-ANN Hybrid test specificity (89.54%) was 10% greater than

of the we_ANN (79.09%). The greatest difference was established in the discrimination capabilities (AUC) where the DT-ANN Hybrid (0.9467) outperformed the we_ANN (0.5704) by 0.376. Verification of both classifiers presented an increase in all values compared to its corresponding test result. Again, the DT-ANN Hybrid outperformed the we_ANN with a mean sensitivity greater than the we_ANN by 16.68%; specificity by 11.88% and the AUC by 15.9% (as shown in Table 4.7).

Table 4.7: Comparing the best performing classifiers for POSTOP between the DT-ANN Hybrid and we_ANN.

Node		11	11
Trial		1	2
ANN		DT-ANN Hybrid	we_ANN
TEST	Sensitivity	75.00	50.00
	Specificity	89.54	79.085
	logsens	0.3627	0.1093
	AUC	0.9467	0.5704
VERIF	Sensitivity	83.38 $\pm$ 2.09	66.67 $\pm$ 7.70
	Specificity	90.75 $\pm$ 0.56	75.40 $\pm$ 0.93
	logsens	0.5134 $\pm$ 0.0406	0.2075 $\pm$ 0.0615
	AUC	0.9586 $\pm$ 0.0026	0.7712 $\pm$ 0.0321

4.3 noPOSTOP ANN Results

4.3.1 noPOSTOP DT-ANN Hybrid Results

The best performing ANN nodes⁵³ from Tr1 were Node 5, 10 and 11 as shown in Table 4.8. The highest test sensitivity was found in Node 5 (90.90%), which had a specificity of 75.16% and an AUC of 0.8333. This classifier verified well, with a sensitivity and specificity of 89.50 ± 2.59 and 73.58 ± 1.05 , respectively; and an AUC verification result of 0.8354 ± 0.0146 . The test results of Node 10 also performed well with a slightly lower test sensitivity (81.82%), increased specificity (81.70%) and slightly lower AUC (0.8277) compared to that of Node 5 (Tr1). This classifier also verified well, with a sensitivity and

⁵³For all noPOSTOP classifier results (DT-ANN Hybrid and we_ANN) see Appendix D

specificity of 79.13 ± 3.02 and 78.47 ± 1.23 , correspondingly. This was also of excellent discrimination as the AUC was 0.8238 ± 0.0128 . Though the test sensitivity of Node 11 was acceptable (77.27%); its verification results had close to a 10% drop in its mean sensitivity (68.30 ± 6.05). Its specificity essentially remained the same at 80.39% (test) and 81.63 ± 1.04 (verification). Though Node 11 had the lowest test AUC (0.7897), this increased to 0.8292 ± 0.0209 in its verification results. Test results observed in Node 5 (Tr2) had a decrease in the sensitivity (86.36%), and a slight increase in specificity (77.12%) and the AUC (0.8437) from the Node 5 (Tr1) test results. Node 10 had a sensitivity of 90.90% (increased by 9.1%) from Tr1, a decreased specificity by 11.77%, and a slightly lower AUC (0.8102). However, Node 10 verifies well on unseen data with a sensitivity of 89.50 ± 2.59 , specificity of 76.48 ± 1.44 , and AUC of 0.8118 ± 0.0140 . Node 11 (Tr2) had a slightly higher sensitivity (86.36%), a decreased specificity (72.54%) and an increased AUC (0.8265) compared to Node 11 (Tr1). Verification of Node 11 (Tr2) gives the lowest sensitivity among all three nodes in Tr2 (68.30 ± 6.05); however, remained consistent with verification results in Node 11 (Tr1). Node 11 (Tr2) had an AUC similar to its first trial (0.8120 ± 0.0236). The best DT-ANN Hybrid chosen for comparison with we_ANN was Node 5 (Tr1) due to its high test sensitivity and specificity, its high specificity ratio (0.4540), high AUC value (0.8333) and good verification results.

Table 4.8: noPOSTOP Best Performing Hybrid ANN without weight-elimination (Node 5, 10 and 11). Verification (VERIF) is the mean value and standard deviation of the 10 verification sets. TRAIN rate is 12.50% and TEST rate is 12.57%. Verification rate was 12.47 ± 1.15 .

Trial#1 (Tr1)	Nodes	5	10	11
TRAIN	Sensitivity	88.57	85.71	77.14
	Specificity	71.84	79.59	79.18
	logsens	0.3890	0.4267	0.3162
	AUC	0.8164	0.8858	0.8306
TEST	Sensitivity	90.90	81.82	77.27
	Specificity	75.16	81.70	80.39
	logsens	0.4540	0.3923	0.3258
	AUC	0.8333	0.8277	0.7897
VERIF	Sensitivity	89.50 ± 2.59	79.13 ± 3.02	68.30 ± 6.05
	Specificity	73.58 ± 1.05	78.47 ± 1.23	81.63 ± 1.04
	logsens	0.4191 ± 0.0372	0.3345 ± 0.0334	0.2466 ± 0.0526
	AUC	0.8354 ± 0.0146	0.8238 ± 0.0128	0.8292 ± 0.0209
Trial#2 (Tr2)	Nodes	5	10	11
TRAIN	Sensitivity	82.86	85.71	80.00
	Specificity	82.45	75.92	75.92
	logsens	0.4136	0.3916	0.3248
	AUC	0.8929	0.8110	0.8377
TEST	Sensitivity	86.36	90.91	86.36
	Specificity	77.12	69.94	72.54
	logsens	0.4114	0.3988	0.3695
	AUC	0.8437	0.8102	0.8265
VERIF	Sensitivity	79.13 ± 3.02	89.50 ± 2.59	68.30 ± 6.05
	Specificity	78.47 ± 1.23	76.48 ± 1.44	76.14 ± 1.07
	logsens	0.3345 ± 0.0334	0.4500 ± 0.0413	0.2217 ± 0.046
	AUC	0.8125 ± 0.0134	0.8118 ± 0.0140	0.8120 ± 0.0236

4.3.2 noPOSTOP ANN with Weight-elimination

The best performing classifiers for noPOSTOP using weight-elimination were Node 10 and 11 (as shown in Table 4.9). Node 10 had the best test sensitivity and specificity rela-

tionship for Tr1 - with a test logsens value of 0.3981, a sensitivity of 81.82 % , a specificity of 82.35% and an AUC of 0.8417. Yet, verification of this model proved to be different, as the logsens verification value dropped down to a mean of 0.1289 ± 0.0285 , the sensitivity to 52.28 ± 4.82 , the specificity to 81.87 ± 1.10 and the AUC to 0.8036 ± 0.0159 . Node 11 (Tr1) showed potential with its test logsens of 0.2369, resulting in a sensitivity of 72.73%, specificity of 72.55% and AUC of 0.7854. Verification showed a slightly higher mean sensitivity of 76.21 ± 3.64 , a drop in specificity of 67.21 ± 0.88 and AUC (0.7475 ± 0.0230). Tr2 showed poor discrimination for Node 10 (Tr2) with a value of 0.6940. Node 11 (Tr2) also decreased to a test AUC of 0.7556 and a verification of 0.7457 ± 0.0225 . Node 10 (Tr2) remained the same as the test AUC (0.6936 ± 0.0260). Node 11 (Tr2) test sensitivity was the same as the Node 11 (Tr1), however there was a slight decrease to the test specificity of Node 11 (Tr2). The verification results of Node 11 (Tr1 and Tr2) were similar to each other and to its corresponding test results. Therefore, this classifier was determined to be the best performing classifier and was chosen as the classifier to compare to the DT-ANN Hybrid.

Table 4.9: noPOSTOP Best Performing ANN with weight elimination (Node 10 and 11). Verification (VERIF) is the mean value and standard deviation of the 10 verification sets. TRAIN rate is 12.50% and TEST rate is 12.57%. Verification rate is 13.55 ± 0.84 .

Trial#1 (Tr1)	Nodes	10	11
TRAIN	Sensitivity	65.71	77.14
	Specificity	81.22	67.76
	logsens	0.2178	0.2478
	AUC	0.8376	0.7772
TEST	Sensitivity	81.82	72.73
	Specificity	82.35	72.55
	logsens	0.3981	0.2369
	AUC	0.8417	0.7854
VERIF	Sensitivity	52.28 ± 4.82	76.21 ± 3.64
	Specificity	81.87 ± 1.10	67.21 ± 0.88
	logsens	0.1289 ± 0.0285	0.2386 ± 0.0273
	AUC	0.8036 ± 0.0159	0.7475 ± 0.0230
Trial#2 (Tr2)	Nodes	10	11
TRAIN	Sensitivity	77.14	74.29
	Specificity	78.37	71.02
	logsens	0.3108	0.2419
	AUC	0.7939	0.7897
TEST	Sensitivity	59.09	72.73
	Specificity	79.74	67.97
	logsens	0.1635	0.2154
	AUC	0.6940	0.7556
VERIF	Sensitivity	61.56 ± 3.59	76.21 ± 3.64
	Specificity	75.75 ± 1.10	69.24 ± 0.95
	logsens	0.1691 ± 0.0249	0.2493 ± 0.0287
	AUC	0.6936 ± 0.0260	0.7457 ± 0.0225

4.3.3 Comparing noPOSTOP DT-ANN Hybrid and we_ANN

The best classifiers for the DT-ANN Hybrid and we_ANN are presented in Table 4.10. These classifiers were selected according to the reproducibility between Tr1 and Tr2, a high sensitivity and specificity, a high logsens value (balance between sensitivity and specificity) and high discrimination capability (AUC). The DT-ANN Hybrid showed

similar results among the nodes per Trial, as well as among nodes between Trials. The best performing DT-ANN Hybrid classifier was chosen as Node 5 of Tr1 (Table 4.8). For we_ANN, Node 11 (Tr2), was chosen as the best performing classifier. Comparison of the test results showed superior results compared to the we_ANN results. The DT-ANN Hybrid sensitivity (90.90%) exceeded that of the we_ANN (72.73%) by 18.17% ; the specificity by 7.19 % and the AUC by 0.0777. Both the DT-ANN Hybrid and we_ANN models selected for comparison verify well, with results that remain close to the test results. It is evident that the DT-ANN Hybrid exceeds the performance of the we_ANN results.

Table 4.10: Comparing the best performing classifiers for noPOSTOP between the DT-ANN Hybrid and we_ANN.

Node		5	11
Trial#		1	2
ANN		DT-ANN Hybrid	we_ANN
TEST	Sensitivity	90.90	72.73
	Specificity	75.16	67.97
	logsens	0.4540	0.2154
	AUC	0.8333	0.7556
VERIF	Sensitivity	89.50 $\pm$ 2.59	76.21 $\pm$ 3.64
	Specificity	73.58 $\pm$ 1.05	69.24 $\pm$ 0.95
	logsens	0.4191 $\pm$ 0.0372	0.2493 $\pm$ 0.0287
	AUC	0.8354 $\pm$ 0.0146	0.7457 $\pm$ 0.0225

4.3.4 Hybrid Classifier Results and Number of Nodes

Figures 4.1 and Figure 4.2 is a visual presentation the number of hidden nodes plotted against the logsens values of the trials which produced the best performing classifiers. For the POSTOP DT-ANN Hybrid this is Tr1 and the noPOSTOP DT-ANN Hybrid classifiers this is Tr2. The results are plotted from hidden node 3 to 11. The ANN-RFW automated search for optimal nodes was terminated at 11 due to the Kolmogorov superposition theory applied to the context of ANN. The best performing nodes is contained in $2*n+1$ nodes; where n is the number of input variables [118]. It is observed that the test

results fall outside of the verification range. It is also noticed that the results depend on the number of hidden nodes.⁵⁴

Figure 4.1: POSTOP Hybrid Tr1 - Plot of the number of hidden nodes against the logsens value. The red points indicate the test results and the blue points indicate the verification results with the standard deviation error bars.

Figure 4.2: noPOSTOP Hybrid Tr2 - Plot of the number of hidden nodes against the logsens value. The red points indicate the test results and the blue points indicate the verification results with the standard deviation error bars.

⁵⁴For final classifier weights and biases see Appendix D

4.4 Assessing PPV and NPV for Predicting Mortality

4.4.1 POSTOP Hybrid PPV and NPV

Table 4.11 shows the PPV and NPV for the best Hybrid classifiers for POSTOP cases. All test NPV were extremely high, ranging from 99.07% to 100.00%. The verification NPVs all had the same mean and std (99.44 ± 0.01), with an exception for Node 11 (Tr2) where the NPV was 100.00%. Though Node 11 (Tr2) had the greatest sensitivity (Table 4.5), it compensated with the lowest specificity among each node; therefore the test PPV (16.33%) was the lowest among the four classifiers. It also had the lowest verification PPV (18.90 ± 2.12). Node 6 (Tr2) had the greatest test sensitivity (24.00%) due to the highest specificity. This was followed by Node 11 (Tr1) with a test PPV of 19.36% and Node 6 (Tr1) with test PPV of 18.75%. Node 6 (Tr1 and Tr2) and Node 11 (Tr1) all had the same verification PPV range (22.06 ± 2.84). All train PPVs were greater than the test and verification PPV because of the *a priori* re-sampling rate (to 5.06%).

Table 4.11: POSTOP Hybrid PPV and NPV. (A) TRAIN rate of 5.06% (B) TEST rate of 3.24% (C) VERIFICATION rate of 3.04 ± 0.38 (mean and standard deviation).

Node		Trial #1		Trial #2	
		6	11	6	11
PPV	A	34.09	34.09	28.00	24.36
	B	18.75	19.36	24.00	16.33
	C	22.06± 2.84	22.06± 2.84	22.06± 2.84	18.90± 2.12
NPV	A	98.38	98.38	98.08	99.41
	B	99.07	99.07	99.10	100.00
	C	99.44± 0.01	99.44± 0.01	99.44± 0.01	100.00±0.00

4.4.2 POSTOP we_ANN PPV and NPV

The test and verification NPV of the POSTOP we_ANN POSTOP classifiers were very high (greater than 97.00%). Our interest was in the PPVs, found in Table 4.12. First it was evident that there was a large ambiguity in Node 6 (Tr1) and Node 5 (Tr2). Though

these nodes have high test PPVs, they failed to reproduce the same PPV in the verification set. Due to the sensitivity and specificity results from the we_ANN, this was expected. Between Node 5 (Tr1) and Node 11 (Tr1 and Tr2), the highest test value was 7.55 % which was well below the POSTOP Hybrid's poorest PPV (Table 4.11) of 16.33% (POSTOP Node 11 - Tr2). According to the results in Table 4.12 the best PPV was found in Node 6 (Tr2) with a test PPV of 66.66% and verification of 42.14 ± 6.83 . Unfortunately, a closer look reveals that this is due to the high test specificity (99.16%) accompanied with a low sensitivity (50.00%) as shown in Table 4.6.

Table 4.12: POSTOP we_ANN PPV and NPV. (A) TRAIN rate of 5.06% (B) TEST rate of 3.24% (C) VERIFICATION rate of 2.71 ± 0.23 (mean and standard deviation).

Node	Trial #1			Trial #2		
	5	6	11	5	6	11
PPV	A	11.93	100.00	16.67	25.00	100.00
	B	7.55	50.00	6.12	41.67	66.66
	C	7.82 ± 0.74	35.47 ± 5.51	8.07 ± 1.07	9.16 ± 0.86	42.14 ± 6.83
NPV	A	97.39	98.75	97.63	95.53	99.24
	B	97.94	97.93	97.47	98.72	98.34
	C	99.36 ± 0.22	98.54 ± 0.21	98.83 ± 0.27	97.66 ± 0.25	98.60 ± 0.24

4.4.3 noPOSTOP Hybrid PPV and NPV

Table 4.13 shows that all the hybrid classifiers have a high NPV. The highest NPVs were found in Node 5 (Tr1) and Node 10 (Tr2) which had a test NPV of 98.29% and 98.17%, respectively. The lowest test NPV was Node 10 (Tr1) at 96.90%. High NPV verification was noticed in both classifiers. The greatest test PPV was found in Node 10 (Tr1) and Node 5 (Tr2) which had a test PPV of 39.13% and 35.19%, respectively. Node 10 (Tr1) and Node 5 (Tr2) both had the same PPV verification range (34.37 ± 3.23). Node 5's (Tr2) test PPV was much closer to the mean PPV found in the verification results compared to Node 10 (Tr1). Node 11 had the lowest PPV verification results of 29.02 ± 3.98 . Although Node 10 (Tr2) had the lowest test PPV (30.30%), its verification mean was higher

(35.17±3.35).

Table 4.13: noPOSTOP Hybrid PPV and NPV. (A) TRAIN rate of 12.50% (B) TEST rate of 12.57% (C) VERIFICATION rate of 12.47±1.15 (mean and standard deviation).

Node	Trial #1			Trial #2		
	5	10	11	5	10	11
PPV	A	31.00	37.50	34.62	40.28	33.71
	B	34.48	39.13	36.17	35.19	30.30
	C	32.55±2.95	34.37±3.23	34.65±4.67	34.37±3.23	35.17±3.35
NPV	A	97.78	97.50	96.04	97.12	97.38
	B	98.29	96.90	96.09	97.52	98.17
	C	98.01±0.46	96.35±0.63	94.80±0.78	96.35±0.63	98.09±0.44

4.4.4 noPOSTOP we_ANN PPV and NPV

The NPV values found in Table 4.14 were lower than that of the NPV of the Hybrid classifiers. The greatest test NPV was found in Node 10 - Tr1 (96.92%) while the lowest test NPV was 92.72% (Node 5 - Tr2). Node 6 (Tr1 and Tr2) both report a test PPV greater than 40.00%; however the sensitivity is only at 50.00% (Table 4.9). Excluding these nodes, the highest PPV values was found in Node 10 (Tr1), which had a PPV of 40.00% but a decrease in its verification PPV (31.15±3.43). Validation of these results indicates a drop in the test PPV of Node 10 (Tr2) to 29.55%. The best PPV values were found in both Node 11 (Tr1) and Node 11 (Tr2) test PPV, which have a value of 27.59% and 24.62%, respectively. Both classifiers verified well at 26.71 ±2.04 (Node 11 - Tr1) and 27.97± 2.05 (Node 11 - Tr2).

Table 4.14: noPOSTOP we_ANN PPV and NPV. A) TRAIN rate of 12.50% (B) TEST rate of 12.57% (C) VERIFICATION rate of 13.55 ± 0.84 (mean and standard deviation).

Node	Trial #1			Trial #2			
	6	10	11	6	10	11	
PPV	A	39.58	33.33	25.47	96.97	33.75	26.80
	B	46.15	40.00	27.59	45.83	29.55	24.62
	C	45.34±2.02	31.15±3.43	26.71±2.04	56.49±3.53	28.48±2.31	27.97± 2.05
NPV	A	93.10	94.31	95.40	98.79	96.00	95.08
	B	93.29	96.92	94.87	92.72	93.13	94.55
	C	93.93±0.66	91.63±0.87	94.75±0.75	93.80±0.79	92.63±0.78	94.90± 0.72

4.5 Prediction of High-Risk Preterm Birth

4.5.1 Decision Tree Results

Three Decision Trees (DTs) were created, one for each of the bootstrapped training set (b_train). As shown in Table 4.15 and Table 4.16, each of the trees used different variables to classify the data, resulting in varying test sensitivity and specificity. For the PAROUS decision tree, each tree was made with a cost of misclassifying positive cases set at a value of 6.5. The NULLIPAROUS DTs were set at a misclassification cost of 6.0. This cost allowed the decision tree to emphasize classification of positive cases, thus favouring the sensitivity as seen in the test results. For the PAROUS database, the first tree used 20 variables, while the other two used 19 variables out of the 69 input variables. For the NULLIPAROUS database, DT 1 used 16 variables, DT2 used 15 and DT3 used 22 variables. Though each tree used different variables to classify the data; there were some variables that appeared in each tree, as highlighted in gray. In the PAROUS database, ten variables were found in each of the classifier: plurality; morbidity: placenta, bleeding, cervix; high blood pressure; maternal race: black and hawaii; previous preterm births, previous low birthweight and maternal age. Certain variables were not applicable to the NULLIPAROUS database, specifically previous preterm births, previous lowbirthweights, previous live births and parity (the number of live births). In the NULLIPAROUS database,

eleven variables were consistently used for each tree: plurality; morbidity: placenta, bleeding, cervix; high blood pressure; maternal race: black, hawaii, pacific; maternal age and mothers who have smoked in the last two years. Eight of these variables were the same as the ones used to create the PAROUS dataset. The PAROUS dataset has test sensitivities that varied among each tree. DT1, DT2 and DT3 have a sensitivity of 80.0%, 76.7% and 79.3%, respectively. The specificity for DT 1 (69.0%), DT 2 (72.8%) and DT 3 (69.7%) vary as well. For the PAROUS DTs, the tree with the best compromise between sensitivity and specificity was DT2, with a test sensitivity of 76.7% and specificity of 72.8%. The variables found in the attribute usage list for DT2 were used to make the attribute intervals. For the NULLIPAROUS database, the difference among the test sensitivity among each tree was not as great as the PAROUS dataset (Table 4.16). Since there was a small difference between the different trees in terms of sensitivity and specificity, DT1 was used as the tree used to make the attribute intervals as it contained less attributes than DT 3.

Attribute Intervals

The attribute intervals were created using the splitting values found in PAROUS DT2 and NULLIPAROUS DT1. The threshold value indicates the splitting value used by the decision tree. If there was more than one splitting value present, then a new variable was created as shown in Table A.5 and Table A.6. For the PAROUS database, there were 26 input variables used to create an ANN classifier; while the NULLIPAROUS database had 27 input variables. The modified input datasets (attribute interval) were labelled as the *h_Parous* (Table A.5) and *h_Nulliparous* (Table A.6) for the appropriate database. This database was distributed as outlined in Section 3.9.5 (page 62) to create the ANN classifiers.

Table 4.15: PAROUS decision tree results and list of attribute usage. (A) Sensitivity (B) Specificity (C) CCR

Decision Tree #		1	2	3
Train	A	82.1	78.3	81.7
	B	69.4	73.0	69.8
	C	71.4	73.8	71.7
Test	A	80.0	76.7	79.3
	B	69.0	72.8	69.7
	C	70.7	73.4	71.3
Attribute Usage	1	100% PLURAL	100% PLURAL	100% PLURAL
	2	95% MORB_PLA	95% MORB_PLA	95% MORB_PLA
	3	87% PREV_PRE	87% PREV_PRE	87% PREV_PRE
	4	82% MORB_BLD	87% MM_HBP	74% MM_HBP
	5	78% BLACK	74% MORB_BLD	72% MORB_BLD
	6	74% MM_HBP	74% MORB_CRV	70% BLACK
	7	70% MORB_CRV	73% BLACK	70% MORB_CRV
	8	23% HAWAII	22% HAWAII	26% PREV_LBW
	9	20% PREV_LBW	19% PREV_LBW	17% MOM_CIG_later
	10	15% MAT_AGE	18% MAT_AGE	14% MM_NOMD
	11	12% MM_NOMD	15% MOM_CIG_later	13% HAWAII
	12	9% MOM_CIG_later	9% DRK4_3L	12% MAT_AGE
	13	7% SMK_3Later	8% BMI	9% MOM_CIG_before
	14	6% PAD_COM	8% STRS_DV3	6% MORB_CAR
	15	5% OTH_TERM	7% MOMSMOKE	5% MOMCIG
	16	5% STRS_PG	6% STRS_JL3	4% BMI
	17	4% ERLY_PNC	5% STRS_BIL	3% OTH_TERM
	18	3% MOM_CIG_before	4% MM_DIAB	2% MORB_DIA
	19	2% STRS_DV3	4% STRS_JOB	2% MARRIED
	20	1% MORB_DIA		

Table 4.16: NULLIPAROUS decision tree results and list of attribute usage.

Decision Tree#		1	2	3
Train	Sensitivity	78.61	77.73	78.40
	Specificity	57.96	58.86	59.65
	CCR	61.5	62.23	62.99
Test	Sensitivity	75.44	74.51	76.57
	Specificity	56.87	58.11	58.03
	CCR	60.21	61.06	61.69
Attribute Usage	1	100% PLURAL	100% PLURAL	100% PLURAL
	2	97% HAWAII	97% HAWAII	97% HAWAII
	3	94% MM_HBP	94% MORB_PLA	94% MORB_PLA
	4	84% MORB_PLA	90% MM_HBP	90% MM_HBP
	5	78% MORB_CRV	78% PACIFIC	78% MORB_BLD
	6	76% MORB_BLD	78% MORB_BLD	66% MORB_CRV
	7	65% BLACK	63% MORB_CRV	65% BLACK
	8	53% MM_NOMD	62% BLACK	53% MM_NOMD
	9	14% MAT_AGE	51% MM_NOMD	39% BMI
	10	12% PACIFIC	13% MAT_AGE	22% MAT_AGE
	11	8% INCOME	4% STRS_WRK	12% PACIFIC
	12	5% STRS_WRK	4% STRS_JOB	12% MOMSMOKE
	13	4% STRS_JOB	4% MORB_KID	11% MARRIED
	14	4% MORB_KID	3% SMK_2YRS	9% MORB_DIA
	15	4% OTH_TERM	3% OTH_TERM	8% MORB_CAR
	16	3% SMK_2YRS		7% PREG_TRY
	17			6% PNC_MTH
	18			4% STRS_WRK
	19			4% DRK43L_A
	20			4% ERLY_PNC
	21			3% SMK_2YRS
	22			2% SEX

4.5.2 HR-PTB Hybrid Classifier Results and Number of Nodes

Figures 4.3 and Figure 4.4 is a visual presentation the number of hidden nodes plotted against the logsens values. It was observed that results remained fairly consistent among each hidden nodes 3 to 11 - except for hidden node 10 for Parous cases and hidden node 6 for the Nulliparous cases. The logsens results were much lower indicating a low sensitivity/specificity tradeoff. Since ANN training is computational expensive and large improvements were not seen; the ANN-RFW automated search for optimal nodes was not

set to Kolmogorov superposition theory ($2*n+1$ nodes; where n is the number of input variables) [118]. It was observed that the test results fell within the verification range except for Node 4 (Parous). The test logsens fell slightly out of the verification range. It was also noticed that the results depend on the number of hidden nodes.

Figure 4.3: Parous Hybrid logsens Plot - Plot of the number of hidden nodes against the logsens value. The red points indicate the test results and the blue points indicate the verification results with the standard deviation error bars.

Figure 4.4: Nulliparous Hybrid logsens Plot - Plot of the number of hidden nodes against the logsens value. The red points indicate the test results and the blue points indicate the verification results with the standard deviation error bars.

4.5.3 Hybrid Classification ANN Results

Table 4.17 shows the best performing classifiers created using the classification ANN for PAROUS cases. All the test results have sensitivities greater than 65.0%. Node 4, 8 and 9 had a specificity > 80%; however Node 11 fell short by 3.07%. Each classifier had similar AUC values of around 0.800 which indicated excellent discrimination. The test results were accompanied by excellent verification results that closely matched the test results. The NULLIPAROUS results (Table 4.18), best performing classifiers were Node 5 and Node 11. Though Node 5 had a test sensitivity of 65.03%, its specificity was at 71.25%. Its AUC (0.7127) was substantially lower compared to the PAROUS AUC values which remained around 0.800. Node 11 had a lower test sensitivity of 61.33%, with a slightly higher specificity (73.64%) and AUC (0.7237) compared to Node 5. All classifiers used to predict HR-PTB in the PAROUS and NULLIPAROUS cases verified well. All test results fell within the range given by the verification results.⁵⁵

⁵⁵For all classifier results see Appendix E

Table 4.17: PAROUS Best Performing Hybrid ANN with WE (Node 4,8, 9 and 11). Verification (VERIF) is the mean and standard deviation of the 10 verification sets. TRAIN prevalence is 16.13% and TEST prevalence is 16.08%. Validation Rate 16.13± 0.1391

	Nodes	4	8	9	11
TRAIN	Sn	66.04	68.52	71.09	73.76
	Sp	83.33	83.20	82.13	77.60
	LS	0.2292	0.2512	0.2707	0.2722
	AUC	0.8055	0.8161	0.8300	0.8138
TEST	Sn	66.34	65.14	67.93	72.51
	Sp	83.92	83.88	81.17	76.93
	LS	0.2343	0.2236	0.2365	0.2570
	AUC	0.7949	0.8077	0.8062	0.8047
VERIF	Sn	61.42 ±4.52	61.80 ±4.50	64.43 ±4.84	69.92± 3.44
	Sp	83.31±2.15	82.96 ±2.51	81.05 ±2.33	77.43± 1.66
	LS	0.1932±0.0340	0.1944±0.0255	0.2082±0.0339	0.2378±0.0287
	AUC	0.7906±0.071	0.7970± 0.0256	0.7983± 0.0112	0.8001 ±0.0214

Sn=sensitivity, Sp = specificity, LS=logsens, AUC = area under the curve

Table 4.18: NULLIPAROUS Best Performing Hybrid Classification (C) ANN with WE (Node 5 and 11). Verification (VERIF) is the mean and standard deviation of the 10 verification sets. TRAIN prevalence is 17.86% and TEST prevalence is 17.40%. Validation Rate 17.72 ± 2.4521.

	Nodes	5	11
TRAIN	Sn	62.77	60.24
	Sp	71.71	74.12
	LS	0.1631	0.1548
	AUC	0.7201	0.7213
TEST	Sn	65.03	61.33
	Sp	71.25	73.64
	LS	0.1757	0.1600
	AUC	0.7127	0.7237
VERIF	Sn	65.51± 4.77	63.54 ± 5.28
	Sp	71.14± 2.73	74.24 ± 2.27
	LS	0.1799 ± 0.0311	0.1786 ± 0.0363
	AUC	0.7310 ± 0.0301	0.7358 ± 0.0341

Sn=sensitivity, Sp = specificity, LS=logsens, AUC = area under the curve

4.5.4 Risk Stratification Results

PAROUS Results The final ANN parameters reached in the Classification (C) ANN of each of the best performing classifiers mentioned in Section 4.5.3 are given for the DT-ANN Hybrid used to classify PAROUS cases (Table A.8) and NULLIPAROUS cases (Table A.9). These final parameter values were used to create a Risk Stratification (RS) ANN, where results are presented in Table 4.19 (PAROUS) and Table 4.20 (NULLIPAROUS). A comparison between the PAROUS Hybrid Classification (C) ANN and Hybrid Risk Stratification (RS) ANN test results (Table 4.21), showed that Node 8 (RS) was the only classifier that validated well compared its classification (C) performance indices. This is highlighted in gray in the table. Node 9 (RS) had a decreased in sensitivity (54.98%) compared to that of Node 9 (C) which had a sensitivity of 67.93%. The specificity increased from 81.17% (C) to 88.58% (RS). Node 11 (RS) had an increased in sensitivity (79.68%) compared to Node 11 (C) which had a sensitivity of 72.51%. However, it was accompanied by a decreased specificity (66.81%). Node 4 (RS) had the largest change - a decreased sensitivity (23.50%) and increased specificity (96.25%) compared to Node 4 (C). All of the RS AUC values had a very small improvement except for Node 11, which had a slight decrease. The highest AUC improvement was in Node 4 which increased from 0.7949 (C) to 0.8195 (RS). Cut-off values were tested for Node 4, 8, 9 and 11 at 0.215, 0.4859, 0.390 and 0.655 respectively. Application of the cut-off value improved the Risk Stratification results for all PAROUS classifiers as shown in Table 4.21. The best outcome in Node 4 (bold font) with the test sensitivity of 65.13% and 84.07% specificity; however all other classifiers had results that were close these values⁵⁶.

NULLIPAROUS Results Interestingly, the performance indices for the NULLIPAROUS RS ANN test results indicated a sensitivity and specificity value far from the Classification ANN test results (Table 4.22). Node 5 (C) had a drop in its sensitivity from 65.03%

⁵⁶For RS classifier biases and weights see Appendix E

to 16.75% (RS). Node 11 (C) also decreased from 73.64% (C) to 29.06% (RS). Both decreased sensitivity were accompanied by an increased specificity. Node 11(RS) had a specificity of 96.4% and Node 11 (RS) had a specificity of 93.85%. The AUC of the RS ANN remained close to the Classification AUC. An adjustment of the cut-off value of Node 5 and 11, 0.1800 and 0.3020 respectively, produced a higher sensitivity however, were still slightly lower than the Classification ANN results.

Table 4.19: PAROUS Best Performing Hybrid Risk Stratification ANN with WE (Node 4,8, 9 and 11). TRAIN prevalence is 16.13% and TEST prevalence is 16.08%

Nodes		4	8	9	11
TRAIN	Sensitivity	28.41	66.44	58.9	80.69
	Specificity	96.66	82.84	87.87	67.21
	logsens	0.03943	0.2306	0.18652	0.2768
	AUC	0.81518	0.81099	0.81935	0.80442
TEST	Sensitivity	25.30	64.94	54.98	79.681
	Specificity	96.25	82.93	88.58	66.81
	logsens	0	0.2181	0.15938	0.26299
	AUC	0.8195.	0.8100	0.8143	0.8014

Table 4.20: NULLIPAROUS Best Performing Hybrid Risk Stratification ANN with WE (Node 5 and 11). TRAIN prevalence is 17.86% and TEST prevalence is 17.40%.

Nodes		5	11
TRAIN	Sensitivity	17.7	30.6
	Specificity	97.512	93.84
	logsens	0.014583	0.04493
	AUC	0.738	0.7370
TEST	Sensitivity	16.75	29.06
	Specificity	96.4	93.85
	logsens	0	0
	AUC	0.7195	0.7210

Table 4.21: PAROUS Test results comparing Hybrid Classification (C) ANN, Risk Stratification (RS) ANN and adjustment of the Cut-off value for the Risk Stratification ANN (Node 4, 8, 9 and 11).

Nodes		4	8	9	11
Classification (C) ANN	Sensitivity	66.34	65.14	67.93	72.51
	Specificity	83.92	83.88	81.17	76.93
	logsens	0.2343	0.2236	0.2365	0.2570
	AUC	0.7949	0.8077	0.8062	0.8047
Risk Stratification (RS) ANN	Sensitivity	25.30	64.94	54.98	79.681
	Specificity	96.25	82.93	88.58	66.81
	logsens	0	0.2181	0.15938	0.26299
	AUC	0.8195	0.8100	0.8143	0.8014
Risk Stratification (RS) Cut-off Adjustment	Threshold	0.215	0.4859	0.390	0.655
	Sensitivity	65.13	65.5	65.14	65.9
	Specificity	84.07	82.8	82.08	81.7

Table 4.22: NULLIPAROUS Comparing Classification (C) ANN with Risk Stratification (RS) ANN (Node 5 and 11) TEST results.

Nodes		5	11
Classification (C) ANN	Sensitivity	65.03	61.33
	Specificity	71.25	73.64
	logsens	0.1757	0.1600
	AUC	0.7127	0.7237
Risk Stratification (RS) ANN	Sensitivity	16.75	29.06
	Specificity	96.45	93.85
	logsens	0	0
	AUC	0.7195	0.7210
Risk Stratification (RS) Cut-off Adjustment	Threshold	0.1800	0.302
	Sensitivity	61.08	60.83
	Specificity	71.14	72.24

4.6 PPV and NPV for HR-PTB

The PPV and NPV from the Hybrid Risk Stratification (RS) ANN are shown in Table 4.23. The PPVs for all Parous cases were greater than 40.00%, where the highest value was 43.93% (Node 4). This node also had the highest NPV value of 92.64%, though

all other classifiers only differed by a few hundredth of a percent. The highest Nulliparous PPV was found in Node 11, with a value of 31.58%. This was 12.35% lower than the best Parous classifier. The best Nulliparous NPV (89.75%) only had a 2.89% difference from the Parous NPV.

Table 4.23: HR-PTB PPV and NPV from Hybrid Risk Stratification (RS) Test Results. Test prevalence was 16.08% (Parous) and 17.40% (Nulliparous).

Node	<i>Parous</i>				<i>Nulliparous</i>	
	4	8	9	11	5	11
PPV	43.93	42.19	41.06	40.83	30.84	31.58
NPV	92.64	92.61	92.47	92.59	89.67	89.75

4.7 Discussion - Mortality Prediction

4.7.1 Predicting mortality in an Adult ICU

Comparing the DT-ANN Hybrid and the we_ANN Results

The 17-variable we_ANN model was able to predict mortality in postoperative and non-postoperative adult ICU cases. Node 11 was classified as the best system for both *POSTOP* and *noPOSTOP*. This is due to its reproducibility of the results between the two trials as well as the ability to verify as shown in the verification results. The ANN with weight-elimination (*POSTOP* Tr1) had acceptable results: a discrimination of 0.7995 ± 0.0272 , sensitivity of 66.67 ± 7.70 and specificity of 78.81 ± 1.05 . Though these results were acceptable, a comparison with the DT-ANN Hybrid showed a substantial increase in all results: Node 11 (Tr1) had a high sensitivity (83.38 ± 2.09), specificity (90.75 ± 0.56) and AUC (0.9586 ± 0.0026) value that surpassed the we_ANN results. The *POSTOP* DT-ANN Hybrid outperformed the we_ANN model. The prominent improvement of the results showed promising use for the DT-ANN Hybrid to predict mortality in an adult ICU.

In terms of the *noPOSTOP* results, the *we_ANN* results also had acceptable verification results in terms of the AUC (0.7475 ± 0.0230), sensitivity (76.21 ± 3.64) and specificity (67.21 ± 0.88) values. In contrast, the best performing DT-ANN Hybrid (Node 5) had almost a 10% increase to the discrimination capability of the *we_ANN* (0.8354 ± 0.0146), a 13% increase to the sensitivity (89.50 ± 2.59) and a 6.37% increase to the specificity (73.58 ± 1.05). The sensitivity and AUC of the DT-ANN Hybrid was seen to have a prominent increase within POSTOP and noPOSTOP cases; thus favouring the use of DT-ANN Hybrids. In addition, the robustness and consistency of the DT-ANN Hybrids were evident when compared to the corresponding node found in *we_ANN* between the two trials. Due to the re-aggregation and re-division of the dataset to run the second trial, it is logical that the training and testing points found in first trial would differ from that of the second trial. The DT-ANN Hybrid seemed to have the ability to deal with the different changes in the data points, as robustness of this was evident in the second trial; both trials had similar results for its corresponding node. This was not the case for the ANN with weight-elimination. Results for this method not only showed large differences between Tr1 and Tr2; but in some cases, classifiers that performed well in the first trial (Node 5 *we_ANN*), created a classifier that was incapable of predicting the cases presented to it in the second trial (Section 4.2.2). An unusual observation was found in the POSTOP *we_ANN* classifiers (Node 6 and 11 (Tr1)) as shown in Table 4.6. A substantial increase was noticed in its verification set compared to that of the test results in terms of sensitivity and AUC. Although it is good to see acceptable verification on previously unseen data, such a magnitude of difference between the test and verification results may be indicative of an unreliable classifier. For example, the POSTOP results were observed to have a large increase from the test results to the verification results (Table 4.6). In the hybrid classifier, an increase was also seen between trials. For example, Node 11 (Tr2) (Table 4.5) had a test sensitivity of 100.00%, a 25.00% increase. This large increase can be attributed to the limited amount of mortality cases found in this dataset where a few correct positive predic-

tions can have a large effect on the percentage. As mentioned in the methodology (Table 3.6), there were only 8 cases in the test set where 6/8 is a 75.00% sensitivity. This shoots up to 100.00% for 2 more correctly predicted positive cases. It is the same for the verification set where there are only 6 cases. Therefore correctly classifying 5/6 cases would give a sensitivity of 83.3% while correctly classifying 6/6 cases would increase the sensitivity to 100%.

4.7.2 Comparing the DT-ANN Hybrid and we_ANN Method

The DT-ANN Hybrid method exceeded the performance of we_ANN. As mentioned in the Literature Review, there are many methods devised to improve the ANN using weight-elimination: variable weight reduction, re-sampling the *a priori* prevalence of rare outcomes before training and adjusting the weight-decay and weight-elimination starting parameters. The we_ANN required parameter tuning after training occurs. This is a time-consuming process as another training run is required before improvements (or degradations) could be seen. The DT-ANN Hybrid did not require weight-elimination to achieve better results than the we_ANN, and only required adjustment of the learning rate (which was set at the minimum value). Not only did the DT-ANN Hybrid outperform the we_ANN, it also required fewer adjustments of starting parameters. Adjustment for optimal performance occurred in two steps - one at the DT portion (adjustment of cost and global pruning) and the second at the ANN portion (adjustment of learning rate and weights). The DT classifier was able to select the significant variables, where classification took less than 0.1 seconds. ANN classification using the DT-ANN Hybrid had high performance at a training rate of only 5.06% from the original 3.37%. This is a much lower re-sampling rate compared to the 20% as suggested by Ennett and Frize (2000) explained in Section 2.8.2. Consequently, reasons for poorer performance of the we_ANN could be attributed to the classifiers inability to identify mortality cases (due to the rare occurrence of these positive outcomes). Additionally, weight-elimination is supposed to effectively

degrade the weights to zero, essentially pruning the network Section F.1.6; however the improved performance using the DT-ANN Hybrid method suggests that prior feature reduction may be needed before ANN weight-elimination is applied.

Best Performing Nodes

The ANN-RFW is an automated program used to determine the optimal number of nodes found in the hidden layer. It is observed that the best performing classifiers for the prediction of the medical outcomes (mortality and high-risk preterm birth) have an ANN architecture that is affected by the number of nodes found in the hidden layer as shown in Figures 4.1 and figure 4.2 (Page 77). The logsens values are shown indicating the relationship between the sensitivity and specificity. It is observed that many of the log-sens tests were not included in the verification range. This can be attributed to the large difference found between the test sensitivity and verification sensitivity (due to the limited amount of positive cases). These figures exhibit how the predictive ability of the ANN classifier is governed by the complexity of the ANN classifier (the number of hidden nodes) [18]. However, the reason behind why the best performing nodes determined performed better compared to the other nodes remains unknown in this work. Such varying results per number of nodes would be a point of interest to examine in the future.

Assessing Diagnosis Accuracy of Mortality using PPV and NPV

As mentioned in earlier chapters, the PPV is the proportion of correctly diagnosed patients with the disease that are identified as positive (Section 3.5.7). The NPV is the proportion of correctly diagnosed patients without the disease out of the patients diagnosed as negative (Section 3.5.8). These are used as direct diagnosis accuracy of a classifier, which is a comprehensive test for physicians. An examination of the equations to calculate these predictive values from the sensitivity and specificity reveal that these values are directly related to the prevalence. A high sensitivity and specificity will not give a high PPV if the prevalence is low. A rarer rate of the disease will result in more assurance that a negative

test indicates no abnormality; while less assurance is placed on a positive test that will indicate abnormality. The PPV and NPV are the posterior probability that a patient will have the disease. The usefulness of the test can be applied by comparing the prior probability (prevalence) with that of the posterior prevalence (PPV) [110]. It is evident from results in Section 4.4, that high PPV results do not necessarily indicate a good test. This is shown with numerous classifiers that have a high PPV, but a low sensitivity. Some of the we_ANN classifiers were not useful due to a low sensitivity and specificity balance and AUC (for example Node 6 noPOSTOP Tr1 and Tr2); however, these classifiers have higher test PPVs (Table 4.14) compared to the useful classifiers. According to these results, it is apparent that diagnostic testing should not rely on the PPV and NPV alone. Although, they give a direct diagnostic accuracy of the classifier, they are subjected to poor sensitivity values and varying prevalence. The contrast between the DT-ANN Hybrid and we_ANN reinforces that the DT-ANN Hybrid system is superior to the we_ANN alone. There is consistency of the PPV and NPV among each classifiers (the we_ANN values have large fluctuations of the PPV) along each node as well as between the first and second trial. In addition, the DT-ANN Hybrid classifiers have greater PPVs and NPVs than that of valid we_ANN classifiers.

Mortality Prediction

The DT-ANN Hybrid classifier has advantages of predicting mortality in an adult ICU compared to current standard mortality prediction systems (as given in Section 2.1 and 2.2.1). Such advantages are as follows: (1) the limited amount of variables that need to be collected (five for both postoperative and non-postoperative cases), compared to 7 to 142 of other clinical mortality prediction systems (Table 2.1); (2) excellent to outstanding discrimination capabilities; and (3) a high specificity followed with a high sensitivity (whereas many standard scoring systems have a high specificity and low sensitivity). For example, a review of standard prediction scoring system by Sinuff (2006) showed the best

specificity to sensitivity as 86% and 67%, respectively, while most of the other systems had 94% (specificity) and 58% (sensitivity). Important predictor variables are emphasized in postoperative and non-postoperative cases. Variables used for mortality prediction in postoperative cases included three APACHE variables (PH, Temperature, and Partial Pressure of Oxygen in the Blood) and 2 Demographic variables (Age and Emergency Surgery). Variables used for mortality prediction in non-postoperative cases included four APACHE variables (Glasgow Coma Score, Fraction of Inspired Oxygen, Mean Arterial Pressure, and Hematocrit) and one demographic variable (Age). Identifying these contributing variables may be useful in patient intervention and resource management. The developed integrated hybrid system presents very promising results; however caution must be taken when applying to other ICU environments as the demographics, equipment maintenance and medical staff may differ, and may ultimately affect the patient outcome.

4.8 Discussion - HR-PTB Prediction

Parous results' using the DT-ANN Hybrid was able to achieve the same sensitivity and specificity as the primary research outlined in Catley's thesis [15]. An integration of Decision Trees, Classification ANN with weight elimination and Risk Stratification gave a model with a 65.13% test sensitivity and 84.07% specificity (Node4). Other nodes (Nodes 8, 9 11) used to classify the Parous cases, also gave similar results; however, Node 4 had the highest sensitivity and specificity balance of a 65.13% and 84.07%(Table 4.21). All classifiers met the clinical expectation set out by Dr. Walker (Section 3.1) as they all had a sensitivity above 65% and a specificity above 80%. Though the sensitivity and specificity did not exceed that of past work (which may be due to the different data processing used in this thesis); the AUC did increase. Catley achieved an AUC of almost 0.800; whereas this integrated DT-ANN Hybrid system had a test AUC of 0.8195. Furthermore, other aspects make this prediction model ideal. First, this ensemble integrated system is a Generalized North American Model. As shown in Table 4.15, none of these features were specific to the

US, thus the model may be applicable to international facilities. Primary work [15] needed 48 variables (including US specific variables) in order to maintain sufficient specificity and sensitivity. Second, the number of variables needed to input into the DT-ANN Hybrid was greatly reduced from 48. This classifier required only 19 variables to be collected by the user (Table 4.15) for the women who have already had children. Reduction in variable input is an ideal factor when designing a decision support system. In addition, the Risk Stratification ANN model not only provides an outcome for individual cases, but also provides a risk level for each case. Therefore, the user may adjust the cutoff point, thus providing a obstetrical screening tool that allows emphasize on the prediction of negative (higher specificity) or positive cases (higher sensitivity). As indicated in the results (Section 4.5.3), the classifier for the Nulliparous database did not perform as well as the Parous database. A reduction in its sensitivity, specificity and AUC were noticed and were all highly expected. In Catley's PhD thesis, she points out that the variable 'previous preterm birth' was the third best indicator of preterm birth, following 'maternal bleeding' and 'plurality'. This study coincides with these findings, as all decision trees used 'maternal bleeding' and 'plurality' as consistent and high contributing factors among each DT classifier. The Parous variables also used 'previous preterm birth' as one of these strong contributors; however, for the nulliparous cases, this variable was not applicable. Another variable that ranked high on the attribute list in the Parous DTs was 'previous low birthweight', which again was not applicable in the Nulliparous database. 'Previous preterm birth' and 'previous lowbirth-weight' are ranked high on the attribute list of Parous cases and are very strong indicators in the prediction of HR-PTB - therefore the poorer accuracy and discrimination ability of the Nulliparous model could be attributed to this factor. As previously mentioned, past work by Catley [15] focused on valid values for the variable 'previous preterm birth', thus her work is predisposed for better prediction. Therefore, this is the first attempt at creating a prediction model for Nulliparous and preliminary findings and requires further analysis. These cases must be further examined as they do not meet clinical expectations set out by

Dr. Walker.

Clinical Diagnostic Accuracy PPV and NPV

The sensitivity and specificity values do not have a direct diagnostic meaning for physicians. This is provided in the PPV and NPV. Clinical standard methods use cervical length (CL) and fetal fibronectin (fFN) as predictors for preterm birth. As mentioned in the literature review, it is difficult to directly compare studies of preterm birth prediction accuracy due to the heterogeneity nature of studies as shown in Table 2.2 and discussed in Section 2.5.3. With this information in mind, not only does the sensitivity and specificity balance of best performing integrated DT-ANN Hybrid system for PAROUS cases exceed the results of all these studies, but as well as the PPV (43.93%) and NPV (92.64%) (Table 4.23). To add, some of these studies (Table 2.2), such as Study 3 to 5, have a higher prevalence of positive cases, thus affecting their respective PPV and NPV value. As expected, the PPV and NPV of Nulliparous cases were lower than the Parous cases. However, its prevalence was lower than most comparative studies conducted and its NPV was higher than all the NPV values in Table 2.2. Again, the user should keep in mind that the PPV and NPV are subject to change according to the rate of HR-PTB prevalence.

Concluding Remarks, Contributing Knowledge and Future Work

5.1 Conclusion

An improved integrated data mining method has been developed and extensively applied to solve two very different medical problems: predicting mortality in an adult ICU and predicting high-risk preterm birth in an obstetrical environment. In an attempt to solve current issues, this work has also raised questions that require further in-depth examination. This is presented in the conclusion, contributing knowledge and future work portion of this chapter.

5.1.1 Accomplished Preliminary Objectives

The first objective was to improve classification or match previous results by integrating the advantages of DTs with back-propagation ANNs. Classifier performances were analyzed to determine if there was an improvement with the use of DT-ANN Hybrids compared to weight-elimination ANNs alone. In order to accomplish the first objective, a structured ensemble hybrid system was developed. Application and analysis of this integrated system was applied to a small homogenous database in an adult ICU. The DT-ANN Hybrid surpassed the results of the weight-elimination ANN. Concluding from the first objective (the ensemble hybrid system was beneficial to an improved accuracy prediction system), provided support to use this system on a difficult, multi-factorial problem - predicting high-risk preterm birth. Though the accuracy prediction did not substantially exceed past attempts, there were several improvements made by the new approach developed in this thesis. This accomplished our second objective - to improve classification of high-risk preterm birth. Furthermore, preliminary results were obtained in predicting HR-PTB for

Nulliparous mothers - allowing a prediction system for a general population, not only to women predisposed to HR-PTB. This thesis work has established the following:

1. The integrated DT-ANN hybrid systems were superior compared to weight-elimination ANNs. This was shown in the increased efficiency and accuracy of the DT-ANN hybrid systems. This should be examined as a potential principle method to be used in further research work.
2. The DT-ANN approach was a robust method of creating an integrated DT-ANN hybrid classifier. Classifiers were reproducible when trained with new training and test points; however this was not the case when using the weight-elimination approach.
3. The need for an ensemble DT method was emphasized, as different bootstrapped training sets produced different trees. Each bootstrapped training set provided different attribute usage among each tree. Such evidence requires the user to examine different trees prior to choosing one as the classifier of choice.
4. There was a difference between test sensitivity and verification sensitivity when predicting rare outcomes. Such differences could be attributed to a limited amount of the rare outcome, thus resulting in a large variance in sensitivity. Consideration of this occurrence should be in place when applying, testing and reporting with this integrated system.
5. Created a mortality prediction system with excellent to outstanding discrimination and high accuracy, exceeding the results of current standard prediction systems. Additionally, only five variables are inputted into the system for predicting mortality in postoperative case and non-postoperative cases, which makes it very convenient for potential users. Contributed to an improved CDSS: The limited amount of variables found in the ICU integrated system solves issues that users have with CDSS such as high cost and extra time needed for data collection. The collection of 5 variables will

decrease the amount of interaction time with the system, thus solving the “burden” of data collection (Section 2.2.2). In addition, less data collection may mean less room for data error and a reduced complexity of the problem that could result in improved generalization.

6. Emphasized the variation in postoperative cases and non-postoperative environment by using the decision tree to classify cases. Identified prominent variables that were significant contributors to mortality in an adult ICU. Variables used for mortality prediction in postoperative cases included three APACHE variables (PH, Temperature, Partial Pressure of Oxygen in the Blood) and 2 Demographic variables (Age and Emergency Surgery). Variables used for mortality prediction in non-postoperative cases included four APACHE variables (Glasgow Coma Score, Fraction of Inspired Oxygen, Mean Arterial Pressure, and Hematocrit) and one demographic variable (Age). This detection could be useful for clinicians as it could mean better resource utilization, earlier means of intervention and successfully identify those that are targeted as positive cases by the system.
7. The integrated DT-ANN Hybrid was able to predict mortality in a homogeneous (POSTOP) and heterogeneous (noPOSTOP) environment with results that exceed the best in current literature.
8. The integrated DT-ANN Hybrid system was able to predict mortality in a more complex, heterogeneous, multi-factorial problem of high-risk preterm birth in Parous and Nulliparous Cases
9. The DT identified significant contributing variables for predicting mortality in the adult ICU database and predicting HR-PTB in the PRAMS database. The DTs used for predicting mortality had variables that remained consistent between each bootstrapped DT. However, the DT for HR-PTB did not. This difference in consistency

stressed the complexity and difficulty that is found when predicting premature births as revealed in Section 2.4.

10. Preliminary results for Nulliparous cases indicate poorer performance results compared to Parous cases. This was expected due to the variables that were not applicable for Nulliparous cases (*previous preterm birth, previous low birthweight, parity and previous live birth*).
11. Created an integrated hybrid model for the prediction of high-risk preterm birth for both mothers who have had previous children and for mothers giving birth to their first child. Past work focused only on mothers who have had previous children. Several improvements were made to past models: reducing the number of input variables to less than 20 from 48 variables while using variables that may be of international value and not US specific. Additionally, significant contributors have been identified, which would be valuable to healthcare professionals. Accurate identification of high-risk preterm birth cases would be useful in resource allocation for those who are in most need.
12. Created a non-invasive prediction system that is applicable to a general population **early** in the pregnancy (<23 weeks of gestation) and maintained clinical expectations. Such a system could be an alternative to costly and invasive methods.
13. DTs provide a prediction method that may be more comprehensive to the medical staff compared to ANNs. Using the developed DT-ANN hybrid model, a physician can see the significant contributing variables as well as their threshold splitting values. Physicians and healthcare professionals may be more receptive to this method as it provides easily understood information at the top level. In addition, common risk factors were identified in for Parous and Nulliparous cases, which may give more confidence in the developed prediction systems.

14. The integrated system not only gave good test results, but also verified well on new patient cases 7864 (Parous) and 5906 (Nulliparous). This is indicative that the system is not over-trained, good generalization has occurred and has verified well on many cases.
15. Provided a risk stratification level for cases, allowing the user to favour sensitivity or specificity as they see fit.

5.2 Contribution to Knowledge

1. This work demonstrated that the DT-ANN Hybrid system can be used to predict outcomes in two completely different environments: mortality in an adult ICU and high-risk preterm birth in an obstetrical environment. To our knowledge, it is the first time that an integrated ensemble Decision Tree Artificial Neural Network system has been applied to predict these outcomes.
2. Demonstrated that the DT-ANN Hybrid does not require that the rare outcome be re-sampled to 20% in order to correctly identify positive cases while maintaining sufficient specificity. Rare outcomes in the POSTOP training set needed only a small increase from 3.37% to 5.06% for successful identification of mortality cases.
3. Showed that the DT-ANN Hybrid required less processing and adjustment of parameters for an improved system. The DT portion did not need normalization in order to create a classifier. This saves a step in choosing the proper normalization methodology for the ANN. Using the attribute interval method, all inputs into the ANN are always between -1 and +1, allowing equal presentation of each variable to the network. Re-sampling was needed for the POSTOP training set, however, the noPOSTOP training set was left as it was.
4. Demonstrated the superior performance of the DT-ANN Hybrid compared to weight-

elimination alone; as determined by the sensitivity and specificity.

5. Established preliminary application of a decision tree and back-propagation MLP with weight-elimination to predict HR-PTB for Nulliparous women thus generalising the system for a general North American population.
6. Achieved excellent results in estimating mortality and highlighted contributing variables. Demonstrated that different variables were needed to predict mortality in both postoperative cases and non-postoperative cases.
7. Demonstrated that the 142 variables found in APACHE IV model may be unnecessary and a time-consuming process that can be eliminated by using the 5-variable integrated hybrid system for the ICU that provided the data. However, the system could be trained on data from any other ICU and it is likely that the number of variables needed would be substantially less than 142.
8. Demonstrated that the DT-ANN hybrid system is able to predict mortality and survival for individual cases. It is also able to predict high-risk preterm birth and low-risk preterm birth for individual cases, as well as provide a risk level per case. This system allows the user to adjust the cut-off threshold of the probability level in order to emphasize sensitivity or specificity.
9. Reduced the complexity of the HR-PTB problem by decreasing the number of input variables. For Parous cases the number of input variables was reduced to 19 from 69 and for Nulliparous cases it was reduced to 16 variables from 65 variables. This is a large improvement from preliminary results by Catley which needed 48 variables. This large decrease of input variables greatly reduces the complexity of the problem, which improves the generalization of the model.

5.3 Future Work

1. Determine what re-sampling rate is optimal to allow the hybrid system to identify rare cases. Although it was originally thought that 20% was needed for weight-elimination ANNs [98]; the mortality cases which were of rare occurrence, only had to be re-sampled to 5.06% using the DT-ANN hybrid system.
2. Explore reasons as to how the number of hidden nodes affects the performance of the classifier.
3. Apply the system on different adult ICU database on a national, international and global level and assess the hybrid system's performance on different ICU environments.
4. Further work is needed to improve the Nulliparous outcomes. These cases do not use primary predictor variables as they do not apply. Preliminary results are slightly under clinical expectations from our clinician partner. However, this is the first time that this has been examined. This requires further exploration in order to improve the results. Suggestions are to separate the cases into attributes that ranked high on the DT list - such as plurality, morbidity of the placenta, bleeding etc. An extensive, ensemble classification method may be essential to improve the results.
5. Due to the low number of variables and high consistency between the significant contributing factors for predicting mortality in an adult ICU, it was logical and simple to combine the variables to make the attribute intervals. The heterogeneity and complexity of predicting high-risk preterm birth is noticeable when contrasted with predicting mortality in an adult ICU. Due to the multi-factorial nature of predicting high-risk preterm births, each bootstrapped training set developed a different tree. Though some variables were consistently found between each bootstrapped

DT, many variables were different. This requires further analysis in predicting multifactorial problems such as preterm birth. A sound method is needed to determine how to make the best attribute interval. Does this involve combining all splitting values of each decision tree? Should only the consistent variables between each tree be used to make the attribute intervals? What is the difference in accuracy prediction if a set of attribute intervals was made for each of the three trees created in this work? Such interesting questions are beyond the scope of this thesis work but may be beneficial in improving the prediction capability of the ensemble DT-ANN Hybrid system.

6. Determine if there are limits to using the DT to first classify the data points. Advantages have been discovered, but further analysis would determine if potential information is lost when knowledge from the DT is integrated with the ANNs.
7. The integrated classification system for HR-PTB is a retrospective analysis and has been tested on the database prevalence of over 16%. In contrast, the rate of high-risk preterm birth is much lower in live births in Canada and the US. Canada reports a rate of preterm birth (<37 weeks) at 8.1% and US reports 12.8% of live births (Section 1.5.1). HR-PTB occurs at a much lower rate than preterm births. Since our classifier has been developed on a higher than normal prevalence of HR-PTB, how closely does this match the prediction for live births. Determine if undersampling of positive cases is required in order to match clinical outcome in prospective studies, or if this 'higher than normal' rate is indicative of predictive outcome.
8. When developing the final system, it may be beneficial to give a visual aspect for users of the system. This may mean presenting the decision tree and splitting variables as a visual aid in addition to the final risk stratification outcome.
9. Throughout the duration of this thesis work, applications for ethical clearance were made and revised numerous times to access the Atlee database, a perinatal database

collected in Nova Scotia (Canada). Collaboration has been made between neonatologists and researchers to access this database for further analysis. To date (November 18, 2009) we have been granted ethical clearance, but are still waiting to access this Canadian database. Examination of this database may reveal differences, consistency and improvements to the conclusions reached using the PRAMS database (see Appendix J for a portion of the Atlee application).

10. A retro-prospective analysis discovered a noteworthy issue to acknowledge: all the data was used for feature selection when creating the DT. After the creation of the attribute intervals (using all the data), the ANN input variables were split into a train, test and validation set. The train and test set were used in the creation of the classifier; while the validation set contained data not seen by the ANN classifier. Since all the data points were used for the DT portion - there may be a possibility that the final results contain optimal bias - or not. However, since the same procedure was used when comparing DT-ANN hybrid with weight-elimination ANNs, the DT-ANN hybrid still remains superior compared to weight-elimination ANN when predicting morality in the adult ICU database. Future work could include examining these results to determine if this method introduces optimal bias.
11. Determine how data allocation into a train, test and verification set affects the final performance results.

Final Statement Preliminary results of the integrated hybrid system proved to be beneficial for the estimation of ICU mortality and preterm birth. The knowledge gained from this thesis work has the possibility to greatly contribute not only to the data mining and artificial intelligence stream, but as well as the medical field. It is the hope that one day this sophisticated, integrated, hybrid system be seen in clinical use and its potential benefits be experienced by those in need.

Bibliography

- [1] J. Prince and J. Links, *Medical Imaging Signals and Systems*. Upper Saddle River, NJ: Pearson Education, 2006.
- [2] J. Webster, *Medical Instrumentation Application and Design*. Hoboken, NJ: John Wiley & Sons, Inc, 1998.
- [3] E.-H. Kluge, "Resource Allocation in Healthcare: Implications of Models of Medicine as a Profession," *Medscape*, vol. 9, p. 57, 2007.
- [4] A. Gullo and P. Lumb, "Scoring Systems and Outcome," in *Intensive and Critical Care Medicine: Reflections, Recommendations and Perspectives*, pp. 117–131, Italy: Springer-Verlag Italia, 2005.
- [5] D. Angus, A. Shorr, A. White, T. Dremsizov, R. Schmitz, and M. Kelley, "Critical Care Delivery in the United States: Distribution of Services and Compliance with Leapfrog Recommendations," *Critical Care Medicine*, vol. 34, pp. 311–314, 2006.
- [6] E. Berner, "Diagnostic Decision Support Systems: Why Aren't They Used More And What Can We Do About It?," *AMIA 2006 Symposium Proceedings*, pp. 1167–1168, 2006.
- [7] J. Norton, P. Barie, R. Bollinger, A. Chang, S. Lowry, S. Mulvihill, H. Pass, R. Thompson, and M. Shirazi, "Chapter 24-Risk Prediction, Disease Stratification, and Outcome Description in Critical Surgical Illness," in *Surgery: Basic Science and Clinical Evidence*, New York: Springer-Verlag, 2001.
- [8] T. Sinuff, N. Adhikari, D. Cook, H. Schünemann, L. Griffith, G. Rocker, and S. Walter, "Mortality Predictions in the Intensive Care unit: Comparing Physicians with Scoring Systems," *Critical Care Medicine*, vol. 34, pp. 878–885, 2006.
- [9] H. Anderson, C. Nugent, S. Wanty, and R. H. Hayashi, "Prediction of Risk for Preterm Delivery by Ultrasonographic Measurement of Cervical Length," *Am J Obstet Gynecol*, vol. 163, pp. 859–67, 1990.
- [10] D. Wong and W. Knaus, "Predicting outcome in critical care: the current status of the APACHE prognostic scoring system," *Can J Anesth*, vol. 38, pp. 378–83, 1991.
- [11] G. Latendresse, "The Interaction Between Chronic Stress and Pregnancy: Preterm Birth from A Biobehavioral Perspective," *American College of Nurse-Midwives*, pp. 8–17, 2009.

- [12] J. Iams, R. Goldenberg, B. Mercer, A. Moawad, E. Thom, and P. Meis, "Previous Preterm Birth, Fetal Fibronectin, and Cervical Length at 24 weeks were Independent Predictors of Preterm Delivery," *American Journal of Obstetrics and Gynecology*, pp. 1035–1040, 1998.
- [13] W. Klosgen and J. Zytchow, "Decision-Tree Discovery," in *Handbook of Data Mining and Knowledge Discovery*, pp. 267–276, New York, NY: Oxford University Press, 2002. section author: R. Kohavi and Quinlan.
- [14] D. Rybchynski, "Design of an Artificial Neural Network Research Framework to Enhance the Development of Clinical Prediction Models," Masters Thesis, University of Ottawa, Ottawa, Ontario, Canada, 2005.
- [15] C. Catley, *An Integrated Hybrid Data Mining System for Preterm Birth Risk Assessment Based on A 'Semantic Web Service for Healthcare' Framework*. PhD thesis, Department of Systems and Computer Engineering— Carleton University, Ottawa, Ontario, 2007.
- [16] D. Zhou and M. Frize, "Predicting Probability of Mortality in the Neonatal Intensive Care Unit," *IEEE*, pp. 2305–2311, 2006.
- [17] S. Mitra and T. Acharya, *Data Mining Multimedia, Soft Computing and Bioinformatic*. Hoboken, NJ: John Wiley & Sons, Inc., 2003.
- [18] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. New York, NY: John Wiley & Sons, Inc., 2nd ed., 2001.
- [19] I. Sethi, "Entropy Nets: From Decision Trees to Neural Networks," *Proceedings of IEEE*, pp. 1605–1613, 1990.
- [20] Z. Zhou and Z. Chen, "Hybrid Decision Tree," *Knowledge-Based Systems–Elsevier*, pp. 515–528, 2002.
- [21] Critical Care Society, "The Role of ICU —What to Expect." website: http://www.canadiancriticalcare.org/patientfamily_icu_role.htm, September 25 2009.
- [22] K. Alameddine, M. and Dainty, R. Deber, and W. Sibbald, "The Intensive Care Unit Work Environment: Current Challenges and Recommendations for the Future," *Journal of Critical Care*, vol. 24, pp. 243–248, 2009.
- [23] R. Kaushal, D. Bates, C. Franz, J. Soukup, and J. Rothschild, "Costs of adverse events in intensive care units," *Critical Care Medicine*, vol. 35, pp. 2479–2483, 2007.
- [24] "iatrogenic", "Concise Medical Dictionary." website, 22 September 2009. Oxford University Press 2007. Carleton University. <<http://www.oxfordreference.com/views/ENTRY.html?subview=Main&entry=t60.e1665>>.

- [25] D. Talmor, N. Shapiro, D. Greenberg, P. Stone, and P. Neumann, "When is Critical Care Medicine Cost-effective? A Systematic Review of the Cost-effectiveness Literature," *Anesth Analg*, vol. 34, pp. 2738–2747, 2006.
- [26] D. Angus and J. Carlet, "Surviving Intensive Care: a report from the 2002 Brussels Roundtable," *Intensive Care Medicine*, vol. 29, pp. 368–377, 2003.
- [27] J. Luc and G. Rubenfeld, "Can Health Care Costs Be Reduced by Limiting Intensive Care at the End of Life," *Am J Respir Crit Care Med*, vol. 165, pp. 750–754, 2002.
- [28] T. Higgins, "Quantifying Risk and Benchmarking Performance in the Adult Intensive Care Unit," *J Intensive Care Med*, vol. 22, pp. 141–156, 2007.
- [29] B. Afessa, O. Gajic, and M. Keegan, "The Large Cost of Critical Care: Realities and Challenges," *Critical Care Clinic*, vol. 23, pp. 639–658, 2007.
- [30] S. Segal and E. Norwitz, "Chapter 20— preterm labour," in *Anaesthetic and Obstetric Management of High-Risk Pregnancy*, Springer- Verlag New York, third ed., 2004. accessed through <http://lib.myilibrary.com.proxy.library.carleton.ca/Browse/open.asp?ID=18890>.
- [31] "preterm birth", "Concise Medical Dictionary." website, 22 July 2009. Oxford University Press 2007. Carleton University. <<http://www.oxfordreference.com/views/ENTRY.html?subview=Main&entry=t60.e8177>>.
- [32] M. Chandiramani, R. Tribe, and A. Shennan, "Preterm labour and prematurity," *Obstetrics, Gynaecology and Reproductive Medicine—Elsevier*, pp. 232–237, 2007.
- [33] K. Joseph, M. Kramer, S. Marcoux, A. Ohlsson, S. Wu-Wen, and R. Allen, A. and Platt, "Determinants of Preterm Birth Rates in Canada from 1981 through 1983 and from 1992 through 1994," *The New England Journal of Medicine*, vol. 339, pp. 1434–1439, November 12 1998.
- [34] Canadian Institute for Health Information, "Too Early, Too Small: A Profile of Small Babies Across Canada." website, 2009. accessed June 09, 2009 <http://secure.cihi.ca/cihiweb/disPage.jsp?cw_page=AR_1106_E.
- [35] M. Davidoff, T. Dias, K. Damus, R. Russell, V. Bettegowda, S. Dolan, R. Schwarz, N. Green, and J. Petrini, "Changes in the Gestational Age Distribution among U.S. Singleton Births: Impact on Rates of Late Preterm Birth, 1992 to 2002," *Seminars in Perinatology— Elsevier*, pp. 8–15, 2006.
- [36] J. Petrini, T. Dias, M. McCormick, M. Massolo, and N. Green, "Increased Risk of Adverse Neurological Development for Late Preterm Infants," *The Journal of Pediatrics*, pp. 1–11, 2008.

- [37] K. Tomashek, C. Shapiro-Mendoza, M. Davidoff, and J. Petrini, "Differences in Mortality between Late-Preterm and Term Singleton Infants in the United States, 1995—2002," *Mosby, Inc*, pp. 450–456, 2007.
- [38] P. Kumar and G. Suresh, "Complications After Preterm Birth: An Overview for Emergency Physicians," *Clinical Pediatric Emergency Medicine*, pp. 191–199, 2008.
- [39] S. Wood, K. Marlow, N. Costeloe, A. Gibson, and A. Wilkinson, "Neurologic and Developmental Disability after Extremely Preterm Birth," *The New England Journal of Medicine*, pp. 378–384, September 2000.
- [40] March of Dimes, "Premature Birth # 1 Cause of Newborn Death—Help Reduce Cost." website, July 15 2009. accessed through http://www.marchofdimes.com/prematurity/21198_10734.asp.
- [41] R. Walker and M. Frize. Carleton University. Ottawa, Ontario. lecture slide, September 2009. Dr. Robin Walker as a guest lecturer for BIOM 5002 (Ethics, Research Methods and Standards for Biomedical Engineering).
- [42] C. Ennett and M. Frize, "Weight-Elimination Neural Networks Applied to Coronary Surgery Mortality Prediction," *IEEE Transactions on Information Technology in Biomedicine*, pp. 86–926, 2003.
- [43] D. Zhou, "Adding Probability to Neural Network Prediction of NICU Mortality," Master's thesis, University of Ottawa, Ottawa, Ontario, 2006.
- [44] G. Bertolini, R. D'Amico, G. Apolone, A. Cattaneo, A. Ravizza, G. Iapichino, L. Brazzi, and R. Maria Melotti, "Predicting Outcome in the Intensive Care Unit Using Scoring Systems: Is New Better? A Comparison of SAPS and SAPS II in a Cohort of 1,393 Patients," *Medical Care – JSTOR*, vol. 36, pp. 1371–1382, 1998.
- [45] M. Capuzzo, V. Valpondi, A. Sgarbi, s. Bortolazzi, V. Pavoni, G. Gilli, G. Candini, G. Gritti, and R. Alvisi, "Validation of Severity Scoring Systems SAPS II and APACHE II in Single-center Populations ," *Intensive Care Medicine*, vol. 26, pp. 1779–1785, 2000.
- [46] X. Castella, A. Artigas, J. Bion, and A. Kari, "A comparison of severity of illness scoring systems for intensive care unit patients: Results of a multicenter, multinational study," *Critical Care Medicine*, vol. 23, pp. 1327–1335, 1995.
- [47] R. Moreno, B. Jordan, and P. Metnitz, "The Changing Prognostic Determinants in the Critically Ill Patient," in *Yearbook of Intensive Care and Emergency Medicine*, SpringerLink, 2007.
- [48] B. Zhu, S. Lemeshow, D. Hosmer, J. Klar, J. Avrunin, and D. Teres, "Factors affecting the performance of the models in the Mortality Probability Model II system and

- strategies of customization: A simulation study,” *Critical Care Medicine*, vol. 24, pp. 57–63, 1996.
- [49] N. Peek, D. Arts, R. Bosman, P. van der Voort, and N. de Keizera, “External validation of prognostic models for critically ill patients required substantial sample sizes,” *Journal of Clinical Epidemiology*, vol. 60, pp. 491–501, 2007.
 - [50] J. Hilden, “The Area Under the ROC curve and Its Competitors,” *Medical Decision Making*, vol. 11, pp. 95–101, 1991.
 - [51] N. Obuchowski, “Receiver Operating Characteristic Curves and Their Use in Radiology,” *Radiology*, vol. 229, pp. 3–8, 2003.
 - [52] F. Bakhshi-Raiez, N. Peek, R. Bosman, E. Jonge, and N. De Keizer, “The Impact of Different Prognostic Models and Their Customization on Institutional Comparison of Intensive Care Unit,” *Critical Care Medicine*, vol. 35, pp. 2553–2560, 2007.
 - [53] J. Zimmerman, A. Kramer, D. McNair, and F. Malila, “Acute Physiology and Chronic Health Evaluation (APACHE) IV Hospital Mortality Assessment for Today’s Critically Ill Patients,” *Critical Care Medicine*, vol. 34, no. 5, pp. 1297–1310, 2008.
 - [54] M. Kuzniewicz, E. Vasilevskis, R. Lane, M. Dean, N. Trivedi, D. Rennie, R. Clay, P. Kotler, and R. Dudley, “Variation in ICU Risk-Adjusted Mortality,” *Critical Care Medicine*, vol. 133, no. 6, pp. 1319–1327, 2008.
 - [55] R. Moreno, P. Metnitz, E. Almeida, B. Jordan, P. Bauer, R. Abizanda Campos, G. Iapichino, D. Edbrooke, M. Capuzzo, and J.-R. Le Gall, “SAPS 3—From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission,” *Intensive Care Medicine—Springer*, vol. 31, pp. 1345–1355, 2005.
 - [56] P. McParland, G. Jones, and D. Taylor, “Preterm Labour and Prematurity,” *Current Obstetrics & Gynaecology Elsevier*, vol. 14, pp. 309–319, 2004.
 - [57] D. Murphy, “Epidemiology and Environmental Factors in Preterm Labour,” *Best Practice & Research Clinical Obstetrics and Gynaecology -Elsevier*, vol. 21, pp. 773–789, 2007.
 - [58] S. Partington, “Seconds Births to Teenage Mothers: Risk factors for Low Birth Weight and Preterm Birth,” *Perspective Sex Reprod Health*, vol. 41, no. 2, pp. 101–109, 2009.
 - [59] C. O’Leary, N. Nassar, J. Kurinczuk, and C. Bower, “The Effect of Maternal Alcohol Consumption on Fetal Growth and Preterm Birth,” *BJOG: An International Journal of Obstetrics & Gynaecology*, vol. 116, pp. 390–400, 2008.

- [60] Centers for Disease Control, "Smoking & Tobacco Use." website: <www.cdc.gov/tobacco/data_statistics/sgr/2004/highlights/cdifferent/index.htm>, 2004.
- [61] March of Dimes, "March of Dimes Study Unveils New Data on the Cost of Having a Baby." website, July 15 2009. accessed through http://www.marchofdimes.com/prematurity/21198_10734.asp.
- [62] P. Kalk, F. Guthmann, K. Krause, K. Relle, M. Godes, G. Goassing, H. Halle, R. Wauer, and B. Hoher, "Impact of maternal body mass index on neonatal outcome," *Eur J Med Res*, vol. 14, no. 5, pp. 216–222, 2009.
- [63] S. Hernandez-Diaz, S. Toh, and S. Cnattingius, "Risk of pre-eclampsia in first and subsequent pregnancies: prospective cohort study," *BMJ*, p. 338, 2009.
- [64] M. Hedderson, A. Ferrara, and D. Sacks, "Gestational Diabetes Mellitus and Lesser Degrees of Pregnancy Hyperglycemia: Association With Increased Risk of Spontaneous Preterm Birth," *The American College of Obstetricians and Gynecologists—Scholars Portal Journal*, vol. 102, pp. 850–856., 2003.
- [65] N. Melamed, R. Chen, U. Soiberman, A. Ben-Haroush, M. Hod, and Y. Yogeve, "Spontaneous and indicated preterm delivery in pregestational diabetes mellitus: etiology and risk factors," *Arch Gynecol Obstet—Scholars Portal Journal*, vol. 278, pp. 129–134, 2008. Israel.
- [66] J. Yang, K. Hartmann, D. Savitz, A. Herring, N. Dole, A. Olshan, and J. Thorp, "Vaginal Bleeding during Pregnancy and Preterm Birth," *American Journal of Epidemiology*, vol. 160, p. 118, 2004.
- [67] "cervical cerclage", "Concise Medical Dictionary." website, 22 July 2009. Oxford University Press 2007. Carleton University. <<http://www.oxfordreference.com/views/ENTRY.html?subview=Main&entry=t60.e1665>>.
- [68] S. Mazaki—Tovi, R. Romero, J. Kusanovic, O. Erez, B. Pineles, F. Gotsch, P. Mittal, N. Than, J. Espinoza, and S. Hassan, "Recurrent preterm birth," *Seminars in Perinatology—Elsevier*, vol. 31, pp. 142–158, 2007.
- [69] R. Pingili, V. Bamigboye, and K. Jegede, "Multiple pregnancy: a blessing or a curse?(Report)," *The Internet Journal of Gynecology and Obstetrics*, vol. 84, July 2008. accessed July 12,2009 – University of Ottawa.
- [70] R. Goldenberg, A. Goepfert, and P. Ramsey, "Biochemical markers for the prediction of preterm birth," *American Journal of Obstetrics and Gynecology*, vol. 192, pp. 36–46, 2005.
- [71] H. Lee Joong, T. Park Chul, and E. Norwitz, "Management of Pregnancies With Cervical Shortening: A Very Short Cervix Is a Very Big Problem," *Reviews in Obstetrics & Gynecology*, vol. 2, pp. 107–115, 2009.

- [72] V. Berghella, E. Hayes, J. Visintine, and J. Baxter, "Fetal Fibronectin Testing for Reducing the Risk of Preterm Birth (Review)," *Cochrane Database Syst Rev.*, vol. 8, October 2008.
- [73] R. Goldenberg, A. Goepfert, and P. Ramsey, "Biochemical Markers for the Prediction of Preterm Birth," *American Journal of Obstetrics and Gynecology-Elsevier*, pp. 36–46, 2005.
- [74] V. Berghella, A. Roman, C. Daskalakis, A. Ness, and J. Baxter, "Gestational Age at Cervical Length Measurement and Incidence of Preterm Birth," *The American Journal of Obstetrics and Gynecology*, vol. 110, no. 2, pp. 311–317, 2007.
- [75] A. Roman, A. Rebarber, A. Sfakianaki, J. MD, Mulholland, D. Saltzman, M. Paidas, V. Minior, and C. Lockwood, "Vaginal fetal fibronectin as a predictor of spontaneous preterm delivery in the patient with cervical cerclage," *American Journal of Obstetrics and Gynecology*, vol. 189, pp. 1368–73., 2003.
- [76] K. Boggess, M. K., A. Murtha, S. Offenbacher, and J. Beck, "Antepartum vaginal bleeding, fetal exposure to oral pathogens, and risk for preterm birth at < 35 weeks of gestation," *American Journal of Obstetrics and Gynecology*, vol. 194, no. 4, pp. 954–60, 2009.
- [77] J. Hibbard, M. Tart, and A. Moawad, "Cervical Length at 16–22 Weeks' Gestation and Risk for Preterm Delivery," *The American Journal of Obstetricians and Gynecologists*, vol. 96, pp. 972–978, 2000.
- [78] Ozdemir, I. and Demirci, F. and Yucel, O. and Erkorkmaz, U., "Ultrasonographic Cervical Length Measurement at 10–14 and 20–24 Weeks Gestation and the Risk of Preterm," *European Journal of Obstetrics & Gynecology and Reproductive Biology—Elsevier*, 2007.
- [79] R. Gomez, R. and Romero, J. Nien, T. Chaiworapongsa, L. Medina, Y. Kim, B. Yoon, M. Carstens, J. Espinoza, J. Iams, and R. Gonzalez, "A short cervix in women with preterm labor and intact membranes: A risk factor for microbial invasion of the amniotic cavity," *American Journal of Obstetrics and Gynecology*, pp. 678–689, 2005.
- [80] S. Hassen, R. Romero, S. Berry, K. Dang, S. Blackwell, M. Treadwell, and H. Wolfe, "Patients with an ultrasonographic cervical length $\leq$.15 mm have nearly a 50% risk of early spontaneous preterm delivery," *Am J Obstet Gynecol-Elsevier*, vol. 182, pp. 1458–1467, 2000.
- [81] D. Anumba, "Management of women with a previous preterm birth," *Obstetrics, Gynaecology and Reproductive Medicine-Elsevier*, vol. 17, pp. 188–191, 2007.
- [82] H. Leiticha and A. Kaider, "Fetal fibronectin—how useful is it in the prediction of preterm birth?," *BJOG: an International Journal of Obstetrics and Gynaecology*, vol. 110, pp. 66–77, 2003.

- [83] K. Duhig, M. Chandiramani, P. Seed, A. Briley, A. Kenyon, and A. Shennan, "Fetal Fibronectin as a Predictor of Spontaneous Preterm Labour in Asymptomatic Women with a Cervical Cerclage," *BJOG*, vol. 112, pp. 799–803, 2009.
- [84] M. Incerti, A. Ghidini, V. Korker, and J. Pezzullo, "Performance of cervicovaginal fetal fibronectin in a community hospital setting," *Arch Gynecol Obstet*, vol. 275, pp. 347–351, 2007.
- [85] M. Lowe, B. Zimmerman, and W. Hansen, "Prospective randomized controlled trial of fetal fibronectin on preterm labor management in a tertiary care center," *General obstetrics and gynecology: obstetrics-Elsevier*, vol. 190, pp. 358–362, 2004.
- [86] J. Iams, D. Casal, J. McGregor, M. Goodwin, U. Kreaden, and R. Lowensohn, "Fetal fibronectin improves the accuracy of diagnosis of preterm labor.," *American Journal of Obstetrics & Gynecology*, vol. 173, pp. 141–145, 1995.
- [87] R. Lopez, J. Francis, T. Garite, and J. Dubyak, "Fetal fibronectin detection as a predictor of preterm birth in actual clinical practice.," *American Journal of Obstetrics & Gynecology*, vol. 182, pp. 1103–1106, 2000.
- [88] The Reproductive Care Program of Nova Scotia, "Fetal Fibronectin (fFN) Testing — "Fetal Fibronectin Guideline for Use in the Management of Preterm Labour"," 2001. accessed through <http://rcp.nshealth.ca/rcp_7632.html>.
- [89] B. Mercer, R. Goldenberg, A. Das, A. Moawad, J. Iams, P. Meis, R. Copper, F. Johnson, E. Thom, D. McNellis, M. Miodovnik, M. Menard, S. Caritis, G. Thurnau, S. Bottoms, and J. Roberts, "The Preterm Prediction study: A Clinical Risk Assessment System," *American Journal Of Obstetrical & Gynecology*, vol. 174, pp. 1885–95, 1996.
- [90] H. Honest, L. Bachmann, R. Sundaram, J. Gupta, J. Kleijnen, and K. Khan, "Review. The accuracy of risk scores in predicting preterm birth—a systematic review," *Journal of Obstetrics and Gynaecology*, vol. 24, pp. 1343–359, 2004.
- [91] W. Maner and R. Garfield E., "Identification of Human Term and Preterm Labor using Artificial Neural Networks on Uterine Electromyography Data," *Annals of Biomedical Engineering*, vol. 35, pp. 465–473, March 2007.
- [92] H. Tan, S. Wu Wen, X. Kuan Chen, K. Demissie, and M. Walker, "Early prediction of preterm birth for singleton, twin, and triplet pregnancies," *European Journal of Obstetrics & Gynecology and Reproductive Biology-Elsevier*, vol. 131, pp. 132–137, 2007.
- [93] J. Iams, "Aprediction and early detection of preterm labour," *The American College of Obstetricians and Gynecologists*, pp. 402–412, 2003.
- [94] R. Li and Z. Wang, "Mining classification rules using rough sets and neural networks," *European Journal of Operational Research*, pp. 439–448, 2004.

- [95] I. Ivanova and M. Kubat, "Initialization of Neural Networks by Means of Decision Trees," *Knowledge-Based Systems— Elsevier*, pp. 333–344, 1995.
- [96] I. Koprinska, G. Pfurtscheller, and D. Flotzinger, "Sleep Classification in Infants by Decision Tree-based Neural Networks," *Artificial Intelligence in Medicine*, pp. 387–401, 1996.
- [97] R. Setiono and W. Leow, "On mapping decision trees and neural networks," *Knowledge-Based Systems— Elsevier*, pp. 95–99, 1999.
- [98] C. Ennett and M. Frize, "Selective Sampling to Overcome Skewed a priori Probabilities," *AMIA*, pp. 225–229, 2000.
- [99] H. Trigg, "An investigation of methods to enhance the performance of artificial neural networks used to estimate icu outcomes," Master's thesis, University of New Brunswick, 1997.
- [100] M. Frize, C. Ennett, M. Stevenson, and H. Trigg, "Clinical decision support systems for intensive care units: using artificial neural networks," *Medical Engineering & Physics—Elsevier*, pp. 217–225, 2001.
- [101] C. Catley, M. Frize, R. Walker, and D. Petriu, "Predicting Preterm Birth Using Artificial Neural Networks," *Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems*, pp. 103–108, 2005.
- [102] "The APGAR score," *Advances in Neonatal Carer*, vol. 6, pp. 220–223, 2006.
- [103] C. Catley, M. Frize, R. Walker, and D. Petriu, "Predicting High-Risk Preterm Birth Using Artificial Neural Networks," *IEEE Transactions on Information Technology in Biomedicine*, vol. 10, pp. 540–548, 2006.
- [104] Y. Tong, M. Frize, and R. Walker, "Estimating Ventilation Using Artificial Neural Networks in Intensive Care Units," *Proceedings Of The First Joint BMES/EMES Conference, IEEE*, p. 1229, 1999.
- [105] J. Jerez-Aragónés, J. Gómez-Ruiza, G. Ramos-Jiménez, J. Muñoz-Pérez, and E. Alba-Conejo, "A combined neural network and decision trees model for prognosis of breast cancer relapse," *Artificial Intelligence in Medicine —Elsevier*, pp. 45–63, 2002.
- [106] M. Kantardzic, *Data Mining Concepts, Models, Methods, and Algorithms*. Hoboken, NJ: The Institute of Electrical and Electronics Engineers Press — John Wiley & Sons, 2002.
- [107] C. Williams and D. Wheeler, "Criteria for ICU Admission and Severity of Illness Scoring," *Surgery -Elsevier*, vol. 27, no. 5, pp. 201–206, 2009.

- [108] Centers for Disease Control and Prevention, “Pregnancy Risk Assessment Monitoring System (PRAMS): Home,” May 5th 2009. accessed on May 05, 2009 through <<http://www.cdc.gov/prams/>>.
- [109] P. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Boston, MA: Pearson Education Inc, 2006.
- [110] D. Altman and A. Bland, “Diagnostic Tests 2: Predictive Power,” *British Medical Journal*, vol. 309, p. 6947, 1994.
- [111] H. Kelly, A. Bull, P. Russo, and E. McBryde, “Estimating sensitivity and specificity from positive predictive value, negative predictive value and prevalence: application to surveillance systems for hospital-acquired infections,” *Journal of Hospital Infection—Elsevier*, pp. 164–168, 2008.
- [112] T. Fawcett, “An Introduction to ROC Analysis,” *Pattern Recognition Letters—Elsevier*, pp. 861–87, 2006.
- [113] W. Ambrosius, *Topics in Biostatistics (Methods in Molecular Biology)*. Humana Press, 1st ed., 2007.
- [114] A. Pan, “MIRG ANN Guide,” 2002. University of Ottawa — School of Information Technology and Engineering.
- [115] M. Dash and H. Liu, “Feature Selection for Classification,” *Intelligent Data Analysis—Elsevier*, pp. 131–156, 1997.
- [116] C. Ennett, M. Frize, and E. Charette, “Improvement and automation of artificial neural networks to estimate medical outcomes,” *Medical Engineering & Physics — Elsevier*, vol. 26, pp. 321–328, 2004.
- [117] C. Ennett, M. Frize, and N. Scales, “Evaluation of the Logarithmic-Sensitivity as a Neural Network Stopping Criterion for Rare Outcomes,” *IEEE*, pp. 338–341, 2003.
- [118] Beiu, “A Novel Highly Reliable Low-Power Nano Architecture When von Neumann Augments Kolmogorov,” *15th IEEE International Conference on Application-Specific Systems, Architectures and Processors*, pp. 167–177, 2004.
- [119] O. Maimon and L. Rokach, *Data Mining and Knowledge Discovery Handbook*. New York, NY: Springer Science & Business Media, Inc, 2005.
- [120] C. Bishop, *Neural Networks for Pattern Recognition*. New York, NY: Oxford University Press, 1995.
- [121] C. Ennett, M. Frize, and R. Walker, “Imputation of Missing Values by Integrating Neural Networks and Case-based Reasoning,” *IEEE*, pp. 4337–4341, 2008.

- [122] P. Braspenning and A. Thuijsman, F.and Weijters, *Artificial Neural Network: An Introduction to Theory and Practice*. New York, NY: Springer-Verlag Berlin Heidelberg, 1995.
- [123] S. Ogulata, C. Sahin, and R. Erol, "Neural Network–Based Computer–Aided Diagnosis in Classification of Primary Generalized Epilepsy by EEG Signals," *Journal of Medical Systems*, vol. 33, pp. 107–112, 2009.
- [124] S. Maslekar, A. Gardiner, J. Monson, and G. Duthie, "The cervix and prematurity: aetiology, prediction and prevention," *Colorectal Disease*, vol. 10, 2009.
- [125] E. Tzanakou, *Supervised and Unsupervised Pattern Recognition: Feature Extraction and Computational Intelligence*. Florida: CRC Press, 2000.
- [126] A. Gupta and S. Lam, "Weight Decay Backpropagation for Noisy Data," *Neural Networks*, pp. 1127–1137, 1998.
- [127] N. Sundararajan and P. Saratchandran, *Parallel architectures for artificial neural networks*. Piscataway, NJ: Institute of Electrical and Electronics Engineers Computer Society Press, 1998.
- [128] R. Research, "Is See5/C5.0 Better Than C4.5?," 2009 accessed on June 02, 2009. <<http://www.rulequest.com/see5-comparison.html>>.
- [129] R. Sarker, H. Abbass, and C. Newton, "Chapter 13-heuristics in medical data mining," in *Heuristics and Optimization for Knowledge Discovery*, Idea Group Publishing, 2002. accessed through <http://library.books24x7.com.proxy.library.carleton.ca>.
- [130] I. Witten and E. Frank, *Data Mining: Practical Machine learning Tools and Techniques*. San Francisco, CA: Elsevier, Inc, 2nd ed., 2005.
- [131] B. Sarkar and S. Sana, "A Hybrid Approach to Design Efficient Learning Classifiers," *Computers and Mathematics with Applications —Elsevier.*, pp. 65–73, 2009.
- [132] B. Chandra and P. P. Varghese, "Fuzzifying Gini Index Based Decision Trees," *Expert Systems with Applications —Elsevier*, pp. 8549–8559, 2009.
- [133] R. Research, "See5: An Informal Tutorial," June 02 2009. <[tp://www.rulequest.com/see5-win.html#FUZZY](http://www.rulequest.com/see5-win.html#FUZZY)>.
- [134] R. Dybowski and V. Grant, *Clinical Applications of Artificial Neural Networks*. New York, NY: Cambridge University Press,, 2001.
- [135] U. Braga–Neto, "Fad and Fallacies in the Name of Small Sample Micro-Array Classification," *IEEE Signal Processing Magazine*, pp. 91–100, 2007.
- [136] D. Verbyla and J. Litvaitis, "Resampling Methods for Evaluating Classification Accuracy of Wildlife Habitat Models," *Environmental Management — SpringerLink*, pp. 783–787, 1989.

- [137] P. Cohen, *Empirical Methods for Artificial Intelligence*. Cambridge, Massachusetts: Massachusetts Institute of Technology, 1995.
- [138] C. Mooney, R. Duval, and M. Lewis–Beck, *Bootstrapping: A Nonparametric Approach to Statistical Inference*. Sage Publications, Inc., 1993.
- [139] Y. Saeys, I. Inza, and P. Larraaga, “A Review of Feature Selection Techniques in Bioinformatics,” *Bioinformatics*, pp. 2507–2517, 2007.
- [140] H. Liu, “Evolving Feature Selection,” *IEEE Intelligent Systems*, pp. 64–76, 2005.
- [141] A. Sharkey, *Combining Artificial Neural Nets – Ensemble and Modular Multi-Net Systems*. London: Springer-Verlag London Limited, 1999.
- [142] A. Kramer and J. Zimmerman, “Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited,” *Critical Care Medicine*, vol. 35, pp. 2052–2056, 2007.
- [143] N. Stallard, “Simple tests for the external validation of mortality prediction scores,” *Statistics in Medicine*, vol. 28, pp. 377–388, 2009.

Tables

Table A.1: Original Adult ICU Variables and their meaning.

	Variable	Meaning
	FIN_ORD	Case Number (REMOVED)
1	CHRONIC	Did patient have chronic condition
2	EMER_SRG	Did patient have emergency surgery (No [-1], Yes [1])
3	SEX	Patient Gender (-1=female,1=male)
4	Age	Patient Age (years)
5	TEMP	Rectal Temperature (°C)
6	MAP	Mean Arterial Pressure (mmHG)
7	HR	Heart Rate (beats/minute)
8	RESP	Respiratory Rate (breaths/min)
9	FIO2	Fraction of Inspired Oxygen
10	PO2	Partial Pressure of Oxygen in the Blood (mmHg)
11	PH	Arterial pH
12	NA	Serum Sodium (mMol/L)
13	K	Serum Potassium (mMol/L)
14	CREAT	Serum Creatinine (micromol/L)
15	HEMA	Hematocrit (Ratio of Red Blood Cell Volume to Total Blood Volume)
16	WBC	White Blood Cell Count (total/mm ³ in 1000's)
17	GCS	Acute Glasgow Coma Score
18	DEATH	OUTPUT: non-survivor [-1] or survivor [1]

Table A.2: POSTOP and noPOSTOP mean and standard deviation used to normalize POSTOP and noPOSTOP database. Values were not rounded to significant figures as these values were used to normalized the variables for each case

Variable Name	POSTOP		noPOSTOP	
	Mean	STD	Mean	STD
1 CHRONIC CONDITION			categorical	
2 EMER_SRG			categorical	
3 SEX			categorical	
4 AGE (years)	64.09	15.12	56.87	19.22
5 TEMP (°C)	36.41	1.14	36.96	1.14
6 MAP(mmHG)	92.70	17.18	89.44	18.13
7 HR (beats/minute)	84.65	22.24	102.57	23.87
8 RESP (breaths/min)	16.09	5.65	21.49	9.01
9 FIO2	0.4532	0.25518	0.5093	0.32998
10 PO2 (mmHg)	129.44	85.11	130.22	100.21
11 PH	7.37	0.074	7.36	0.121
12 NA (mMol/L)	137.63	3.49	137.44	6.17
13 K (mMol/L)	4.28	0.563	4.25	0.785
14 CREAT (micromol/L)	97.82	37.254	133.12	88.764
15 HEMA	0.34841	0.053603	0.37687	0.072984
16 WBC (total/mm ³ in 1000's)	13.32	6.168	14.94	20.079
17 GCS	10.60	4.738	11.25	4.380

Table A.3: ICU ANN Starting parameters.

	DT-Hybrid	we_ANN
	(weight elimination OFF)	(weight elimination ON)
Learning Rate	0.00005	0.00005
Learning Inc	1.01	1.01
Learning Dec	0.99	0.99
Lambda (weight-decay)	NA	0.0001
Lamda_inc	NA	0.75
Lambda_dec	NA	1.25
Wnot	NA	0.0001
Momentum	0.5	0.5
Err_ratio	1.02	1.02

Table A.4: PRAMS input variable list for Decision Tree including confounding variables (no. = Number, avg = Average, mat = Maternal). (A) Variable (B) Explanation of variable (C) Variable Origination: variables originated from (1) PRAMS questionnaire, (2) birth certificate or, (3) was created from data processing steps in this thesis work.

A	B	C
1 MAT_AGE	mat age	1
2 PLURAL	plurality single, twin, triplet (specify)	1
3 PREV_LB	previous live birth	2
4 PARITY	parity (from PRE_LB) is the no of previous children	1
5 PRE_PRE	previous no of children	1
6 PRE_LBW	previous low birthweight	2
7 DRK_2YRS	mat drinking in the last two years	2
8 MOMDRINK	mat drinking during pregnancy	2
9 MOMDRKS	mat drinking: no of drinks during pregnancy	2
10 B_MOMDRKS	mat drinking: 3 months before pregnancy	2
11 NO_DRKS_BEFORE	mat drinking: avg. no of drinks 3 months before pregnancy	2
12 NO_BINGES_BEFORE	no of binges 3 months before pregnancy	2
13 DRK43L_A	mat drinking: last 3 months of pregnancy	2
14 DRK4_3L	mat drinking: avg no. of drinks in last 3 months of pregnancy	2
15 DRK4_3LB	no of binges in last 3 months of pregnancy	2
16 SMK_2YRS	mat smoking: at least 100 cigarettes in the past 2 years	2
17 SMK_3BEFORE	mat smoking: last 3 months of pregnancy	2
18 SMK_3BEFORE	mat smoking: 3 months before pregnancy	2
19 SMK_3LATER	mat smoking: last 3 months of pregnancy	2
20 MOM_CIG_LATER	mat smoking: no of cigarettes smoked in last 3 months of pregnancy	2
21 MOM_CIG_before	mat smoking: no of cigarettes smoked 3 months before pregnancy	2
22 MOMSMOKE	mat smoking	1
23 MOMCIG	mat smoking: no of cigarettes per day	2
24 SEX	sex of the child	1
25 MARRIED	is the mother married	1
26 MAT_ED	mat education	1
	mat race	1
27 BLACK	black	3
28 WHITE	white	3
29 INDIAN	indian	3
30 HAWAII	hawaii	3
31 ALASKA	alaska	3
32 PACIFIC	pacific	3

34	OTHER	other - Chinese, Japanese, Other Asian, Filipino, Other Non-White	3
35	OTH_TERM	history of other terminations	1
36	MM_DIAB	mat diabetes (prepregnancy & gestational)	1
37	MM_FEVER	mat fever	1
38	INSURE	insurance just before pregnancy	2
39	PREG_TRY	tried to get pregnant	2
40	VITAMIN	vitamin intake in the month before pregnancy	2
41	MOM_LB	weight in pounds just before pregnancy	2
42	BMI	body mass index just before pregnancy	3
43	PNC_MTH	month of first prenatal care visit	1
44	ERLY_PNC	access to prenatal care as early as wanted	2
45	MORB_NAU	morbidity: nausea, vomiting, dehydration during pregnancy	2
46	MORB_DIAB	morbidity: diabetes	2
47	MORB_KID	morbidity: kidney or bladder (urinary tract) infection	2
48	MORB_PLA	morbidity: problems with placenta	2
49	MORB_CRV	morbidity: cervix had to be sewn shut (incompetent cervix, cerclage	2
50	MORB_CAR	injury in car accident	2
51	MM_HBP	hypertension (prepregnancy, gestation, eclampsia)	1
52	MM_BLEED	vaginal bleeding	1
53	MM_NOMD	no medical risk factor	1
54	MORB_BP	high blood pressure during pregnancy	2
55	MORB_BLD	vaginal bleeding during pregnancy	2
56	STRS_FM3	stress: family	2
57	STRS_DV3	stress: separation or divorce from husband or partner	2
58	STRS_MOV	stress: moved to a new address	2
59	STRS_HOM	stress: homeless	2
60	STRS_JOB	stress: husband or partner lost job	2
61	STRS_WRK	stress: mother lost job	2
62	STRS_ARG	stress: argument	2
63	STRS_BIL	stress: bills that could not be paid	2
64	STRS_DH3	stress: someone very close died	2
65	STRS_FT4	stress: was in a physical fight	2
66	STRS_DRG	stress: problems with drinking or drugs	2
67	STRS_COM	stress: combination - partner did not want pregnancy and or problems with jail	3
68	PAD_COM	physical abuse by husband or other	3
69	INCOME	income	2

Table A.5: PAROUS DT splitting values to create attribute intervals for the prediction of HR-PTB. Threshold indicates the splitting value used by the DT. New Variables are new variables that were created if there is more than one splitting value present as outlined in Section 3.7.3. Prevalence was the percentage of positive cases (HR-PTB). Variables from 6 to 19 were left in their original format of -1 and +1.

Variable Name	Threshold	New Variable	Prevalence
1 PLURAL	≤ 1 1		4.8
2 MAT_AGE	≤ 18	MA1	1.8
	>18 and ≤ 24	MA2	25.7
	>24 and ≤ 37	MA3	65.5
	>37	MA4	7.1
3 MOM_CIG_later	≤ 2	MCL1	
	>2 and ≤ 14	MCL2	
	>14	MCL3	
4 DRK4_3L	1	DRK1	95.2
	2 3	DRK2	4.7
	4 5 6	DRK3	0.2
5 BMI	≤ 19.226 >19.226		90.4
6 MORB_PLA			8.7
7 PREV_PRE			16.1
8 MM_HBP			6.3
9 MORB_BLD			19.4
10 MORB_CRV			3.0
11 BLACK			18.4
12 HAWAII			3.3
13 PREV_LBW	No [-1] YES [+1]		14.1
14 STRS_DV3			10.1
15 MOMSMOKE			12.8
16 STRS_JL3			4.8
17 STRS_BIL			30.2
18 MM_DIAB			3.9
19 STRS_JOB			13.2

Table A.6: NULLIPAROUS DT splitting values to create attribute intervals for the prediction of HR-PTB. Threshold indicates the splitting value used by the DT. New Variables are new variables that were created if there is more than one splitting value present as outlined in Section 3.7.3. Prevalence was the percentage of positive cases (HR-PTB). Variables from 4 to 16 were left in their original format of -1 and +1.

Variable Name	Threshold	New Variable	Prevalence
1 PLURAL	≤ 11		2.8
2 MAT_AGE	≤ 17	MA1	9.1
	>17 and ≤ 28	MA2	62.8
	>28 and ≤ 33	MA3	18.3
	>33	MA4	9.8
3 OTH_TERM	0	OT0	81.0
	1	OT1	13.9
	2	OT2	3.2
	3	OT3	1.0
	4	OT4	0.4
	5	OT5	0.1
	6	OT6	0.1
	7	OT7	0.1
4 MORB_PLA			7.3
5 HAWAII			2.8
6 MM_HBP			10.8
7 MORB_CRV			3.0
8 MORB_BLD			18.1
9 BLACK			17.7
10 MM_NOMD	No [-1] YES [+1]		62.8
11 PACIFIC			3.6
12 INCOME			91.0
13 STRS_WRK			10.3
14 STRS_JOB			12.8
15 MORB_KID			18.8
16 SMK_2YRS			25.6

Table A.7: DT-ANN Hybrid starting parameters. Weight Elimination ON.

Learning Rate	0.00005
Learning Inc	1.01
Learning Dec	0.99
Lambda (weight-decay)	0.0001
Lamda_inc	1.001
Lambda_dec	0.999
Wnot	0.5
Momentum	1.02
Err_ratio	1.02

Table A.8: PAROUS Final Hybrid Network Parameters used to create the Hybrid Risk Stratification (RS) ANN.

Hidden nodes	4	8	9	11
Final learning rate	0.0005997	0.0002735	0.0004388	0.0000537
Learning rate increment	1.001	1.001	1.001	1.0207
Learning rate decrement	0.7828	1.0389	0.999	0.999
Final weight decay constant	0.0001	0.0001	0.0001	0.0000775
Weight decay constant increment	0.9083	1.001	0.8986	1.001
Weight decay constant decrement	0.999	0.999	0.999	0.999
Final weight scale (wo)	0.999	0.999	0.999	0.999
Momentum	0.257	0.5	0.008	0.5
Final error ratio	1.25	1.375	1.25	1.4375

Table A.9: NULLIPAROUS Final Hybrid Network Parameters used to create Hybrid Risk Stratification (RS) ANN.

Hidden nodes	5	11
Final learning rate	0.0004777	0.0003221
Learning rate increment	1.001	1.001
Learning rate decrement	0.9652	0.999
Final weight decay constant	00.0001	0.0000944
Weight decay constant increment	0.7656	0.7541
Weight decay constant decrement	1.25	1.25
Final weight scale (wo)	0.5269	0.4994
Momentum	0.5	0.5
Final error ratio	1.2638	1.375

Data Formatting

B.1 Data Pre-processing

The first step of knowledge discovery is data pre-processing, executed before creating a pattern classification model. The data analyst must ensure that relevant information is retained as information loss will result in decreased or erroneous performance. Though a formal definition of data pre-processing does not exist, the basic idea is to clean the “raw data”.⁵⁷ The first step defines and determines error types: errors that may exist in medical databases are data collection errors (i.e. measurement, improper format, misunderstanding of the question), and the presence of missing values and outliers [119]. Once errors have been identified, methods are taken to correct them. This involves the following data manipulation: deletion, transformation and aggregation [109]. Additional processing steps are normalization and reduction of the dimensionality of the input data.⁵⁸ Reduction of the data can occur by discarding a subset of the original inputs. Medical databases may also contain confounding variables (where the values essentially contain the same information); these are therefore removed [120]. All these data pre-processing methods were used to prepare the raw data.

B.1.1 Missing Values

Medical database may contain missing values for many reasons: human error, participants do not want to answer questions, technological error, or the irrelevance of the question. The following techniques were applied in this thesis work.

⁵⁷**Raw Data:** data that has not been analyzed for errors

⁵⁸**Normalization:** a linear or non-linear transformation that reduces the magnitude of the interval range

1. **Eliminate data objects or attributes:** Individual cases that are missing data may be eliminated. If too many values for a variable are missing, then the variable itself may be deleted. One has to be careful as loss of information may occur.
2. **Inconsistency may be corrected [106]:** Some data may be missing because the information does not apply to that question. For example the first question may ask “How many children have you had before your present birth?” If the answer is none (zero), then the response for the following question “Out of these children, how many were delivered prematurely”, may be missing. A value that a classifier recognizes should replace the missing value.
3. **Imputation of Missing Values Hybrid Method:** In medical databases, imputation using the mean/mode/max/min method may be invalid. Ennett *et al* propose a method of integrating Neural Networks (NN) and Case-based Reasoning (CBR). The CBR is used to retrieve cases that are most similar to the case of interest, while the NN uses the weights as a relative match weight of the CBR. The missing data is inputted by using the calculated mean of the missing parameter in the ten closest matching cases [121].

B.2 PRAMS Data Processing

The PRAMS database is sufficiently large so if inconsistencies could not be corrected, the cases were deleted. It was not necessary to attempt to replace missing values due to abundance of data.

PREVIOUS PRETERM BIRTH: Values of PREV_PRE were labelled so that 0 = did not occur; -1 = No; and 1 = YES. In order for No or Yes to be valid, the mother should have had previous children. This should match the variable PRE_LB (*number of previous live children - parity*). Parity should be greater than 0. If parity is 0, then previous preterm birth could not have occurred. PREV_LB (*did mother have previous live birth*) is related

to parity therefore this variable must match with parity. Parity should be greater than 0 if PREV_LB is +1 (yes).

Condition	Cases	Assumption
If PREV_PRE=1 and PRE_LB=0, then switch PREV_PRE to 0	741	PREV_PRE is labelled as “did not occur” since the mom had no previous children.
Select PRE_LB=0. Frequency of PREV_PRE are: (0=741 cases), (2=284 cases), (Missing=48104 cases)	48104	PREV_PRE have missing values because the question did not apply to the mother. Missing over 97.9% of values.
If PREV_PRE=missing & PRE_LB=0, then switch PREV_PRE=0.	48104	Switch all missing PREV_PRE cases to 0 where parity=0
	284	Ambiguity: If parity equals 0, then PREV_PRE could not have occurred. These cases are switched to PREV_PRE = 0
If PREV_PRE=no then PRE_LB should be >1	138	There are 138 cases missing for parity in this condition. List PREV_PRE as missing
If PREV_PRE=yes, then PRE_LB should be >1.	284	There are 284 parity cases listed as 0 and 29 missing cases. These PREV_PRE cases will be listed as missing
When PREV_PRE=yes and PRE_LB= yes there were 261 cases listed as PRE_LB=0.	261	These will be switched to missing cases and leaves 2809 cases as missing. These values are switched to missing because we cannot assume how many previous children a mother has had. It should be greater than 1, but this is unknown.
If PREV_PRE=0, then PRE_LB should equal 0. If PREV_LB=yes, then create these as missing cases	755	There are 755 missing cases for PREV_PRE and PRE_LB when PREV_LB=yes. Impossible to assume if PREV_PRE occurred. IF PREV_LB=no then PREV_PRE should be 0 and PRE_LB should be 0. This moved 2505 to PREV_LB=2(yes)
Now when PREV_LB=no, there were 201 missing cases for PRE_LB and PREV_PRE	201	Switch these both PRE_LB and PREV_PRE to 0

CIGARETTE AND ALCOHOL USE: The PRAMS database has many variables related to cigarette and alcohol use. Analysis of the data revealed ambiguity between the smoking variables. It also revealed ambiguity between the drinking variables. Further analysis concluded that this was because some variables originated from the PRAMS questionnaire and some from the birth certificate variables. As shown, below, the PRAMS questionnaire has more detail:

Variables from Birth Certificate:

MOMSMOKE (Did Mom Smoke)
 MOMCIG (How many cigarettes)
 MOMDRINK (Did mom drink during pregnancy)
 MOMDRKS (Number of Drinks)

Variables from PRAMS questionnaire:

SMK_2YRS (Have you smoked at least 100 cigarettes in the past 2 year?)
 SMK_3BEFORE (Did mom smoke 3 months before pregnancy)
 MOM_CIG_before (In the 3 months before you got pregnant, how many cigarettes or packs of cigarettes did you smoke on an average day)
 SMK_3Later (Did mom smoke in the last 3 months of pregnancy)
 MOM_CIG_later (In the last 3 months of your pregnancy, how many cigarettes or packs of cigarettes did you smoke on an average day)
 B_MOMDRK (Did mom drink 3 months before pregnancy)
 No_DRKS_Before (Average number of drinks before pregnancy)
 NO_BINGES_BEFORE (Number of binges 3 months before pregnancy)
 DRK43L_A (Did mom drink in the last 3 months of pregnancy)
 DRK4_3L (Average number of drink in last 3 months of pregnancy)
 DRK4_3LB (Number of binges in the last 3 months of pregnancy)

Questions from the birth certificate variables were not as specific as the PRAMS questionnaire. For example: the birth certificate variables do not reveal if maternal smoking was before, or during pregnancy. Additionally, it was noticed that the birth certificate variables do not seem to match that of the PRAMS questionnaire variables. Therefore, all confounding variables between the birth certificate variables and the PRAMS questionnaire were inputted into the decision tree.

ABUSE:

PAD_COM: combination of physical abuse of both husband or other

IF ((PAD_HUS=1) & (PAD_OTH=1)) PAD_COM=0 (No).

IF ((PAD_HUS=2) | (PAD_OTH=2)) PAD_COM=1 (Yes).

EXECUTE.

MATERNAL RACE: Convert to variables WHITE, BLACK, INDIAN, ALASKA, HAWAII, PACIFIC, OTHER (Chinese, Japanese, Filipino, Other non-white).

Modeling Cost Functions and Deriving a Training Algorithm

Zhou and Frize's work adds risk stratification estimates to the MLP ANN. In order to train the network using maximum likelihood estimation, the cost function should be modeled as a likelihood function. This occurs by modeling a cost function and deriving a training algorithm as described in Zhou's Master's thesis [43] and Zhou and Frize paper [16]. Let D be the training data D composed of patient records in the PRAMS database C.1:

$$\{ \langle x_1, y_1 \rangle, \langle x_2, y_2 \rangle, \dots, \langle x_m, y_m \rangle \} \quad (\text{C.1})$$

Where x_m represents the values of a set of input variables for the m^{th} patient and y_m corresponds to an outcome. The outcome can take a value of 1 (high-risk preterm birth) and 0 (low-risk preterm birth). Rewrite the likelihood function in the following form C.2:

$$L(\Theta) = f(\langle x_1, y_1 \rangle, \langle x_2, y_2 \rangle, \dots, \langle x_m, y_m \rangle; \Theta) = \prod f(x_i, y_i | \Theta) \quad (\text{C.2})$$

Where $f(x_i, y_i | \Theta)$ denotes the probability of y_i , given the input as x_i in the i^{th} patient record. The conditional probability that $y_i = 1$, given x_i will be denoted as $P(y_i = 1 | x_i)$. This value can be given by the neural network output $O(x_i)$. It follows that $O(y_i = 0 | x_i) = 1 - O(x_i)$. Next rewrite the $f(\langle x_i, y_i \rangle | \Theta)$ in the likelihood function as C.3:

$$f(x_i, y_i | \Theta) = O(x_i)^{y_i} [1 - O(x_i)]^{(1-y_i)} \quad (\text{C.3})$$

Assumptions are made that the observations of patient records are independent and the

likelihood function on the training data with m records is obtained as C.4:

$$\prod O(x_i)^{y_i} [1 - O(x_i)]^{(1-y_i)} \quad (\text{C.4})$$

Which can be rewritten in a log-likelihood form as C.5:

$$L(\Theta) = \ln [L(\Theta)] = \sum y_i \times \ln O(x_i) + (1 - y_i) \times \ln(1 - O(x_i)) \quad (\text{C.5})$$

The final equation is the logarithmic form of the likelihood function used as the cost function. Training of this ANN requires maximizing this likelihood function.

Mortality Results

POSTOP Hybrid Trial #1. Verification Rate is $\mu=3.04$, $\sigma=0.382$, $CI = 0.31$ where μ is mean, σ is standard deviation and CI is 99% confidence interval. (A) Train (B) Test (C) VERIFICATION.

Nodes		3			4			5			6			7		
Nodes		8			9			10			11					
		μ	σ	CI	μ	σ	CI	μ	σ	CI	μ	σ	CI	μ	σ	CI
A	Sensitivity	28.57			71.43			71.43			71.43			33.33		
	Specificity	99.49			92.64			92.64			92.64			95.18		
	LogSens	0.0415			0.3362			0.3362			0.3362			0.0552		
	AUC	0.9297			0.8500			0.9312			0.9321			0.8996		
B	Sensitivity	50.00			75.00			75.00			75.00			62.50		
	Specificity	98.33			89.96			89.54			89.12			95.40		
	LogSens	0.1469			0.3658			0.3627			0.3596			0.2462		
	AUC	0.9383			0.8426			0.9427			0.9417			0.8776		
C	Sensitivity	33.24	4.17	3.40	83.38	2.09	1.70	83.38	2.09	1.70	83.38	2.09	1.70	33.24	4.17	3.40
	Specificity	99.59	0.22	0.18	90.75	0.56	0.46	90.75	0.56	0.46	90.75	0.56	0.46	97.33	0.46	0.37
	LogSens	0.0593	0.0171	0.0139	0.5134	0.0406	0.0330	0.5134	0.04	0.0330	0.5134	0.0406	0.0330	0.0577	0.0166	0.013
	AUC	0.9646	0.0030	0.0024	0.9579	0.0027	0.0022	0.9586	0.00	0.0021	0.9586	0.0026	0.0021	0.9324	0.0071	0.005
Nodes		8			9			10			11					
A	Sensitivity	71.43			71.43			90.48			71.43					
	Specificity	92.64			92.64			84.52			92.64					
	LogSens	0.3362			0.3362			0.5685			0.3362					
	AUC	0.9319			0.8851			0.8603			0.9321					
B	Sensitivity	75.00			62.50			100.00			75.00					
	Specificity	89.12			89.54			84.10			89.54					
	LogSens	0.3596			0.2226			0.7986			0.3627					
	AUC	0.9417			0.9027			0.9119			0.9467					
C	Sensitivity	83.38	2.09	1.70	83.38	2.09	1.70	83.38	2.09	1.70	83.38	2.09	1.70			
	Specificity	90.75	0.56	0.46	90.75	0.56	0.46	86.48	0.93	0.76	90.75	0.56	0.46			
	LogSens	0.5134	0.0406	0.0330	0.5134	0.0406	0.0330	0.4635	0.0346	0.0282	0.5134	0.040	0.0330			
	AUC	0.9577	0.0026	0.0021	0.9270	0.0046	0.0038	86.40	0.90	0.7	0.9586	0.002	0.0021			

POSTOP Hybrid Trial #2. Verification Rate is $\mu=3.04$, $\sigma=0.382$, $CI = 0.31$ where μ is mean, σ is standard deviation and CI is 99% confidence interval. (A) Train (B) Test (C) VERIFICATION.

Nodes		3			4			5			6			7		
Nodes																
		μ	σ	CI	μ	σ	CI	μ	σ	CI	μ	σ	CI	μ	σ	CI
A	Sensitivity	71.43			38.10			52.38			66.67			38.10		
	Specificity	88.83			100.00			96.19			90.86			96.19		
	LogSens	0.3122			0.0793			0.1595			0.2695			0.0755		
	AUC	0.7399			0.9085			0.7492			0.8649			0.8864		
B	Sensitivity	75.00			37.50			62.50			75.00			50.00		
	Specificity	87.87			98.75			92.05			92.05			93.72		
	LogSens	0.3504			0.0753			0.2325			0.3819			0.1373		
	AUC	0.7163			0.9297			0.7798			0.8394			0.8577		
C	Sensitivity	100.00	0.00	#NUM	33.24	4.17	3.40	49.86	6.26	5.10	83.38	2.09	1.70	33.24	4.17	3.40
	Specificity	88.79	0.63	0.52	99.59	0.22	0.18	97.22	0.34	0.28	90.75	0.56	0.46	97.33	0.46	0.37
	LogSens	0.9512	0.0245	0.0200	0.0593	0.0171	0.0139	0.1481	0.047	0.038	0.5134	0.0406	0.0330	0.0577	0.0166	0.0135
	AUC	0.9624	0.0051	0.0041	0.9484	0.0070	0.0057	0.7405	0.032	0.026	0.9600	0.0026	0.0022	0.9324	0.0071	0.0058
Nodes		8			9			10			11					
Nodes																
		μ	σ	CI	μ	σ	CI	μ	σ	CI	μ	σ	CI	μ	σ	CI
A	Sensitivity	28.57			47.62			52.38			90.48					
	Specificity	100.00			97.21			95.69			85.03					
	LogSens	0.0418			0.1285			0.1582			0.5763					
	AUC	0.9037			0.7498			0.8501			0.9003					
B	Sensitivity	37.50			50.00			50.00			100.00					
	Specificity	98.75			93.72			91.63			82.85					
	LogSens	0.0753			0.1373			0.1331			0.7656					
	AUC	0.9218			0.7191			0.9124			0.9098					
C	Sensitivity	33.24	4.17	3.40	33.24	4.17	3.40	16.62	2.09	1.70	100.00	0.00	#NUM!			
	Specificity	99.59	0.22	0.18	97.74	0.35	0.28	96.71	0.49	0.40	86.53	0.85	0.69			
	LogSens	0.0593	0.0171	0.0139	0.0580	0.0168	0.0136	0.0129	0.003	0.002	0.8716	0.0277	0.0226			
	AUC	0.9479	0.0082	0.0067	0.7374	0.0322	0.0262	0.7615	0.018	0.015	0.9374	0.0078	0.0063			

POSTOP we ANN Trial #1. Verification Rate is $\mu=2.71$, $\sigma=0.2340$, $CI = 0.19$ where μ is mean, σ is standard deviation and CI is 99% confidence interval. (A) Train (B) Test (C) VERIFICATION.

Nodes	3			4			5			6			7		
A	Sensitivity	52.38		47.62		61.91		76.19		28.57					
	Specificity	100.00		55.58		75.64		100.00		79.70					
	LogSens	0.1688		0.0636		0.1698		0.4749		0.0321					
	AUC	0.8325		0.5160		0.7597		0.9236		0.5611					
B	Sensitivity	25.00		87.50		50.00		37.50		62.50					
	Specificity	99.58		56.49		79.50		98.75		79.92					
	LogSens	0.0311		0.2591		0.1100		0.0753		0.1879					
	AUC	0.6193		0.7199		0.6017		0.6645		0.7035					
C	Sensitivity	14.67	7.89	6.43	52.00	7.89	6.43	83.33	6.09	4.96	48.00	7.89	6.43	64.67	6.32
	Specificity	78.15	41.19	33.55	56.59	0.78	0.63	72.61	0.84	0.69	97.57	0.24	0.20	77.10	0.71
	LogSens	0.0127	0.0073	0.0059	0.0812	0.0261	0.0213	0.3416	0.07	0.0641	0.1383	0.0553	0.0451	0.1980	0.0540
	AUC	0.5699	0.0302	0.0246	0.5430	0.0412	0.0336	0.8021	0.04	0.0345	0.8913	0.0306	0.0249	0.8539	0.0167
Nodes	8			9			10			11					
A	Sensitivity	76.19		23.81		19.05		61.91							
	Specificity	26.40		98.73		56.60		83.50							
	LogSens	0.0743		0.0277		0.0094		0.1956							
	AUC	0.4721		0.6562		0.2291		0.8158							
B	Sensitivity	87.50		37.50		50.00		37.50							
	Specificity	31.80		99.16		59.00		80.75							
	LogSens	0.1239		0.0757		0.0759		0.0587							
	AUC	0.5829		0.6791		0.4383		0.4686							
C	Sensitivity	83.33	6.09	4.96	14.67	7.89	6.43	0.00	0.00	#NUM!	66.67	7.70	6.27		
	Specificity	26.36	0.97	0.79	98.61	0.25	0.20	0.00	0.00	#NUM!	78.81	1.05	0.86		
	LogSens	0.0903	0.0150	0.0123	0.0128	0.0073	0.0060	0.0000	0.000	#NUM!	0.2214	0.065	0.0532		
	AUC	0.5022	0.0434	0.0353	0.6486	0.0409	0.0333	0.1242	0.016	0.0137	0.7995	0.027	0.0222		

POSTOP we ANN Trial #2. Verification Rate is $\mu=2.71$, $\sigma=0.2340$, $CI = 0.19$ where μ is mean, σ is standard deviation and CI is 99% confidence interval. (A) Train (B) Test (C) VERIFICATION.

Nodes		3			4			5			6			7		
A	Sensitivity	μ	σ	CI	μ	σ	CI	μ	σ	CI	μ	σ	CI	μ	σ	CI
	Specificity	71.43			76.19			14.29			85.71			42.86		
	LogSens	99.49			100.00			97.72			100.00			59.90		
	AUC	0.3847			0.4749			0.0093			0.7244			0.0552		
B	Sensitivity	37.50			37.50			62.50			50.00			75.00		
	Specificity	99.16			100.00			97.07			99.16			66.53		
	LogSens	0.0757			0.0765			0.2533			0.1487			0.2251		
	AUC	0.7100			0.6797			0.7568			0.7390			0.7286		
C	Sensitivity	μ	σ	CI	μ	σ	CI	μ	σ	CI	μ	σ	CI	μ	σ	CI
	Specificity	83.33	6.09	4.96	0.00	0.00	#NUM	18.67	1.72	1.40	50.00	8.16	6.65	100.00	0.00	#NUM!
	LogSens	96.64	0.36	0.30	0.00	0.00	#NUM	94.83	0.47	0.38	98.09	0.24	0.20	56.95	0.67	0.54
	AUC	0.6394	0.2846	0.2318	0.0000	0.0000	#NUM	0.0159	0.003	0.0024	0.1540	0.0595	0.0485	0.3661	0.0067	0.0054
Nodes		8			9			10			11					
A	Sensitivity	28.57			0.00			76.19			57.14					
	Specificity	54.32			99.49			99.49			79.44					
	LogSens	0.0209			0.0000			0.4695			0.1502					
	AUC	0.4377			0.4975			0.8686			0.7072					
B	Sensitivity	87.50			25.00			25.00			50.00					
	Specificity	59.00			100.00			99.16			79.08					
	LogSens	0.2759			0.0312			0.0309			0.1093					
	AUC	0.7351			0.6250			0.7037			0.5704					
C	Sensitivity	μ	σ	CI	μ	σ	CI	μ	σ	CI	μ	σ	CI	μ	σ	CI
	Specificity	33.33	7.70	6.27	14.67	7.89	6.43	64.67	6.32	5.15	66.67	7.70	6.27	75.40	0.93	0.76
	LogSens	57.42	1.17	0.96	99.12	0.25	0.20	96.64	0.35	0.29	75.40	0.93	0.76	0.2075	0.0615	0.050
	AUC	0.0325	0.013	0.0110	0.0129	0.0074	0.0060	0.2841	0.088	0.0719	0.2075	0.0615	0.050	0.7712	0.0321	0.026

noPOSTOP Hybrid Trial #1. Verification Rate is $\mu=12.47$, $\sigma=1.15$ CI = 0.94 where μ is the mean, σ is the standard deviation and CI is the 99% confidence interval. (A) Train (B) Test (C) VERIFICATION

Nodes		3			4			5			6			7		
A	Sensitivity	77.14			88.57			88.57			85.71			65.71		
	Specificity	78.37			70.20			71.84			54.69			89.80		
	LogSens	0.3108			0.3740			0.3890			0.2355			0.2545		
	AUC	0.8412			0.8173			0.8164			0.6336			0.8265		
B	Sensitivity	63.64			86.36			90.91			72.73			40.91		
	Specificity	81.05			77.12			75.16			56.21			88.89		
	LogSens	0.2004			0.4114			0.4540			0.1660			0.0803		
	AUC	0.7910			0.8168			0.8333			0.5958			0.7019		
C	Sensitivity	74.45	3.81	3.10	83.60	2.86	2.33	89.50	2.59	2.11	67.40	2.36	1.92	36.47	3.37	2.74
	Specificity	81.21	1.50	1.23	74.56	1.25	1.02	73.58	1.05	0.85	47.40	1.49	1.22	86.12	0.76	0.62
	LogSens	0.3019	0.0403	0.0328	0.3562	0.0357	0.0291	0.4191	0.0372	0.0303	0.1127	0.0074	0.0060	0.0604	0.0121	0.009
	AUC	0.8454	0.0141	0.0115	0.8292	0.0151	0.0123	0.8354	0.0146	0.0119	0.4610	0.0172	0.0140	0.7118	0.0104	0.008
Nodes		8			9			10			11					
A	Sensitivity	74.29			91.43			85.71			77.14					
	Specificity	77.55			55.51			79.59			79.18					
	LogSens	0.2769			0.2813			0.4267			0.3162					
	AUC	0.7226			0.7454			0.8858			0.8306					
B	Sensitivity	63.64			95.46			81.82			77.27					
	Specificity	71.90			64.05			81.70			80.39					
	LogSens	0.1690			0.3919			0.3923			0.3258					
	AUC	0.6289			0.7943			0.8277			0.7897					
C	Sensitivity	63.84	5.50	4.48	89.50	2.59	2.11	79.13	3.02	2.46	68.30	6.05	4.93			
	Specificity	74.49	1.28	1.04	63.03	1.01	0.82	78.47	1.23	1.00	81.63	1.04	0.85			
	LogSens	0.1818	0.0396	0.0323	0.3238	0.0294	0.0239	0.3345	0.0334	0.0272	0.2466	0.0526	0.0428			
	AUC	0.7374	0.0166	0.0136	0.7552	0.0180	0.0147	0.8238	0.0128	0.0105	0.8292	0.0209	0.0170			

noPOSTOP Hybrid Trial #2. Verification Rate is $\mu=12.47$, $\sigma=1.15$ CI = 0.94 where μ is the mean, σ is the standard deviation and CI is the 99% confidence interval. (A) Train (B) Test (C) VERIFICATION

Nodes		3			4			5			6			7					
A	Sensitivity		62.86			85.71			82.86			88.57			82.86				
	Specificity		84.90			71.43			82.45			73.06			82.45				
	LogSens		0.2082			0.3527			0.4136			0.4007			0.4136				
	AUC		0.7841			0.8271			0.8929			0.8839			0.8632				
B	Sensitivity		54.55			90.91			86.36			95.46			86.36				
	Specificity		85.62			66.67			77.12			70.59			77.12				
	LogSens		0.1491			0.3678			0.4114			0.4644			0.4114				
	AUC		0.7127			0.8024			0.8437			0.8605			0.8552				
C	Sensitivity	μ	47.09	σ	3.09	CI	2.52	μ	74.20	σ	5.66	CI	4.61	μ	79.13	σ	3.02	CI	2.46
	Specificity		85.78		1.40		1.14		68.93		0.96		0.78		78.47		1.23		1.00
	LogSens		0.1064		0.0160		0.0130		0.2340		0.0436		0.0355		0.3345		0.0334		0.0272
	AUC		0.7565		0.0166		0.0135		0.7709		0.0204		0.0166		0.8125		0.0134		0.0109
Nodes		8			9			10			11								
A	Sensitivity		71.43			62.86			85.71			80.00							
	Specificity		77.96			87.76			75.92			75.92							
	LogSens		0.2525			0.2190			0.3916			0.3248							
	AUC		0.7226			0.8496			0.8110			0.8377							
B	Sensitivity		68.18			59.09			90.91			86.36							
	Specificity		71.24			84.97			69.94			72.55							
	LogSens		0.1969			0.1790			0.3988			0.3695							
	AUC		0.6340			0.7732			0.8102			0.8265							
C	Sensitivity	μ	63.84	σ	5.50	CI	4.48	μ	51.77	σ	2.10	CI	1.71	μ	89.50	σ	2.59	CI	2.11
	Specificity		74.49		1.28		1.04		84.54		0.97		0.79		76.48		1.44		1.18
	LogSens		0.1818		0.0396		0.0323		0.1297		0.0107		0.0087		0.4500		0.0413		0.0336
	AUC		0.7374		0.0166		0.0136		0.7700		0.0118		0.0096		0.8118		0.0140		0.0114

noPOSTOP we_ANN Trial #1 Verification Rate is $\mu=13.55$, $\sigma=0.837$ CI = 0.68 where μ is the mean, σ is the standard deviation and CI is the 99% .confidence interval. (A) Train (B) Test (C) VERIFICATION

Nodes		3			4			5			6			7		
A	Sensitivity	82.86			42.86			80.00			54.29			48.57		
	Specificity	100.00			90.20			51.84			88.16			83.67		
	LogSens	0.6346			0.0910			0.1861			0.1535			0.1100		
	AUC	0.9038			0.6559			0.7514			0.8313			0.7654		
B	Sensitivity	54.55			59.09			77.27			54.55			81.82		
	Specificity	95.43			88.24			56.21			90.85			83.01		
	LogSens	0.1741			0.1891			0.1912			0.1621			0.4039		
	AUC	0.7273			0.7583			0.7450			0.7616			0.8467		
C	Sensitivity	μ 43.27	σ 3.54	CI 2.89	μ 46.91	σ 4.12	CI 3.36	μ 76.78	σ 3.13	CI 2.55	μ 63.77	σ 2.67	CI 2.18	μ 54.96	σ 3.63	CI 2.96
	Specificity	94.83	0.60	0.49	88.65	0.69	0.56	51.96	1.29	1.05	87.97	0.44	0.36	83.64	0.91	0.74
	LogSens	0.1005	0.0202	0.0165	0.1108	0.0231	0.0188	0.1704	0.0189	0.0154	0.2289	0.0250	0.0204	0.1481	0.0232	0.018
	AUC	0.7194	0.0228	0.0186	0.6958	0.0228	0.0186	0.6602	0.0262	0.0213	0.8121	0.0194	0.0158	0.7844	0.0129	0.010
Nodes		8			9			10			11					
A	Sensitivity	57.14			85.71			65.71			77.14					
	Specificity	93.06			100.00			81.22			67.76					
	LogSens	0.1883			0.7244			0.2178			0.2478					
	AUC	0.8707			0.9187			0.8376			0.7772					
B	Sensitivity	50.00			54.55			81.82			72.73					
	Specificity	91.50			92.81			82.35			72.55					
	LogSens	0.1328			0.1672			0.3981			0.2369					
	AUC	0.8252			0.7512			0.8417			0.7854					
C	Sensitivity	μ 60.94	σ 5.57	CI 4.54	μ 50.21	σ 3.48	CI 2.83	μ 52.28	σ 4.82	CI 3.93	μ 76.21	σ 3.64	CI 2.96			
	Specificity	88.21	1.02	0.83	92.44	0.94	0.77	81.87	1.10	0.89	67.21	0.88	0.72			
	LogSens	0.2079	0.0467	0.0380	0.1370	0.0219	0.0178	0.1289	0.0285	0.0232	0.2386	0.0273	0.0222			
	AUC	0.7943	0.0182	0.0148	0.7639	0.0194	0.0158	0.8036	0.0159	0.0129	0.7475	0.0230	0.0187			

noPOSTOP we_ANN Trial #2. Verification Rate is $\mu=13.55$, $\sigma=0.837$ CI = 0.68 where μ is the mean, σ is the standard deviation and CI is the 99% confidence interval. (A) Train (B) Test (C) VERIFICATION

Nodes		3			4			5			6			7		
A	Sensitivity	57.14			62.86			80.00			91.43			31.43		
	Specificity	100.00			89.80			62.45			99.59			70.61		
	LogSens	0.2103			0.2269			0.2405			0.9586			0.0343		
	AUC	0.7700			0.8639			0.7984			0.9820			0.5014		
B	Sensitivity	36.36			45.46			54.55			50.00			54.55		
	Specificity	93.46			88.89			58.82			91.50			79.09		
	LogSens	0.0656			0.1022			0.0917			0.1328			0.1337		
	AUC	0.6289			0.7466			0.6310			0.6812			0.7170		
C	Sensitivity	μ 36.63	σ 3.84	CI 3.13	μ 46.29	σ 4.95	CI 4.03	μ 76.78	σ 3.13	CI 2.55	μ 60.98	σ 4.41	CI 3.59	μ 31.83	σ 2.37	CI 1.93
	Specificity	93.14	0.46	0.38	89.31	0.89	0.72	60.34	0.60	0.49	92.63	0.85	0.70	67.70	1.40	1.14
	LogSens	0.0674	0.0153	0.0125	0.1092	0.0256	0.0208	0.2083	0.0218	0.0177	0.2229	0.0403	0.0328	0.0337	0.0049	0.004
	AUC	0.6361	0.0197	0.0160	0.8060	0.0176	0.0143	0.6722	0.0204	0.0166	0.8096	0.0184	0.0150	0.5437	0.0119	0.009
Nodes		8			9			10			11					
A	Sensitivity	45.71			68.57			77.14			74.29					
	Specificity	51.43			85.31			78.37			71.02					
	LogSens	0.0532			0.2619			0.3108			0.2419					
	AUC	0.4353			0.8635			0.7939			0.7897					
B	Sensitivity	59.09			63.64			59.09			72.73					
	Specificity	57.52			83.01			79.74			67.97					
	LogSens	0.1066			0.2076			0.1635			0.2154					
	AUC	0.5410			0.7144			0.6940			0.7556					
C	Sensitivity	μ 51.36	σ 3.71	CI 3.02	μ 71.25	σ 4.84	CI 3.95	μ 61.56	σ 3.59	CI 2.92	μ 76.21	σ 3.64	CI 2.96			
	Specificity	49.59	1.58	1.29	82.13	1.22	1.00	75.75	1.10	0.90	69.24	0.95	0.78			
	LogSens	0.0661	0.0111	0.0091	0.2756	0.0498	0.0405	0.1691	0.0249	0.0203	0.2493	0.0287	0.0234			
	AUC	0.4729	0.0196	0.0160	0.7839	0.0273	0.0222	0.6936	0.0260	0.0212	0.7457	0.0225	0.0183			

**Hybrid POSTOP BIAS and Weights -
Classification ANN Best Performing Classifier Node 11 (Trial #1)**

W1 =

-0.8520	-0.3887	-0.6639	-0.7751	0.6070
-0.4532	0.0416	0.1542	0.0122	0.4637
-0.4782	1.0417	-0.8641	1.0895	-0.2526
0.4353	0.8289	-0.1101	0.5931	1.0013
0.6453	0.5926	-0.5065	-0.7452	0.4229
0.3122	-0.8871	-0.9880	-0.6492	0.4652
-0.8276	-0.7668	0.5776	-1.0107	-0.8452
-0.2077	0.1027	-0.8904	0.8962	-0.4059
0.2605	-0.2670	-0.8364	-0.2527	-0.7471
0.3525	0.1039	-0.0603	0.7132	-0.1848
-0.3515	-1.2191	-0.8498	0.7813	-0.4092

B1 =

0.6821
-0.3142
0.2845
0.1261
0.2712
-0.4051
-0.1520
0.5768
0.0349
-0.3567
0.0136

W2 =

[1.2457 0.1448 1.0980 -0.0172 -0.0611 0.6262
0.2329 0.4936 -0.3661 -0.3510 -0.9277];

B2 = 0.3022

**Hybrid noPOSTOP BIAS and WEIGHTS –
Classification ANN Best Performing Classifier Node 5 (Trial #1)**

W1 =

1.6431	-0.7047	1.6520	-0.4943	1.8931
-7.4075	-2.1347	-5.7403	0.8916	-8.0259
3.7299	-4.9195	4.4013	-5.5596	-6.2706
-15.5603	16.5625	14.2341	16.5184	-13.7109
6.1893	-0.2032	6.0757	-5.6565	-2.5799

B1 =

-2.0370
6.3463
-4.5637
16.9593
-6.3421

W2 =

-9.1573 6.9172 -6.4414 5.2113 -2.9871

B2 =

-20.3712

High-Risk Preterm Birth Results

PAROUS HYBRID: Verification Rate is $\mu=16.13$, $\sigma=0.1391$, $CI = 0.11$ where μ is mean, σ is standard deviation and CI is 99% confidence interval. (A) Train (B) Test (C) VERIFICATION

Nodes		3			4			5			6			7		
A	Sensitivity	73.66			66.04			74.06			69.70			71.98		
	Specificity	78.15			83.33			76.71			82.02			79.66		
	LogSens	0.2743			0.2292			0.2700			0.2567			0.2663		
	AUC	0.7954			0.8055			0.8023			0.8105			0.8017		
B	Sensitivity	73.90			66.34			71.51			66.14			67.73		
	Specificity	78.46			83.92			76.17			81.51			78.80		
	LogSens	0.2783			0.2343			0.2444			0.2225			0.2244		
	AUC	0.7989			0.7949			0.7901			0.7958			0.7692		
C	Sensitivity	μ 69.79	σ 4.22	CI 3.44	μ 61.42	σ 4.52	CI 3.68	μ 68.20	σ 5.49	CI 4.47	μ 63.65	σ 4.53	CI 3.69	μ 61.68	σ 5.75	CI 4.68
	Specificity	78.79	1.89	1.54	83.31	2.15	1.75	76.37	2.40	1.96	82.06	2.20	1.79	78.31	2.26	1.84
	LogSens	0.2434	0.0346	0.0282	0.1932	0.0340	0.0277	0.2204	0.0393	0.0320	0.2052	0.0301	0.0246	0.1795	0.0388	0.031
	AUC	0.7930	0.0255	0.0208	0.7906	0.0171	0.0139	0.7891	0.0230	0.0187	0.7954	0.0170	0.0138	0.7563	0.0266	0.021
Nodes		8			9			10			11					
A	Sensitivity	68.52			71.09			72.08			73.76					
	Specificity	83.20			82.13			15.35			77.60					
	LogSens	0.2512			0.2707			0.0367			0.2722					
	AUC	0.8161			0.8300			0.3592			0.8138					
B	Sensitivity	65.14			67.93			72.31			72.51					
	Specificity	83.88			81.17			17.23			76.93					
	LogSens	0.2236			0.2365			0.0418			0.2570					
	AUC	0.8077			0.8062			0.3791			0.8047					
C	Sensitivity	μ 61.80	σ 4.50	CI 3.66	μ 64.43	σ 4.84	CI 3.94	μ 70.47	σ 3.34	CI 2.72	μ 69.92	σ 3.44	CI 2.80			
	Specificity	82.96	2.51	2.05	81.05	2.33	1.90	16.39	1.09	0.89	77.43	1.66	1.35			
	LogSens	0.1944	0.0313	0.0255	0.2082	0.0339	0.0276	0.0376	0.0039	0.0032	0.2378	0.0287	0.0234			
	AUC	0.7970	0.0256	0.0209	0.7983	0.0112	0.0091	0.3749	0.0218	0.0178	0.8001	0.0214	0.0174			

NULLIPAROUS HYBRID: Verification Rate is $\mu=17.72$, $\sigma=2.4521$, $CI = 2.00$ where μ is mean, σ is standard deviation and CI is 99% confidence interval. (A) Train (B) Test (C) VERIFICATION

Nodes		3			4			5			6			7		
A	Sensitivity	μ	σ	CI	μ	σ	CI	μ	σ	CI	μ	σ	CI	μ	σ	CI
	Specificity	65.66	69.85		65.18	64.04		62.77	71.71		25.18	95.86		55.06	79.13	
	LogSens	0.1750	0.1750		0.1529	0.1529		0.1631	0.1631		0.0302	0.0302		0.1368	0.1368	
	AUC	0.7227	0.7227		0.6972	0.6972		0.7201	0.7201		0.7388	0.7388		0.7255	0.7255	
B	Sensitivity	65.52	68.24		68.97	65.18		65.03	71.25		23.89	94.34		50.25	77.53	
	Specificity	68.24	0.1686		65.18	0.1788		71.25	0.1757		94.34	0.0265		77.53	0.1077	
	LogSens	0.1686	0.1686		0.1788	0.1788		0.1757	0.1757		0.0265	0.0265		0.1077	0.1077	
	AUC	0.7137	0.7137		0.7133	0.7133		0.7127	0.7127		0.7263	0.7263		0.6834	0.6834	
C	Sensitivity	μ	σ	CI	μ	σ	CI	μ	σ	CI	μ	σ	CI	μ	σ	CI
	Specificity	67.66	5.25	4.27	70.27	5.31	4.33	65.51	4.77	3.88	27.22	6.42	5.23	53.62	8.13	6.63
	LogSens	0.1872	0.0352	0.0286	0.1880	0.0308	0.0251	0.1799	0.0311	0.0253	0.0377	0.0202	0.0165	0.1310	0.0399	0.0325
	AUC	0.7305	0.0298	0.0243	0.7203	0.0263	0.0214	0.7310	0.0301	0.0245	0.7460	0.0267	0.0218	0.7062	0.0327	0.0266
Nodes		8			9			10			11					
A	Sensitivity	60.36	73.00		56.75	79.70		58.31	77.82		60.24	74.12				
	Specificity	73.00	0.1523		79.70	0.1484		77.82	0.1531		74.12	0.1548				
	LogSens	0.1523	0.1523		0.1484	0.1484		0.1531	0.1531		0.1548	0.1548				
	AUC	0.7238	0.7238		0.7276	0.7276		0.7389	0.7389		0.7213	0.7213				
B	Sensitivity	62.07	71.51		55.67	77.79		58.62	76.49		61.33	73.64				
	Specificity	71.51	0.1582		77.79	0.1372		76.49	0.1515		73.64	0.1600				
	LogSens	0.1582	0.1582		0.1372	0.1372		0.1515	0.1515		0.1600	0.1600				
	AUC	0.6988	0.6988		0.7083	0.7083		0.7173	0.7173		0.7237	0.7237				
C	Sensitivity	μ	σ	CI	μ	σ	CI	μ	σ	CI	μ	σ	CI			
	Specificity	61.76	6.10	4.97	55.82	5.67	4.62	57.78	5.51	4.49	63.54	5.28	4.30			
	LogSens	0.1623	0.0373	0.0304	0.1412	0.0334	0.0272	0.1506	0.0339	0.0276	0.1786	0.0363	0.0296			
	AUC	0.7212	0.0298	0.0243	0.7191	0.0315	0.0256	0.7359	0.0319	0.0260	0.7358	0.0341	0.0278			

PAROUS Hybrid Risk Stratification ANN - Classifier Node 4

GW1 =

Columns 1 through 17

0.9443	-0.5159	0.1453	-0.0382	-0.7709	0.8977
-0.4182	-0.6032	0.2691	-0.3349	0.3766	-0.2840
-0.7764	0.1213	-1.0070	0.7950	0.2266	-0.6553
-0.8745	-0.1564	0.4017	0.2138	-0.6055	0.8334
-0.3825	0.3941	-0.0354	0.8458	-0.9348	-0.2518
-0.5538	0.5369	0.4200	-0.1020	-0.2685	-0.0816
0.7673	0.6510	-1.2112	-0.2306	0.4709	-0.1738
-0.5355	-0.2947	0.1916	-0.1053	-0.1276	0.6059
-1.0510	-0.0154	-0.1070	1.0976	0.6897	0.5874
0.3424	-1.4511	-0.7637	-0.4548	-0.2605	0.1544
0.1566	0.7044	0.5685	0.7473	-0.0989	-0.4297
0.9724	-0.1115				

Columns 18 through 26

0.6229	0.8737	-0.0676	-0.7772	0.8681	0.9356
0.5054	-1.0069	-0.4603	0.6031	0.3054	0.7560
-0.8539	0.0228	0.5059	-0.6002	-0.2799	-0.6581
0.2541	0.4285	-0.8405	-0.8372	-0.8478	-0.0488
0.2008	-0.3219	1.2048	-1.2526	-0.0123	-1.1091
-0.5987	-0.6653	-0.0838	-0.0340	-0.0951	-0.0404

GB1 =

0.6149
0.6018
-0.9893
-1.4987

GW2 = -0.0067 0.1350 -1.1964 -3.3474

GB2 = 0.9114

PAROUS Hybrid Risk Stratification ANN - Classifier Node 8

GW1 =

Columns 1 through 17

0.7944	-0.6620	-0.5340	0.3045	-0.5162	-0.1168	-.7072
-0.9888	-0.5784	0.9146	-0.7528	0.8483	-0.7654	0.3998

0.7435	0.1171	0.9518	-0.3339	0.2220	-0.3886	0.9034
-0.1066	0.8423	0.9224	0.0102	0.7838	-0.2629	-1.1408
-0.0364	0.8714	0.1858	0.1353	0.4725	-0.8824	0.2167
0.5207	-0.4800	0.2983	-1.2298	-0.8973	-0.5639	-1.0355
0.6111	0.8605	-0.1098	0.8667	0.8436	-0.0652	-0.6256
1.6049	-0.4802	0.5532	0.4370	-0.3811	-0.0357	2.5661
1.8632	1.3046	0.2737	-0.7710	0.0285	-0.3929	0.0652
-0.4984	-0.2757	1.8334	-0.6204	0.1307	-0.6101	0.3549
0.7907	-0.0517	0.2825	0.9531	0.0909	0.5182	0.3078
0.2028	0.9639	0.0469	-0.7119	-0.1488	0.5806	-1.0655
-0.3391	-1.0022	-0.0652	-0.5659	0.2983	2.8939	-0.7796
-0.5311	-0.2764	0.3872	0.1165	-0.6971	-0.3441	-0.7473
-0.2079	0.4377	-1.2985	-0.0105	0.8220	0.0029	0.8422
-0.3021	-1.2995	-1.6753	-0.0019	-1.0784	1.4570	0.0972
-1.1238	0.7582	1.6290	0.4077	-0.7705	0.7900	0.4736
-0.0289	-0.1409	-0.2642	-0.1113	0.1895	0.8451	0.1356
-0.6577	-0.0538	-0.0883	-0.0176	0.4999	0.2711	0.3772
-1.6249	0.6491	0.3681				

Columns 18 through 26

-0.8388	-0.8089	1.0172	-0.3795	-0.6080	0.7865	0.4349
-0.7163	0.3583	-1.0092	-0.1787	1.2568	0.6848	1.7987
0.6880	0.8444	-0.5500	-0.1352	-0.8535	-0.7563	0.0860
-0.0428	-1.4538	0.3360	0.0154	-0.4788	0.2679	-0.9892
0.4817	1.2561	1.2057	1.7541	-0.0974	-0.3155	0.2527
-0.8915	1.9877	-0.2417	4.1115	0.7737	0.4176	-0.3865
-0.5213	0.5249	0.2506	1.7406	0.0507	-0.0761	0.5194
0.6326	0.5281	0.2301	0.2712	-0.4838	-1.1519	-0.2211
0.5656	0.2707	-0.2667	0.9103	-0.1395	-0.6757	0.6105
0.1650	-0.4145	-0.4146	0.3342	-0.9052	1.3405	0.6610
0.9564	0.4801					

GB1 =

-1.0293
 0.9332
 -1.4190
 1.2944
 1.0993
 -0.4566
 0.1080
 -0.9651

GW2 = 0.3379 1.2525 -0.5492 1.8406 1.9332 1.5409
 -0.2165 0.1100

GB2 = -1.3219

PAROUS Hybrid Risk Stratification ANN - Classifier Node 9

GW1 =

Columns 1 through 17

-0.8182	0.0559	0.0530	0.3747	-0.6456	0.6233	0.9114
-0.7149	0.8037	-0.7215	0.7059	0.6202	-0.6056	-0.8430
-0.8337	0.6266	-1.0587	0.0820	-0.3459	-0.1453	-0.1490
-0.6578	0.4684	0.8119	0.0443	-0.3980	0.3104	-0.6856
-0.4390	0.3076	0.5051	0.5282	-0.0848	-0.3299	0.3624
-0.1795	-0.2003	0.4102	0.5868	0.8971	-0.4779	0.1271
-0.9811	-0.6085	0.8305	-0.7800	0.2210	-0.4677	0.0138
-1.0585	0.3698	-0.4385	-0.1931	0.0280	-0.4052	0.2249
0.7746	1.0416	0.0209	-0.3768	0.6435	0.1597	0.1887
-0.5110	-0.8639	0.7483	0.0321	-0.2842	-0.4425	-0.4069
-0.8633	0.0590	-0.5984	0.0041	0.2609	0.7995	-0.8303
0.4356	0.2586	-0.5045	0.1057	-0.5912	0.5935	0.5263
0.3827	0.9691	-0.1023	0.3833	-0.3349	-1.5619	-0.5682
-0.3680	0.6400	-0.6046	1.1177	-0.6809	-0.8934	-0.4222
-0.2732	0.0050	1.1751	0.7671	0.6310	0.5449	0.8308
0.4589	-1.7477	0.7156	0.9249	0.1382	0.8601	-0.3095
-0.0074	-1.1122	0.8171	0.7271	-0.7179	-0.7419	-0.7760
0.3602	-0.5018	-0.2888	0.3718	-0.9107	-0.7865	0.4478
-0.8399	-0.3940	-0.1624	-0.8311	0.5281	-0.6719	-0.8013
-0.5497	-0.5571	0.2786	0.6322	-0.2979	0.5778	0.2749
-0.4618	0.4832	-0.9267	0.2990	-0.5901	0.8162	-0.5390
-0.6712	0.8792	-0.2641	0.7711	-0.5213	0.7938	

Columns 18 through 26

0.1605	-0.6107	0.7358	-1.0177	-0.8458	-0.7135	0.0016
-0.4435	-0.2931	-1.0033	0.4449	0.9477	1.1378	1.1956
0.3941	0.0058	-0.2129	-0.1345	1.0074	0.2839	0.6558
0.7493	0.1438	0.0747	0.4595	-0.3711	0.7515	0.4751
-0.6625	1.5081	0.3930	0.2556	-0.9524	-0.1755	0.3053
-0.7200	0.1461	-0.3193	-0.8323	0.3095	0.7245	0.8302
1.0254	0.1088	-0.8877	-1.0171	-0.1134	0.0206	.4431
-0.0178	0.2669	0.1266	-0.6028	0.5150	-0.8958	0.6995
-0.4114	0.6369	-0.7169	-0.7829	0.0886	0.1801	0.5317
-0.5540	0.0406	-0.0443	0.7680	-0.6279	0.0875	-0.6208
-0.1692	0.5884	-0.1622	0.2098	0.9123	-0.1525	-0.0374
0.6181	-0.1506	0.8216	0.8445			

GB1 =

-0.8327
 -0.0119
 0.8118
 0.7905
 0.9423
 0.9264
 -0.7201
 -0.1989
 -0.2388

GW2 = -0.7960 1.5700 1.4149 1.6366 -0.4215 -1.5065
 -0.9892 -0.0546 0.0273

GB2 = -0.0818

PAROUS Hybrid Risk Stratification ANN - Classifier Node 11

GW1 =

Columns 1 through 17

-0.1385	-0.3798	-0.6207	0.3556	0.2619	0.9235	-0.1651
-0.5884	-0.2851	0.1926	0.2899	0.5649	0.8284	0.0245
0.7954	0.1632	0.3943	0.0654	0.6890	-0.8890	0.8691
0.6943	1.4138	0.8994	-0.3874	1.0252	0.0215	-1.0271
-0.0596	0.9946	1.0494	-0.0172	-0.3409	0.4988	-0.0200
-0.5114	0.5097	-0.1011	1.4591	0.8761	-0.3592	0.2668
0.4830	0.5531	-1.1705	0.0967	1.0806	1.0187	-0.3812
-0.6297	0.7014	1.4091	-0.3226	-1.1007	-0.1460	2.5046
1.4161	0.7211	-1.1671	0.2501	0.7668	-0.5718	0.2955
0.7894	0.7703	1.8069	1.0717	0.8487	0.3439	-0.3758
-0.1713	-0.8504	-1.4403	-0.9132	-0.8686	-0.3040	0.1930
-1.1510	0.0522	0.3477	-0.0171	-0.1877	-0.4861	-0.2966
-1.2386	0.9561	0.1310	-0.5388	1.4317	0.2276	-0.4365
0.5977	-1.0036	0.1517	0.3363	0.1751	0.3482	0.2712
0.6564	1.2473	0.5268	1.6381	1.3203	0.4546	0.8232
0.0401	-0.5003	-0.2763	0.4479	-1.1964	0.3932	1.4882
0.5338	-0.1617	0.9114	-0.0937	-0.5010	0.9210	0.3828
1.2371	0.9286	1.2938	0.5404	-3.5449	-2.4634	-1.6030
-0.9773	0.8477	1.6978	0.5473	0.4495	2.5690	0.5295
-1.6023	2.7858	0.3610	-1.2408	0.3130	-0.6635	-0.1587
1.5801	0.6331	0.7341	0.9085	0.7601	0.1062	-0.6240
0.0916	0.3232	-0.3238	1.5146	-1.0676	-0.1448	0.9983
0.3232	-0.2245	1.0759	-1.0906	-1.2228	-0.4115	-0.4547

1.0743	0.8875	0.5103	0.1352	0.0263	0.9278	-1.0446
1.8632	0.7045	-0.1551	0.5968	0.2647	0.8652	-0.1319
0.4031	-0.0809	0.3418	0.3579	1.3777	-0.5482	1.1745
0.5253	0.9878	0.8632	-0.2099	0.6874		

Columns 18 through 26

1.2746	-0.1826	0.3599	-0.7656	-0.4177	-0.7160	0.5486
0.4684	0.9953	1.5290	-0.1320	1.6055	1.5456	1.0578
0.8459	0.1267	0.8165	0.2743	0.0584	-0.2076	1.1133
-0.5659	-0.5516	0.7376	0.0766	-0.3984	-0.3091	1.0046
-1.2790	1.8965	0.2979	1.4336	-0.0184	0.7809	0.6543
1.1364	-1.7921	0.2668	-1.1847	-1.2443	-0.5425	-0.3090
-0.6852	-0.6154	0.2965	-0.4611	0.1694	0.2994	-0.0676
0.1305	0.6354	0.5381	-0.5160	1.2388	-0.4208	-0.5770
-0.1431	-0.7527	-0.6430	-0.0675	0.0089	0.3065	0.8836
-2.6202	-0.6292	-3.4434	-0.1982	-3.1236	-0.0197	0.5007
-0.9889	1.3762	-1.0171	-0.2601	2.6143	1.7819	0.0237
0.1363	0.2580	-0.2933	0.0088	-0.2014	-0.8582	-0.7899
0.3926	-1.1633	-0.7494	0.2599	0.6574	-0.0735	0.0832
0.3716	0.0942	0.2753	-0.5647	0.8881	0.9801	-0.3554
0.9055						

GB1 =

0.6735
 -0.7058
 -1.0420
 -1.1752
 1.0626
 -0.1682
 0.5001
 -2.0790
 -0.6261
 -0.5491
 -0.4946

GW2 =

1.1899	-0.5416	0.6932	-0.2357	0.5598	-0.9209	-
0.4491	-1.7671	2.2761	-0.4705	-0.6159		

GB2 = 0.0979

NULLIPAROUS Hybrid Risk Stratification ANN - Classifier Node 5

GW1 =

Columns 1 through 17

0.0745	0.4683	0.1347	-0.0956	-1.2187	-0.1909	-0.2226
0.3641	0.5348	0.3092	-0.3251	0.0894	-0.8374	-1.0612
0.3996	1.1317	-0.6120	-0.3496	-0.0669	-0.2380	-0.5434
-0.1880	1.0111	0.9329	-0.9882	-0.1953	-0.5863	0.0930
-0.4965	-0.8077	-0.6069	0.1144	0.1479	-0.9075	0.5177
-0.7565	-0.9173	-0.1548	-0.7290	0.5579	-0.2204	0.5180
-0.0434	0.9783	0.7069	-0.8131	0.6811	-0.3663	0.9341
-0.2266	0.0189	-0.2017	-0.4722	0.0831	0.4280	1.0246
-0.2685	-0.5824	-0.5209	0.0736	0.8068	-0.1825	0.2905
0.7423	-0.9204	0.1911	0.6450	0.1799	-0.8005	-0.5652
0.1104	-0.4167	-1.5881	-0.2506	-0.7801	0.1122	0.8641
-0.6733	-0.4556	0.6692	0.7115	0.3940	-0.4095	-0.3306
0.9206						

Columns 18 through 26

-0.6166	-0.1948	-0.8992	-0.0806	-0.0383	-1.7013	0.6878
-0.0757	-0.3058	0.6916	0.3013	-0.6667	0.8612	0.4796
0.2226	-0.6882	0.5763	-0.3601	-0.6328	-0.5548	0.5210
-0.5580	1.5369	-0.1122	-0.0513	-0.5834	-0.1105	-0.3480
0.0940	0.9872	0.6016	0.1666	0.0768	0.5672	-0.7299
-0.5426	-0.3378	0.4819	-0.7986	-0.8273	-0.1717	-1.2491
-0.3295	-0.0676	0.1815				

GB1 =

0.4701

0.2930

-0.4650

-0.0149

0.4423

GW2 = -1.6090 0.8192 -1.3362 1.4679 -1.3552

GB2 = 0.2866

NULLIPAROUS Hybrid Risk Stratification ANN - Classifier Node 11

GW1 =

Columns 1 through 6

0.2457	-0.4509	-0.3825	0.2248	-0.2546	0.4894
-0.4933	0.0526	-0.7608	0.6079	0.1169	0.7106
-0.1887	-0.2733	0.2994	-0.4311	1.7572	0.3764
0.6355	-0.9222	-0.7247	-0.9664	0.5306	0.4709
0.9009	0.1027	-0.6497	-0.5224	-0.7179	0.1756
0.0418	-0.4247	-0.5812	0.9126	-0.7781	-0.3044
0.9180	0.5006	0.7512	-0.4032	-0.5721	0.3333
0.1375	-0.0886	0.8479	-0.3642	-0.5934	0.0344
-0.8126	-0.1883	-0.0499	-0.3771	-0.6910	-0.3884
0.3066	-0.1501	-0.3328	0.7446	-0.2731	-0.3434
-0.6982	0.3525	0.2025	0.2360	-1.0206	0.5932

Columns 7 through 12

-0.6559	-0.5947	-0.2516	0.3601	-0.0300	0.9344
0.6201	-0.0551	0.4344	-0.8324	-0.4049	-0.4696
-0.1140	-0.1603	-0.4566	0.1163	-0.7640	-0.0815
-0.0663	-0.4771	-0.3999	0.1040	0.0148	-0.8091
-0.4685	-0.3213	0.6771	-0.5787	-0.4102	0.7671
0.3126	-0.0916	-0.2178	-0.7251	0.6014	-0.1362
0.7262	-0.6425	0.0898	0.9683	0.6240	-0.3900
-1.0573	1.2549	-0.3752	-0.0393	0.6965	0.0706
0.5718	0.9680	0.6215	0.4598	-1.0195	0.2458
0.0015	0.0513	0.5192	0.0439	0.7336	-0.2678
-0.7129	1.0117	0.3467	0.7042	-0.4754	0.8040

Columns 13 through 18

1.0201	0.2956	0.2924	0.5126	0.5075	0.4279
0.3623	0.6511	-0.7397	-0.9772	-0.0729	0.6601
0.7959	1.0289	-0.6253	-0.9261	0.4318	-0.3105
0.1136	-0.1153	0.7100	0.3717	0.0770	-0.5296
0.2774	0.2255	0.4218	0.0474	-0.8349	-0.5623
-0.6669	0.1630	1.0719	-0.5289	0.8881	-1.0434
0.6117	-0.3401	-0.2198	0.2918	0.0393	-0.1212
1.0864	0.1656	-0.7477	0.8256	-0.3162	-0.1909
0.6139	-0.1093	-0.1348	-0.7752	0.1166	-1.3002
-0.5483	0.7559	-0.2869	0.8221	0.1658	0.6140
0.0146	0.9251	0.9663	-0.9827	0.1355	-0.4117

Columns 19 through 24

-0.0680	0.1733	-0.6291	-0.3961	-0.6079	0.8928
0.4192	1.1896	0.7804	0.2512	0.4687	-0.3071
0.2501	0.9469	-0.1215	-0.1211	1.1285	0.2400
-0.3301	0.9402	-1.1891	1.1648	-0.9740	-0.1598
-0.0914	-0.4726	-0.7207	0.0692	-0.2139	-0.3004

0.9061	-0.1555	-1.0807	0.4095	-0.6055	-0.0072
-0.5235	-0.0427	-1.2463	-0.7399	-0.1988	-0.2456
-0.7533	-0.2066	0.0966	-0.0027	-1.1074	0.2513
-0.7332	0.0416	0.2743	1.3771	-0.7040	-0.2103
-0.8706	-0.1410	0.5674	-0.4534	-0.9604	-0.2742
1.1843	-0.2484	-0.5758	-0.4003	0.0863	0.3269

Columns 25 through 26

0.5943	0.9335
0.6950	0.1381
-0.8660	-0.0369
0.1463	0.2095
-0.2090	0.0127
-0.6187	0.1623
0.5064	0.5240
-0.8595	0.0631
-0.7149	0.0023
0.9501	-0.6615
-0.2596	-0.0941

GB1 =

0.4633
-0.0418
-0.8066
-0.2664
0.6200
0.8475
0.9360
-0.3454
-0.7872
0.2532
0.2448

GW2 =

-0.5113	0.8279	2.1084	-0.7267	-0.6262	-0.1634
-0.2097	-1.1432	-0.7487	0.3537	-1.3115	

GB2 = 0.5884

Artificial Neural Networks (ANNs)

F.1 Artificial Neural Networks (ANNs)

There are many names associated with ANNs such as connectionist model, neuromorphic system and parallel distribution systems. For the purpose of simplicity, ANN will be used. ANNs gets its name from its use of a system structure analogous to that of biological neurons [122, 13]. The processing units of an ANN, input, weights and output values, provide a mathematical model of numerous neurons connected in a network [17]. There are numerous types of ANNs, such as feed-forward, recurrent neural networks and probabilistic neural network. In a feed-forward neural network, the nodes in one layer are connected only to the nodes of the next layer allowing the forward flow of information. In a recurrent neural network, the nodes are linked with the next layer and to that of its previous layer [109]. This allows data and information to flow in both directions. These two ANN types can be visualized in figure F.1. The ANN structure used is a feed-forward, backpropagation multi-layer perceptron (explained in greater detail in section F.1.2). The complex nature of the ANN allows this pattern classification approach to model the complexities of the input and output relationship [109].

F.1.1 ANN Application

ANNs have been used for the prediction of disease outcome (cancer, epilepsy [123]), micro array for gene expression and medical treatment. Using patient information (age, sex, physiological condition), medical data and data collected from medical instruments (ie: EEG signals [123], electromyography), as input variables, the uses of this pattern recognition method are endless. An ANN's predictive power is evident in medical uses. An example is the for prediction of significant pathology outcome for patients in clinics with

Figure F.1: Two layer neural network types and their information connection pathway (created from [109, 106]).

colorectal symptoms. The ability to predict 90% (95% CI 84-96) in the internal validation set was of significant value as it exceeded that of clinical accuracy (75.0%). Used on an external patient set (n=50) in primary care, the model accuracy was 89.0%. Application of this model with greater prediction accuracy could contribute to a reduction in unnecessary lower gastrointestinal endoscopies [124].

F.1.2 The Multi-Layer Perceptron (MLP)

The complexity of an ANN lies within its hidden layer giving rise to a pattern classification system containing numerous layers. Each layer of a multi-layer perceptron contributes to the final output. The input variables are entered in the input layer, while the output layer contains the value for the output variable. The hidden layers are found between the input and output. The ANN can have any amount of input layers, hidden layers and output layers [120]. The simplest type of ANN is the single layer network (2-layer network) composed of an input layer and an output layer. A double layer network (3-layer network), is composed of an input layer, a hidden layer and output layer (as shown in Figure F.2) which was used for our research purpose. Each layer contains nodes which are units that receive and/or send signals to the next layer. The input layers send signals to the next layer which then receives the signal, processes it and outputs a response to the next layer. The output of the previous layer becomes the input for the next layer.

Figure F.2: A figure depicting the input layer, hidden layer and output layer creating a three-layer network (created from [18])

When the output of the layer is not seen by the user, the layer is considered hidden [120]. The computational power is found within the connections between the nodes as each node is associated with a connection weight. The input variables are then weighted with the connection weights, summed or calculated according to ANN mathematical equation [125]. The output is passed through an activation function which has a non-linear relationship [125]. Figure F.3 is a model of the input, output, weight and activation process. The activation function can be a hard limiter (or step), a threshold logic (ramp), a linear or a sigmoidal logistic function.

Figure F.3: Neural Network structure and its components —nodes containing input variables (I_i), weights (W_i), bias (B), the activation function (F) and output nodes (courtesy of [106]).

F.1.3 Activation Function

In order to activate and pass the signal through to the next layer, the input values must meet the requirements of the activation function. Computations are given (see equations below) for the weighted sum of the input values forms the scalar net activation (net) - where the value $x_0=1$ and $w_0=1$.

$$net_j = \sum_{i=1}^d x_i w_{ji} + w_{j0} = \sum_{i=0}^d x_i w_{ji} \equiv w_j^t x \quad (F.1)$$

The values of i is the units in the input layer, j in the hidden layer, w_{ji} is the input to hidden layer weights at the hidden layer j [18]. The hidden unit outputs a nonlinear function of its activation $f(net)$.

$$y_j = f(net_j) \quad (F.2)$$

The most common activation function (f) are the logistic sigmoid and hyperbolic tangent function [109]. The function used in this research is the hyperbolic tangent function shown in figure F.4.

Figure F.4: A hyperbolic tangent function activation function

F.1.4 Feed-forward and Learning

A feed-forward neural network has two primary modes of operation: feed-forward and learning. The feed-forward portion consists of a pattern that is received by the input units, passes a signal through the network, and the signal received by the output units. The

learning portion is where the output of the network is compared to the desired target. The difference between the two is the error, which is minimized when the output value matches that of the target [120]. This learning portion, called backpropagation, occurs in two steps; the first is the calculation of derivatives of the error function (with respect to the weights) back into the network and the second where the derivative will be used to adjust the values of the weights [120, 18].

F.1.5 Backpropagation

The ANN-RFW uses gradient descent as its backpropagation method. In a gradient search, an error function is defined (sum of squares error) and evaluation of the derivatives of the error function (with respect to weights and bias in the networks) occurs. A weight vector that minimizes an error function $E(w)$ is calculated. There may be many minima that satisfy the condition where gradient of E in a weight space equal 0; $\nabla E = 0$. The desired value is the global minimum where the values of the error function is the smallest, though many other local minima exist. To begin its search, an initial weight vector is chosen at random and then gradient descent algorithm searches through the weight space [120, 18].

$$w^{\tau+1} = w^{\tau} + \Delta w^{\tau} \quad (F.3)$$

where τ is the iteration step and Δw^{τ} is the weight vector increment. The choice of the initial weights will determine which local minimum the algorithm will converge to. A disadvantage of this algorithm is the possibility of remaining in a local minimum. The weights are then updated towards the rate of greatest error decrease which is the negative gradient $w^{(\tau)}$. At each step the weight value is updated until it satisfies the criteria [120, 18, 122]. $\Delta w^{(\tau)} = -\eta \nabla E|_{w^{(\tau)}}$; where η is the learning rate. Re-evaluation of the gradient occurs at each step. For a sequential or pattern based gradient descent, for one pattern at a

time, the error function is evaluated and weights updated using the equation below [120].

$$\Delta w^{(\tau)} = -\eta \nabla E^n|_{w^{(\tau)}} \quad (\text{F.4})$$

where n is the different patterns in the training set. Successive steps will lead to a decrease of the value of E (provided that the learning rate is small). This will eventually lead to a weight vector satisfying the condition $\nabla E = 0$ [120, 18, 13, 122].

Learning Rate

The learning rate of the system plays a role on how fast the system reaches minimal error, if the system will reach minimal error [18]. This is shown in figure F.5. If the learning rate value η is smaller than the optimal learning rate η_{opt} then convergence will occur (a). If $\eta_{opt} < \eta < 2\eta_{opt}$, the system will oscillate but convergence occurs in multiple steps and training will be slow (b). If $\eta > 2\eta_{opt}$ the system will diverge and the minimum error will not be reached (c) [18]. Convergence occurs when the local minimum error is reached. An ideal situation is when local minimum error is reached within one learning step; however this is a rare occurrence [18, 120].

Figure F.5: Learning rate of ANN by gradient descent where w is the error and w^* has the minimum error of 0 [18].

F.1.6 Weight Decay and Weight Elimination

Weight-decay and weight elimination are used for regularization of the network [120]. Weight-decay decays the connection weights towards zero which effectively reinforces larger weight and weakening small weights [126]. By weakening smaller weights (used to code noisy data); the network becomes more robust and reduces the effects of noise. The weight-decay constant (λ) determines how strongly the weights are penalized [43]. In cases of weight-decay, the weights that are not needed for error reduction are decreased to a value that can ultimately be eliminated. Weights that are necessary to solve a problem will not decay indefinitely [18]. A larger λ indicates weights must be closer to zero to be considered part of the noise, thus increasing the pressure on small weights to further be reduced [43]. Weight elimination attempts to reduce weights to zero [14, 42] effectively removing them from the neural network. This is essentially a form of pruning the network [43] which has been used to improve ANN classification results [100]. Weight decay and weight-elimination can be applied by using the ANN starting parameters as outlined in (Section F.2.1).

F.1.7 Steps

The following is a summary of the steps used to train a three-layer MLP [127, 18, 120]: (1) Initialize weights to small random values, (2) Select a training vector (input and corresponding output) from a training set and enter into the inputs of the ANN, (3) Use the feed-forward phase to calculate the actual output, (4) Adjust weights in order to minimize the error (difference between the target and actual outputs), (5) Repeat Step 2 for all training vectors, (6) Proceed through steps until the error is within an acceptable range

F.2 Ann Research Framework (ANN-RFW)

In 1993, MIRG developed its first program create ANNs. Since then, the program has improved by many MIRG researchers. Using Matlab® as a platform, numerous improvements have been made to increase the functionality and performance of this program.⁵⁹ There are many parameters to consider when determining the ANN architecture. The purpose of the ANN-RFW is to improve the efficiency and effectiveness of developing an optimal ANN structure to classify medical outcomes. The current ANN program (ANN Research Framework 1.3) was updated by Dawn Rybchynski (2005) and was used in this thesis. The ANN-RFW program uses an automated, exhaustive method of finding the optimal architectural structure by running the training vectors through a structure with no hidden nodes (0) to the maximum number of nodes ($2*n+1$ hidden nodes, where n is the number of inputs) [14].

F.2.1 Parameters

The following network parameters are used for the ANN classifier design. Starting parameters are set by the user before running training and with ANN training, the final optimal parameters are obtained (from the MIRG ANN GUIDE by Andy Pan) [114].

learn rate (lr): The value for the learning rate was chosen to gain a balance between speed of training and to ensure convergence (as mentioned in section F.1.5) [120, 18]. The learning step increments can be adjusted using the learn rate increment (lr_inc) and the learn rate decrement (lr_dec).

weight-decay constant (λ): This determines how strongly the weights are penalized. The weight-decay constant can be adjusted by modifying the lambda increment (λ_{inc}) and lambda decrement (λ_{dec}) values [120].

⁵⁹**Matlab:** An object-oriented programming environment created by The MathWorks, Inc <http://www.mathworks.com/>

weight scale (wnot): weight-elimination scale factor defined the sizes of ‘large’ and ‘small’ weights. When w_0 is small, this term favours small weights. The small weights will be forced to zero resulting in fewer large weights leading to a form of network pruning (i.e. weight-elimination) [42]. A large w_0 limits the large size of weights and causes many small weights to remain (i.e. weight-decay) [14, 120].

momentum (momentum): A proportion of the previous weight-change is added to the new value (in the gradient descent formula), thereby giving the algorithm some “inertia” through the weight-space and prevents it from getting caught in local minima or plateaus. This can lead to faster convergence without oscillation [14, 120].

error ratio (err_ratio): The error ratio is a control parameter used to make adaptive changes in backpropagation on the learning rate, the weight-decay constant, and the momentum term.

F.2.2 Early Stopping Criterion

An early stopping criterion is a means to control the effective complexity of a network, balancing the error of the training data to that of the validation data [120]. Typically, the error decreases during the training of the classifier. The validation set also has a decrease in error and then a gradual increase in error. This increase of error indicates that though the error of the training set decreases, the model does not validate well. This is indicative of over-fitting, a situation where the model performs well only on training cases. This situation is unacceptable as a model developed to be used for the prediction of new, unknown clinical cases; therefore an over-fitting would be a useless system. In order to avoid over-fitting an early stopping criterion stops model adjustment when the error of the validation set is at a minimum. This point is the model’s optimal performance [18, 120].

Decision Trees

G.1 Decision Trees (DTs)

There are numerous algorithms to create a DT - the most popular being Classification And Regression Trees (CART), ID3, C4.5 and C5. Though all algorithms follow basic steps (i.e. splitting criteria, pruning, noise handling and missing values), the method of differentiating features are unique within each DT method. C4.5 is a successor of ID3 and with a readily available sources code, while C5.0 (See5) is the newest commercially available DT algorithm available from RuleQuest Research and was acquired by MIRG for this work [13, 128].

G.1.1 What is a Decision Tree?

The basic components of a decision tree are the root node, subsequent nodes and terminal leaves. The root node is at the top of the DT where the first splitting criterion is made. It is connected to successive branches of subsequent nodes, leading to further splitting criteria at each node. At each node, a value of the splitting criterion is made which includes values that fall within the threshold. These values are assigned to one node, while values that fall outside the threshold are assigned to another node. The final outcome is assigned at the terminal node (leaf) which has no further splitting criterion. At the terminal node a category label is read. The split is made with a least two values [18]. This can be visualized in Figure G.1.

G.1.2 First Step —Constructing a Decision Tree

The tree construction algorithm splits the data locally at each node - using a greedy top-down method. The greedy algorithm selects a test that maximizes a heuristic splitting

Figure G.1: Diagram of a decision tree and its components (created from [18])

criterion and tests locally [106].⁶⁰ Although, local optimum may be reached, it does not necessarily guarantee global optimum. A decision tree is obtained using the divide-and-conquer algorithm. The whole dataset is contained in the root node, and selected subsets are partitioned from the top down [17]. The feature that is selected as the root node is the feature that results in maximum decrease in entropy or maximum gain in information [17, 106, 13]. Different tests are used to partition the data into subsets. The use of different variables and training vectors results in the creation of a completely different tree

Entropy Impurity (Information Impurity):

The DT is built to allow each leaf node to be as pure as possible. Minimizing the impurity is used to assess the quality of the DT. A popular method is the entropy impurity given in equation below. The data is split at each node by obtaining minimum impurity at each node [18].

$$I(S) = - \sum_{i=1}^l p_i \log_2 p_i \quad (G.1)$$

⁶⁰**Heuristic:** defined as a process that achieves faster solutions or optimal solutions for a problem[129]

where p_i is the *a priori* probability (frequency of class i at node(S)) and l is the number of classes. If all the patterns fall within the same category, the impurity is minimized to 0 and maximized at when both patterns are equally likely. There are several mathematical measures of impurity proposed (Gini and Misclassification Impurity). All of which peak at equal class frequency [18].

Gain Criterion:

Information gained after being partitioned into subsets, S_i , by test B . The test B that maximizes $G(S, B)$ is chosen [13]. The gain criterion is maximized by a test which each subset (S_i) contains one case.

$$G(S, B) = I(S) - \sum_{i=1}^t \frac{|S_i|}{|S|} I(S_i) \quad (G.2)$$

Gain ratio:

The gain criterion is biased towards tests with numerous outcomes, thus the gain ratio is more suitable for selecting the test. This particular ratio ignores information regarding the class and instead incorporates the number and size of the daughter nodes of which the attribute splits the dataset [130]

$$P(S, B) = - \sum_{i=1}^t \frac{|S_i|}{|S|} \log \left(\frac{|S_i|}{|S|} \right) \quad (G.3)$$

G.1.3 Second Step —Pruning Options

In theory, a DT can be grown so that it fits the training data with 100% accuracy and minimized impurity. The grown tree is often complex, which leads to over-fitting of the classifier with poor generalization on the test data. The DT classifier needs to be pruned in order to perform well on new data [18]. See5 offers pruning options to the initial tree.

Smaller values cause more of the initial tree to be pruned while larger values cause less pruning [128])

Pre-pruning:

Also known as “forward pruning”, pre-pruning builds the decision tree up to a certain point. The Chi-squared equation (χ^2) or information gain are the early stopping criteria used to assess the goodness of a split [17]. Sometimes stopped splitting lacks sufficient beneficial splits since the splits may be stopped “too early” to provide overall optimal recognition accuracy [18].

Post-pruning:

Also known as “backward pruning”, post-pruning builds the decision tree and then prunes it. This method is usually the choice for most DT algorithms. The following pruning options are assessed; subtree replacement, subtree raising or leaving the subtree as it is. A subtree replacement assesses whether or not making subtrees into a leaf node will improve the accuracy on test data. The implementation of subtree raising begins at the leaf node and assesses the tree bottom-up [130]. This involves merging or joining subsequent leaf nodes. The procedure that leads to a satisfactory (small) increase in impurity between pairs of nodes declares the combined node as a leaf [18].

G.1.4 Third Step —Rule Induction Based on the Pruned Tree

The final pruned decision tree contains a path from the root node to each leaf. Each successive pathway yields a potential rule [131].

G.1.5 Fuzzy Thresholds

Real life data may not consist of such sharp cut-off values that exist within many decision tree classification algorithms. Sometimes a slight change of a feature value may

result in misclassification. The hard cut-off thresholds of a splitting criterion can be softened using fuzzy thresholds leading to gradual transitions between attribute values [132]. Thresholds are divided into three ranges - a lower bound, upper bound and a central value. A value falling within the lower and upper bound follow classification of $\leq$ or $>$ respectively. Investigation is performed for a value that falls in between bounds (central bound) and results are combined probabilistically [133].

G.1.6 Prior frequency and Cost

It is not safe to assume that a certain category is represented with the same frequency both in the training and test data. When frequencies are different, weighting the samples to correct for the prior frequency is needed in order to control the tree creation. Also, the cost of misclassification may be different for each category and therefore may be modified using a cost matrix [18]. Since this may assist with better classification of skewed outcomes, it was used in the development of a decision tree classifier.

G.1.7 Advantages of a DT

There are many advantages of using decision trees. They have the ability to deal with missing cases and nominal data.⁶¹ In addition, attribute divisions are easy to interpret. DTs can be more comprehensive for physicians as they are easily understood when compared to other data mining methods such as neural network or genetic algorithms which are both “black box methods” [18].⁶² Tree generation using See5 is quick, ranging from 1 second to 10 seconds when classifying over 90,000 cases. A logical approach does not depend on the statistical distribution of the attribute or independence of the attribute [106].

⁶¹**Nominal data:** data that is discrete, with no order or metric system

⁶²**Black Box Methods:** manner of which an output is generated and patterns placed on data is not easily comprehended by a non-specialist[134]

G.1.8 Disadvantages of a DT

DT classifiers are not robust.⁶³ The accuracy is greatly affected by noise, training cases and the size of the training data [17]. In addition, DTs do not have the ability to handle complicated relationships between features as the separation uses simple axis-parallel decision boundaries [17].

⁶³robust methods are insensitive to outliers

Additional Model Estimation Techniques

H.1 Model Estimation

The optimal model is created by using a portion of the data to train (training set) the system and another portion to validate findings (test set). The test result is used to evaluate the model and in order for the results to be reliable, the data sets must be sufficiently large and independent. Division of the dataset assumes that the divided sets represent the same, unknown distribution.⁶⁴ For larger datasets this may be true. In the case of smaller datasets, the choice of division may have an impact on the accuracy of the model [106] as shown in Table H.1.

H.2 Bootstrap Methods

Bootstrapping is a powerful procedure for small-sample datasets as often there is insufficient data to make statistical calculations or error estimates. This method is used to make statistical inferences by drawing a sample size n from the original dataset D and calculating the statistics for each of these samples [137]. This method re-samples with replacement.⁶⁵ This is repeated many times (200-1000) times in order to calculate the bootstrap standard error, confidence intervals and test error estimates of the original dataset [137, 138]. For situations that were applicable, a modified version of the bootstrap method was used to create datasets used to verify the model. These sets are not used in training or testing of the model, but are needed to determine how well the model performs on new cases. For smaller databases with rare outcomes (such as the adult ICU) bootstrapping proves to be a valuable method.

⁶⁴the distribution of the dataset is not a factor of the division process

⁶⁵Each sample size, n is then placed back into the set and a new sample is drawn. Though each sample set is the same size, different data points may be drawn for each set.

Table H.1: Methods used for creating and testing a classifier

Method	Characteristics	Advantages	Disadvantages
Re-substitution	All available data used for training and testing [106]	Simplest method. Good for small dataset[135]	Test error rate estimation is optimistically biased because same data is used to train model [136]
Holdout	Half or two-thirds of data used for training and the remaining is used for testing [106]	Numerous training and testing set will improve pessimistic error [106]	Training and test sets are independent and error estimate pessimistic [106]. Different partitioning gives different error [106]
Leave-one-out (LOO) / Jackknife	Model designed using (n-1) samples for training and tested on the remaining sample [18]	Yield precise estimates even for 'moderate sample sizes' [13]	Large computational time. n models are designed and compared [137]
n-fold Cross Validation	A mixture between holdout and LOO. Dataset is divided into P disjoint datasets [106]	Each subset is used for training and remaining for testing [18]. Popular good for small number of samples [106]	Needs a lot of data [137]
Bootstrap	Re-samples available data with replacement, to generate several new artificial data sets (200-1000) [137]	Provides bootstrap error estimates. Accurate estimates for small class sizes Outperforms cross-validation estimates [135]. Good for small data [106]	Large computational time [137]

H.3 Ensembles for Feature Selection

Selecting the important features for classification can improve generalization of the model and provide faster, more cost-effective models [139]. DTs are an embedded method of feature selection.⁶⁶ The information is embedded in its node structures and DTs are fast, flexible and are able to handle multiple types of data which makes it a popular tool for feature selection. However, DTs are unstable classifiers and benefits from the use of ensemble methods. Ensemble methods explore the outcomes of numerous classifiers, rather than just one. Creation of ensemble classifiers can occur by varying the training data either through the use of re-sampling methods (cross validation, bootstrapping, bagging) and/or using a subset of the total variables [141]. Though ensemble methods are computationally time-consuming, the best performing classifier (or combination of classifiers) can be discovered using this method.

H.4 Average Squared Error

The average squared error (ASE) (also referred to as the mean squared error (MSE)) is a measure of error of the predicted value compared to that of the true value. The equation is the expected value (mean) of the square of the difference of the predicted ($\hat{\theta}$) value from that of the true value (θ) [106].

$$\text{ASE} = \mathbf{E}[(\hat{\theta} - \theta)^2] \quad (\text{H.1})$$

$$\text{ASE} = \frac{(\text{predicted} - \text{actual})^2}{\sum_{i=\text{allcases}} (\text{predicted} - \text{actual})^2} \quad (\text{H.2})$$

Ideally the value of the ASE should be minimized; however, in classification of cate-

⁶⁶**Embedded Method:** models feature dependencies [139] and infers the importance of the interaction between features [140]

gorical results, it is difficult to report a numeric ASE value. ASE is a calculation of the distance error but in a categorical classification of values $[-1,1]$ with a midpoint of 0, any value that deviates from the actual value should have the same error. For example, if a result of 1 is desired, a predicted value of 0.9 and 0.1 would both have the same misclassification error. However, with the ASE equation, 0.1 would have a greater ASE value and therefore be indicated as more incorrect than the value of 0.9 [14].

H.5 Hosmer-Lemeshow

The Hosmer-Lemeshow test (H-L test) is a statistical test designed to see how well a logistic regression model fits the data but is also used as a method of external validation of the data [49]. Since current mortality prediction systems rely on a form of logistic regression (multivariate), the H-L test is used to report the validity of the mortality prediction tools. These models generate a probability of mortality which, when summed for a population, the mean predicted mortality can be compared to the observed mortality [142]. This test is used to assess the calibration of the model, an agreement between the observed and predicted risk. This is used as an indication of the “goodness of fit” of the model or how well the model fits the data. The Null Hypothesis of the $\hat{C}$ statistics states that the model does not fit the data [28]. To compute the $\hat{C}$ -statistic the following formula is used [49]:

$$\hat{C} = \sum_{g=1}^n \frac{(O_g - n_g \bar{\pi}_g)^2}{n_g \bar{\pi}_g (1 - \bar{\pi}_g)} \quad (\text{H.3})$$

The n refers to the number of groups. The data is usually partitioned into 10 equally sized subsets based on deciles of predicted risk [49]. O_g is the number of events observed in group g , e_g is the number of events expected in group g and $\bar{\pi}_l$ is the mean estimated probability, always in the group l [4]. Since the distribution of the H-L test is approximately chi-squared, a χ^2 test is used to actual outcome versus predicted outcome for each subgroup [143]. The degrees of freedom (df) are as follows: 8 df for the training sample, 9 df on the

internal validation and 10 df on external data [49]. A P-value >0.05 indicates rejection of the Null Hypothesis, indicating a good fit [28]. A high $\hat{C}$ value relates to a small P-value which implies a rejection of the Null hypothesis; however, the $\hat{C}$ value is subjective to the division of the group (deciles, quartiles) and the sample size [28, 143]. Therefore results of this test is always accompanied by another form of statistical testing, usually the AUC of the ROC.

Appendix J. Atlee Database Application
JOINT DATA ACCESS COMMITTEE APPLICATION - FULL PROPOSAL

Purpose and objectives of the project - Research and Education

Our research group (MIRG) has identified artificial neural networks (ANNs) as an ideal data mining tool to predicted clinical outcomes and risk factors using large medical databases. At the present time, MIRG has used ANN to produce models that predict high risk preterm birth, mortality and length of stay and artificial ventilation duration in the NICU. In addition to this, ANNs have been used to identify risk factors for complications types for NICU patients.

ANN has been used by MIRG with two different perinatal databases (PPESO and PRAMS) to predict pre-term births. The results to date match the best existing performance system (the invasive fibronectin test) with 66 % sensitivity and 85 % specificity. PPESO was used to predict delivery type (caesarean or vaginal) and APGAR5. The ATLEE database may provide input variables that lead to additional improvement of the predictability of such a system. The improved system could become an effective Computer Decision Support System (CDSS) to help the management of future births in Nova Scotia and elsewhere. CDSSs are designed to provide caregivers with clinical outcomes which complement their own experience and knowledge. The CDSS contains medical knowledge from various cases, more than one doctor will experience in a lifetime. For example, in a previous MIRG research project using the PPESO database, there were over 17,688 cases. Implementation of CDSS has the potential to support the decisions made by the medical staff and also provides a complementary technique to costly and invasive methods.

Significance/ benefits to be derived

This research project will further expand and contribute to the existing CDSS research by enhancing the CDSS and using it to predict the following five outcomes - preterm birth, delivery mode (caesarean section or vaginal), Apgar score at 5 minutes after birth, and likelihood of an infant being admitted to NICU. Accurate predictions of such outcomes will allow the medical staff (doctors and nurses), parents and family to prepare for the predicted outcome. Such preparations may include access to the proper medical equipment, having the required medical team along with preparing for the physical and emotional repercussions of these outcomes. In addition, by identifying such risk factors, attempts can be made to eliminate them whenever possible.

The results to date of our previous study of high-risk preterm birth match the best existing performance system (the invasive fibronectin test) with 66% sensitivity and 85 % specificity. Anything better than is achieved now by using current risk assessment forms (such as in prenatal records) would be useful from a clinical perspective – the current approaches have very poor sensitivity and specificity. From the clinical standpoint, even the results we have already achieved would greatly impact practice by reducing the number of women subjected to unnecessary high risk care and increasing the proportion of those who will have preterm birth identified for appropriate high risk care. With a richer database such as the Atlee, it is possible we will see further significant improvements in sensitivity and specificity.

Detailed methodology for the use of RCp/Rh Program data

Objectives: The data will be used to predict the five listed outcomes of interest. Risk factors will be identified using data mining methods and variables that contribute to the following outcomes will be used in the minimal dataset.

- a) **Preterm birth** – preterm birth is described as birth before 37 weeks of gestation which has increased more than 30% in the States. Factors influencing this increase may be due to the changes in obstetric practice (labor induction as well as Caesarean delivery before the 37 gestation period). A US and a Canadian study demonstrated that late-preterm birth have a high risk of death within their first year compared to term infants. They also experience higher incidence of health problems such as respiratory distress, apnea, hypoglycemia and hypothermia (Tomashek *et al.*, 2007)
- b) **Delivery method (Caesarian or vaginal)** – Used to prevent maternal and/or fetal morbidity, caesarean method requires delivery of a fetus through surgical incision through the abdominal wall (Caesarean Delivery, 2008). In 2006, over 31.1% births in the US were delivered by cesarean delivery which is an increase from the 20.7% from 1996 (MacDorman, 2008). This method may reduce certain risks of breech position, it may also increase other risk factors associated with preterm birth such as respiratory morbidity (Hansen, *et al.*, 2008).
- c) **The probability of the infant being delivered to the Neonatal Intensive Care Unit (NICU)** – Infants placed in the NICU have a comprised immune system and are more vulnerable to contraction of an infection. The intrinsic risk factors for infection are due to the underdeveloped immune system, skin barrier and gastrointestinal tract. The low birth weight and the severity of the illness also contribute to the risk. Extrinsic risk factor related to rigorous treatment needed to sustain the infant are invasive and are of high contribution to infection (Saiman, 2002).
- d) **Mortality** – Death of an infant within 28 days after birth is defined as neonatal mortality. Early neonatal death is described as death before a 7 day period while late neonatal death occurs between 7 to 28 days. There are certain factors that contribute to neonatal death and these variables have been identified and used to predict mortality. According to the World Health Organization, mortality is also an indication of maternal and infant health status (WHO, 2008). Previous MIRC work used 13 variables as the set of risk indicators used to predict mortality (Zhou, 2006).
- e) **APGAR5** – The APGAR score represents the status of the newborn and the response to resuscitation. A score of 0, 1 and 2 is assessed with for each of the five components – Appearance (color); Pulse (heartbeat); Grimace (reflex irritability); Activity (muscle tone) and Respiration (breathing). The score reported at 5 minutes have indicates certain health status of the newborn where a score of 7 or over indicates a good health status (The APGAR Score, 2006).

1. Data Processing and Procedure:

The data will first be formatted so that compatibility with the software is possible. Cases with missing values will either be deleted or if they are few, the missing values will be imputed using the method successfully developed by MIRC (Ennett, 2003 and 2008). Then the data is passed through an ANN which maximises sensitivity and then again in an ANN which maximizes specificity. (Catley, 2005 and

2006).

2. Details of the Analysis Approach

Figure 1 Artificial Neural Network Structure (Duda, Hart and Stork 2001)

This ANN consists of an input layer, a hidden layer and an output layer (Figure 1). Each layer has nodes (Figure 1 – circles indicate nodes). The input nodes are the variables that contribute to the outcome (target nodes). The links from each layer (input and hidden) are connected and accompanied by weights associated at each layer (Duda, Hart and Stork 2001). The input value is multiplied by the weights and is passed through a mathematical function (hyperbolic tangent sigmoid transfer function- Figure 2). The resultant value carries a pattern signal into the hidden nodes. The hidden nodes become the input nodes for the output layer and follow similar mathematical functions towards the output layer.

Figure 2 Hyperbolic tangent sigmoid transfer function

The two primary modes of operation of this ANN is FEED-FORWARD and LEARNING. FEED-FORWARD indicates that the pattern moves forward from input to output. LEARNING indicates that the output of the network is compared to the desired target (the error) and the ANN is

adjusted so that error is minimized (Bishop 1995). This type of ANN is a *multilayer perceptron feedforward backpropagation* with the hyperbolic tangent transfer function.

ANN Analysis

Based on the information acquired from a large medical database, the ANN is trained on a portion of the data (usually one-third), then tested on another third and finally verified with the last third.

TRAINING: Relevant data and patterns are recognized by the ANN model. An ANN model is developed that simulates complex interactions between the variables. ANNs have been shown to be a valuable tool in classifying data with complex, non linear association between variables and have been useful in predicting phenomena's related to many factors (Catley *et al.*, 2006).

TESTING: The ANN model developed is tested on another portion of data to determine the real performance result.

At maximum performance, weights are extracted to determine which variables are most predictive of the outcome of interest. This will result in an optimal minimal dataset used to input into the prototype CDSS.

Performance Measures

The success of the ANN test set performance is based on certain criteria: sensitivity, specificity, accuracy and area under the Receiver Operating Characteristic (ROC) curves. When estimating medical outcomes, the key measures of performance - the sensitivity and specificity - must adhere to clinical criteria standards and are needed for clinical application (Catley *et al.*, 2006).

SENSITIVITY: The true positive rate (number of positive cases that are predicted correctly) measures how well the model classifies the positive outcome

SPECIFICITY: The true negative rate (number of negative cases that are predicted correctly) measures how well the model classifies the negative outcome.

Diagnostic accuracy is maximized when both the sensitivity and specificity is maximized. ROC curves provide a visual model that shows the tradeoff between sensitivity and specificity (Figure 3).

ROC Curve for PRAMS PREV_{PREA}(May 20) Classification Using DLBP with 3 hidden nodes.

Figure 3. An example of an ROC curve. Train: 0.7918. Test: 0.7761

The values of the TPR (Sensitivity) are graphed along the Y-axis and the X axis has the values of False Positive Rate (1- Specificity). The area under the curve (AUC) is calculated and a value of greater than 0.75 is highly desired. The Positive Predictive Rate and Negative Predictive Rate will also be calculated and reported.

The results of previous work, predicting high risk preterm birth, had a sensitivity of 66% and a specificity of 85%. This outcome has met the clinical criteria set out by the physician involved (Dr. Robin Walker) in the PRAMS data project. Results are also based on average (mean) performance and standard deviation of multiples different verification sets for each outcome. The results of the CDSS must follow or exceed clinical medical standards in order to become of interest for clinical use. Our current approach may be improved through the additional experience gained by using the ATLEE database and its set of variables. It is also possible that a combination of variables from the various data sources may provide the best predictions.

Austin, C., Boxerman, S. *Information Systems for Healthcare Management*, 6th ed. Chapter 7. Chicago. Health Administration Press. 2003.

Bishop, C. *Neural Networks for Pattern Recognition*. New York: Oxford University Press, 1995.

Cesarean Delivery. *Emedicine for WebMD*. <<http://www.emedicine.com/med/topic3283.htm>> (accessed July 20, 2008)

Cardoso, J. *Semantic Web Services: Theory, Tools and Applications*. Hershey PA. IGI Publishing. 2007. <<http://library.books24x7.com.proxy.library.carleton.ca/toc.asp?bookid=20775>>

Cardoso, J; Sheth, A. *Semantic Web Services, Processes, and Applications*. New York. Springer Science+Business Media, LLC. 2006. <<http://www.mylibrary.com.proxy.library.carleton.ca/Browse/open.asp?ID=72420>>

Catley, C., M. Frize, R.C. Walker, and D.C. Petriu. "Predicting preterm birth using artificial neural networks." *IEEE*, 2005 : 103-108.

Catley, C., Frize, M., Walker, R., Petriu, D. "Predicting High-Risk Preterm Birth Using Artificial Neural Networks". *IEEE TRANSACTIONS on Information Technology in Biomedicine*. vol. 10, no. 3, pp. 540-549, July 2006.

Communications Security Establishment Canada. *CSEC Approved Cryptographic Algorithms for the Protection of Protected Information and for Electronic Authentication and Authorization Applications within the Government of Canada*. July 2008. <http://www.cse-cst.gc.ca/its-sti/publications/itsa-asti/itsa11d-eng.html> (accessed July 8, 2009).

CSEC. *Government of Canada (GC) Policy for the Protection of Classified Information Using Suite B Algorithms*. November 2006. <http://www.cse-cst.gc.ca/its-sti/publications/itsb-bsti/itsb40-eng.html> (accessed July 9, 2009).

Duda, R., P. Hart, and D. Stork. *Pattern Classification*. New York, NY: John Wiley & Sons, Inc., 2001.

Ennett, C., Frize M., Charette, E. "Improvement and automation of artificial neural networks to estimate medical outcomes". *Medical Engineering & Physics*. vol. 26, issue 4. pp 321-328. May 2004.

Ennett, C., M. Frize, and R. Walker. "Imputation of missing values by integrating neural networks and case-based reasoning." *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE, 2008*: 4337-4341.

Hansen, AK., Wisborg, K., Uldbjerg, N. and Henriksen, TB. "Risk of respiratory morbidity in term infants delivered by elective caesarean section: cohort study". *British Medical Journal*. pp. 85-87. January 2008.

MacDorman, M., Menacker, F, Declercq, E. "Cesarean Birth in the United States: Epidemiology, Trends, and Outcomes". *Clinics in Perinatology*. Volume 35, Issue 2. pp. 293-307. June 2008.

NIST. *Approved Algorithms*. July 28, 2008.
http://csrc.nist.gov/groups/ST/toolkit/block_ciphers.html (accessed July 5, 2009).

Puentes, J., Coatrieux, G., Lecornu, L. Secured Electronic Patient Records Content Exploitation (Chapter 12), in *Healthcare Knowledge Management: Issues, Advances, and Successes*, Bali, R., Dwivedi, A. (Editors), Springer Verlag, pp. 162-178. 2006.

Puentes, J., Batrancourt, B., Atif, J., Angelini, E., Lecornu, L., Zemirline, A., Bloch, I., Coatrieux, G., Roux, C. "Integrated Multimedia Electronic Patient Record and Graph-Based Image Information for Cerebral Tumors". *Computers in Biology and Medicine*, Volume 38, N° 4, pp. 425-437. 2008.

Saiman, L. "Risk factors for hospital-acquired infections in the neonatal intensive care unit". *Seminars in Perinatology*. vol 26, no. 5, pp. 315-321, October 2002.

Simon, S. (2008). *Explanation of ROC curves*. Last accessed: May 13, 2009 from <http://www.cmh.edu/stats/ask/roc.asp>.

"The APGAR score". *Advances in Neonatal Care*. vol 6, no. 4, pp. 220-223. August 2006.

Tomashek, K.M., C. Shapiro-Mendoza, M. Davidoff, and J. Petrini. "Differences in Mortality between Late-Preterm and Term Singleton Infants in the United States, 1995–2002." *Mosby, Inc*, 2007: 450-456.

WHO Neonatal Mortality Rate. *World Health Organization*.

<http://www.who.int/whosis/indicators/2007MortNeoBoth/en/> (accessed July 15, 2008)

Zhou, D., Frize, M. "Predicting Probability of Mortality in the Neonatal Intensive Care Unit". *IEEE*. pp 2308-2311. 2006.