MODELING PATIENT FLOW IN A NETWORK OF INTENSIVE CARE UNITS (ICUs)

by

Yazan Roumani

BS, Indiana University of Pennsylvania, 2003

MBA, Indiana University of Pennsylvania, 2004

MS, University of Pittsburgh, 2006

Submitted to the Graduate Faculty of

Joseph M. Katz Graduate School of Business in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2013

UNIVERSITY OF PITTSBURGH

JOSEPH M. KATZ GRADUATE SCHOOL OF BUSINESS

This dissertation was presented

by

Yazan Roumani

It was defended on

April 11, 2013

and approved by

Jerrold H. May, Professor, Business Administration

Jennifer Shang, Professor, Business Administration

David P. Strum, Associate Professor, Anesthesiology

Youxu Cai Tjader, Part-time Lecturer, Business Administration

Dissertation Advisor: Luis G. Vargas, Professor, Business Administration

MODELING PATIENT FLOW IN A

NETWORK OF INTENSIVE CARE UNITS (ICUs)

Yazan Roumani, PhD

University of Pittsburgh, 2013

Beginning in 2012, the Department of Health and Human Services (HHS) started adjusting payment for specific conditions by 30% for hospitals with 30-day patient readmission rates higher than the 75th percentile (HHS.gov, 2011). Furthermore, starting in 2013, HHS requires hospitals to publish their readmission rates (HHS.gov, 2011). It is also estimated that by 2013, healthcare expenditures in the United States will account for 18.7% of the Gross Domestic Product (GDP) (Centers of Medicare and Medicaid Services and US Bureau of Census, 2004). Yet the US healthcare system still suffers from congestion and rising costs as illustrated by hospital congestion.

One way to reduce congestion and improve patient flow in the hospital is by modeling patient flow. Using queueing theory, we determined the steady state solution of an open queueing network, while accounting for instantaneous and delayed feedback. We also built a discrete event simulation model of patient flow in a network of Intensive Care Units (ICUs), while considering instantaneous and delayed readmissions, and validated the model using real patient flow data that was collected over four years. In addition, we compared several statistical and data mining techniques in terms of classifying patient status at discharge from the ICU (highly imbalanced data) and identify methods that perform the best.

Our work has several contributions. Modeling patient flow while accounting for instantaneous and delayed feedback is considered a major contribution, as we are unaware of any

patient flow study that has done so. Validating the discrete event simulation model allows for the implementation and application of the model in the real world by unit managers and administrators. The simulation model could be used to test different scenarios of patient flow, and to identify optimal resource allocation strategies in terms of number of beds and/or staff schedules in order to maximize patient throughput, reduce patient wait time and improve patients' outcome. Moreover, identifying high risk patients who are more likely to die in the ICU ensures that those patients are receiving appropriate and timely care, so their risk of death is reduced.

TABLE OF CONTENTS

ACCKNOWLEDGMENTS XVI

1.0 INTRODUCTION 1

1.1 BACKGROUND 1

1.2 PATIENT FLOW 3

1.3 THE INTENSIVE CARE UNIT (ICU) 3

1.3.1 Readmission to ICU 5

1.4 IMBALANCED DATA 7

1.5 OBJECTIVES 7

1.6 RESEARCH CONTRIBUTIONS 10

1.7 CHAPTER SUMMARY 11

2.0 LITERATURE REVIEW 13

2.1 QUEUEING NETWORKS 13

2.2 ANALYSIS OF QUEUEING NETWORKS 14

2.3 QUEUEING NETWORK MODELS 15

2.4 QUEUEING NETWORKS WITH BLOCKING 16

2.4.1 Decomposition method 17

2.5 QUEUEING NETWORKS WITH FEEDBACK FLOWS 20

2.6 QUEUEING NETWORKS WITH BLOCKING AND FEEDBACK FLOWS 22

2.7 QUEUEING NETWORK MODELS IN HEALTHCARE 23

2.8 CLASSIFYING IMBALANCED DATA SETS 26

2.9 CHAPTER SUMMARY 29

3.0 DATA AND ANALYSES 31

3.1 DATA 31

3.2 QUEUEING NETWORK WITH FEEDBACK... 39

3.2.1 Open queueing network (one M/M/1 node) with instantaneous feedback 39

3.2.1.1 Solving the open queueing network (one M/M/1 node) with instantaneous feedback 40

3.2.2 Open queueing network (two M/M/1 nodes) with instantaneous and delayed feedback 41

3.2.2.1 Solving the open queueing network (two M/M/1 nodes) with instantaneous and delayed feedback.. 42

3.3 SIMULATION MODEL 46

3.3.1 M/M/s model 46

3.3.1.1 Coronary Care Unit (CCU)... 46

3.3.1.2 Cardiac Intermediate Care Unit (CICU).. 47

3.3.1.3 Cardiothoracic Surgical Intensive Care Unit (CTSICU) 47

3.3.1.4 Neurotrauma Surgical Intensive Care Unit (NTSICU).............................. 47

3.3.2 G/M/s model 48

3.3.2.1 Coronary Care Unit (CCU)... 48

3.3.2.2 Cardiac Intermediate Care Unit (CICU) 48

3.3.2.3 Cardiothoracic Surgical Intensive Care Unit (CTSICU) 49

3.3.2.4 Neurotrauma Surgical Intensive Care Unit (NTSICU) 49

3.4 CLASSIFYING PATIENTS' STATUS AT DISCHARGE 49

3.4.1 Classification methods 50

3.4.2 Classification measures 51

3.4.2.1 Specificity, recall, precision and F-measure 51

3.4.2.2 Confusion entropy (CEN) 51

3.4.2.3 Hand's measure 53

3.5 CHAPTER SUMMARY 55

4.0 RESULTS 56

4.1 QUEUEING NETWORK WITH FEEDBACK 56

4.1.1 Local balance approach 58

4.2 STEADY STATE SOLUTION FOR A QUEUEING NETWORK WITH INSTANTANEOUS FEEDBACK 58

4.3 STEADY STATE SOLUTION FOR A QUEUEING NETWORK WITH INSTANTANEOUS AND DELAYED FEEDBACK 59

4.3.1 Numerical example of the steady state solution 60

4.4 SIMULATION MODEL RESULTS 61

4.4.1 Comparing mean service time 61

4.4.2 Comparing mean time between external arrivals 62

4.5 CLASSIFICATION RESULTS 63

4.5.1 Logistic regression 64

4.5.2 Discriminant analysis ... 65

4.5.3 Classification and Regression Tree (CART) models ... 66

4.5.4 C5 ... 67

4.5.5 Support Vector Machine (SVM) ... 68

4.5.6 Specificity ... 69

4.5.7 Recall ... 70

4.5.8 Precision ... 71

4.5.9 F-measure ... 72

4.5.10 Confusion entropy (CEN) ... 73

4.5.11 Hand's measure ... 74

4.6 SUMMARY OF CLASSIFICATION RESULTS ... 75

4.7 CHAPTER SUMMARY ... 75

5.0 CONCLUSIONS AND FUTURE RESEARCH ... 77

5.1 CONCLUSIONS ... 77

5.2 CONTRIBUTIONS AND IMPLICATIONS ... 78

5.3 FUTURE WORK ... 79

5.3.1 Non-Markovian queueing network ... 80

5.3.2 Patient classes ... 80

5.3.3 Classifying readmitted patients ... 82

5.3.4 Admission/discharge policies ... 83

5.4.4.1 Analytic Hierarchy Process (AHP) ... 86

5.4 CHAPTER SUMMARY ... 86

APPENDIX A ... 88

APPENDIX B **89**

APPENDIX C **91**

BIBLIOGRAPHY **93**

LIST OF TABLES

Table 1. Data descriptive statistics 32

Table 2. Transition probability matrix of cardiac patients 34

Table 3. Mean, standard deviation, squared coefficient of variation (CV^2) of service time in Coronary Care Unit (CCU), Cardiac Intermediate Care Unit (CICU), Cardiothoracic Surgical Intensive Care Unit (CTSICU) and Neurotrauma Surgical Intensive Care Unit (NTSICU) 34

Table 4. Mean, standard deviation, squared coefficient of variation (CV^2) of time between external arrivals to Coronary Care Unit (CCU), Cardiac Intermediate Care Unit (CICU), Cardiothoracic Surgical Intensive Care Unit (CTSICU) and Neurotrauma Surgical Intensive Care Unit (NTSICU) 37

Table 5. Summary table for calculation of specificity, recall and precision 51

Table 6. Expected performance of specificity, recall, precision, the F-measure and confusion entropy (CEN) as the misclassification cost ratio (MCR) is increased 53

Table 7. Expected performance of Hand's measure for different choices of Beta distribution parameters 54

Table 8. Assigned numerical values for parameters in equations 4.8 and 4.9 60

Table 9. Comparison of mean service time in Coronary Care Unit (CCU), Cardiac Intermediate Care Unit (CICU), Cardiothoracic Surgical Intensive Care Unit (CTSICU) and Neurotrauma Surgical Intensive Care Unit (NTSICU) between actual data and simulation runs...................... 62

Table 10. Comparison of mean time between external arrivals to Coronary Care Unit (CCU), Cardiac Intermediate Care Unit (CICU), Cardiothoracic Surgical Intensive Care Unit (CTSICU) and Neurotrauma Surgical Intensive Care Unit (NTSICU) between actual data and simulation runs.. 62

Table 11. Hand measure values for logistic regression, discriminant analysis, Classification and Regression Tree (CART) models, C5 and Support Vector Machine (SVM) using different Beta distribution parameters.. 74

Table 12. Descriptive statistics of patient classes... 81

Table 13. Transition probability matrix of class 1 patients 81

Table 14. Transition probability matrix of class 2 patients 82

LIST OF FIGURES

Figure 1. General diagram of patient flow in a hospital (ER='Emergency Department', OR='Operating Room', ICU='Intensive Care Unit', PACU='Post Anesthesia Care Unit') 33

Figure 2. Flow diagram of cardiac patients in the queueing network (CCU='Coronary Care Unit'; CICU='Cardiac Intermediate Care Unit'; CTSICU='Cardiothoracic Surgical Intensive Care Unit'; NTSICU='Neurotrauma Surgical Intensive Care Unit'; PACU='Post Anesthesia Care Unit'; Wards; CATH='Cardiac Catheterization Unit') 33

Figure 3. Measured distribution of Coronary Care Unit (CCU) service time 35

Figure 4. Measured distribution of Cardiac Intermediate Care Unit (CICU) service time 35

Figure 5. Measured distribution of Cardiothoracic Surgical Intensive Care Unit (CTSICU) service time 36

Figure 6. Measured distribution of Neurotrauma Surgical Intensive Care Unit (NTSICU) service time 36

Figure 7. Measured distribution of time between external arrivals to Coronary Care Unit (CCU) 37

Figure 8. Measured distribution of time between external arrivals to Cardiac Intermediate Care Unit (CICU) 38

Figure 9. Measured distribution of time between external arrivals to Cardiothoracic Surgical Intensive Care Unit (CTSICU) 38

Figure 10. Measured distribution of time between external arrivals to Neurotrauma Surgical Intensive Care Unit (NTSICU) 39

Figure 11. Open queueing network with instantaneous feedback 40

Figure 12. Open queueing network with instantaneous and delayed feedback 42

Figure 13. Logistic regression's specificity, recall, precision, the F-measure and confusion entropy (CEN) as the misclassification cost ratio (MCR) is increased 64

Figure 14. Discriminant analysis' specificity, recall, precision, the F-measure and confusion entropy (CEN) as the misclassification cost ratio (MCR) is increased 65

Figure 15. Classification and Regression Tree (CART) models' specificity, recall, precision, the F-measure and confusion entropy (CEN) as the misclassification cost ratio (MCR) is increased 66

Figure 16. C5's specificity, recall, precision, the F-measure and confusion entropy (CEN) as the misclassification cost ratio (MCR) is increased........ 67

Figure 17. Support Vector Machine's (SVM) specificity, recall, precision, the F-measure and confusion entropy (CEN) as the misclassification cost ratio (MCR) is increased........ 68

Figure 18. Specificity of logistic regression, discriminant analysis, Classification and Regression Tree (CART) models, C5 and Support Vector Machine (SVM) as the misclassification cost ratio (MCR) is increased 69

Figure 19. Recall of logistic regression, discriminant analysis, Classification and Regression Tree (CART) models, C5 and Support Vector Machine (SVM) as the misclassification cost ratio (MCR) is increased 70

Figure 20. Precision of logistic regression, discriminant analysis, Classification and Regression Tree (CART) models, C5 and Support Vector Machine (SVM) as the misclassification cost ratio (MCR) is increased 71

Figure 21. F-measure of logistic regression, discriminant analysis, Classification and Regression Tree (CART) models, C5 and Support Vector Machine (SVM) as the misclassification cost ratio (MCR) is increased 72

Figure 22. Confusion Entropy (CEN) of logistic regression, discriminant analysis, Classification and Regression Tree (CART) models, C5 and Support Vector Machine (SVM) as the misclassification cost ratio (MCR) is increased... 73

Figure 23. Hand's measure values of logistic regression, discriminant analysis, Classification and Regression Tree (CART) models, C5 and Support Vector Machine (SVM) as Beta distribution parameters are changed.. 74

ACCKNOWLEDGMENTS

I am grateful to Dr. Luis Vargas, my dissertation advisor for all his support and mentorship. I also would like to thank my dissertation committee members, Dr. Jerrold May, Dr. Shang, Dr. Strum and Dr. Tjader for all their help and feedback. I am also grateful to my wife and my family for all their support and encouragement throughout my doctoral career.

1.0 INTRODUCTION

This chapter introduces the problem of congestion and rising costs in the US healthcare system as exemplified by hospital congestion. One of the main reasons for hospital congestion is the lack of coordination among the various units within the hospital. Modeling and studying patient flow within the hospital presents a promising solution to this problem. One particular hospital unit that might benefit from studying patient flow is the Intensive Care Unit (ICU). The chapter identifies some of the gaps and limitations in the existing literature on ICUs, and points out the importance of studying patient flow in a network of ICUs and accounting for readmissions. The chapter also discusses the problem of classifying imbalanced data sets. The objectives and contributions of our work are specified as well.

1.1 BACKGROUND

Healthcare is one of the most important domestic industries in the United States (Cochran et al., 2006). In 2002, US healthcare expenditures reached $1.5 trillion (about 15% of Gross Domestic Product (GDP)) (Cochran et al., 2006). By 2013, the percentage is projected to increase to 18.7% (Centers of Medicare and Medicaid Services and US Bureau of Census, 2004). Furthermore, in 2003, total national health expenditures increased at a rate four times the rate of inflation (Smith et al., 2005).

Despite the above, the US healthcare system still suffers from congestion and rising costs. According to Carter (2002) "in my experience, one of the major causes of inefficiency in the healthcare system is what I call 'localized people.' People working in the healthcare system are very knowledgeable about their own area but have relatively little understanding of what goes on in the next department. Doctors and nurses in the Emergency Department or in operating rooms do not really understand or sympathize with the problems faced by ward staff. People in hospitals have little appreciation of issues in long-term and home care. Occasionally, there are issues about 'my work is more important than yours' or 'my problems are bigger than yours.' This is where Operational Research professionals can play an important role."

One ramification of healthcare system congestion and rising costs is hospital congestion. Although about one-third of healthcare expenditures go toward hospitals (Cochran et al., 2006), hospital congestion is a common problem in the United States. More than 60% of hospitals are operating "at capacity" or "over capacity" (Managing Patient flow, 2004). Congestion can be attributed to several factors including: efficiency demands at the hospital level; healthcare system level and physician practices level; higher expectations from customers; rising costs due to medical premiums and lower reimbursement from insurance companies and federal and state agencies (Managing Patient Flow, 2004).

Initially, congestion was considered an emergency department phenomenon, but it is now recognized as a system problem (Managing Patient Flow, 2004). Consequently, hospital management has recognized the importance of coordination among the various units within a hospital, rather than looking at units independently from one another (Managing Patient Flow, 2004). One way to coordinate hospital units is by studying patient flow within the hospital.

1.2 PATIENT FLOW

According to Koo et al. (2010), patient flow is “the movement of patients through a set of activities, services, or locations in a healthcare facility.” During the flow, patients require various healthcare resources, such as beds, physicians, and equipment (Koo et al., 2010). Clinically, patient flow represents the progress of patient’s health status (Cote, 2000). Moreover, patient flow represents the underlying environment in healthcare, and understanding it is critical for the successful implementation of operations management tools and methodologies (Cote, 2000).

From a managerial perspective, studying patient flow in the hospital should help administrators in better predicting expected demand, and therefore better manage patient admissions and scheduling, bed capacity, and staff scheduling (Cote, 2000). From a clinical perspective, studying patient flow should help administrators, healthcare providers and patients to better understand the disease progress and recovery process (Cote, 2000). Consequently, hospital throughput should improve, physicians and staff’s idle time and reduce patients’ waiting time should. Hence, hospital costs and negative patient outcomes such as morbidity and mortality should be reduced.

1.3 THE INTENSIVE CARE UNIT (ICU)

One of the most important units in the hospital is the Intensive Care Unit (ICU). The ICU is a specialized department in the hospital that is exclusively dedicated to critically ill patients. Patients are usually monitored continuously and closely by highly trained nurses as well as physicians. ICUs are usually equipped with very expensive and highly specialized machines. In fact, ICUs account for 15-20% of hospital costs (Gruenberg et al., 2006; Pronovost et al., 2004). In terms of size (i.e., number of beds), ICUs vary widely depending on the size of the hospital. In some cases, a hospital may have several ICUs, each dedicated to special kinds of patients such as a neonatal ICU (NICU) and a surgical ICU (SICU).

The cost of caring for ICU patients in the United States is estimated to account for 1-2% of the gross national product (Gruenberg et al., 2006). Given the much higher costs for patient care in the ICU, ICUs usually operate at a high level of occupancy, which, in turn, leads to longer waiting times in the upstream units, and lower patient throughput (KC et al., 2012). For instance, the SICU has been identified as a bottleneck in the cardiac care process (KC et al., 2012). High occupancy in the ICU is also associated with increased likelihood of mortality and readmission (Chrusch et al., 2009).

Patients arrive to the ICU from several places. Some patients are admitted to the ICU right after surgery while others arrive from the emergency department (ED) or are directly admitted to the ICU (Dobson et al., 2010). Furthermore, some patients are transferred from another hospital that does not have the capability to handle the case (Dobson et al., 2010). Therefore, managing patient flow in the ICU is critical, as inadequate capacity in the ICU can lead to negative outcomes, such as cancellation of scheduled surgeries, transfer of patients and early discharge from the ICU (Green, 2002; Costa et al., 2003). Clinically, such negative outcomes might cause the patient’s health status to deteriorate, and may increase the risk of morbidity and/or mortality. Moreover, if a patient is discharged early from the ICU, his/her probability of being readmitted increases (KC et al., 2012).

While there are healthcare modeling studies of patient flow, those studies are often confined to a single department within the hospital (Cochran et al., 2006; Vanberkel et al., 2009). For example, there have been numerous studies of ICUs (Ridge et al., 1998; Kim et al., 1999; Kim et al., 2000), but all of those studies looked at the ICU independently, without considering other units within the same hospital. Although helpful, those studies do not provide a comprehensive picture of patient flow and throughput in the hospital. Therefore, studies of patient flow in the ICU and ICU-related units (the operating room (OR), the emergency department (ED), etc.) are needed, because improving efficiency and patient throughput in the ICU should significantly reduce hospital's costs, and significantly improve hospital's throughput, as well as patient health outcomes.

1.3.1 Readmission to ICU

Beginning in 2012, the Department of Health and Human Services (HHS) started adjusting payment for specific conditions by 30% for hospitals with 30-day patient readmission rates higher than the 75th percentile (HHS.gov, 2011). Furthermore, starting in 2013, HHS requires hospitals to publish their readmission rates (HHS.gov, 2011). On average, 7% of patients are readmitted to the ICU (Rosenberg et al., 2000).

Readmission is often associated with increased costs and increased risk of morbidity and mortality (Alban et al., 2006; Chrusch et al. 2009; Limathe et al., 2009). As a result, numerous studies have attempted to build models for predicting and identifying reasons for readmission to the ICU (Rosenberg et al. 2000; Bradell et al., 2003; Alban et al., 2006; Conlon et al., 2008; Litmathe et al., 2009). However, the results from those studies are different from each other, and, in some cases, contradicting (Rosenberg et al., 2000).

Rosenberg et al. (2000) reviewed eight studies that looked at ICU readmissions. Despite the fact that the readmission rates were relatively similar across the eight studies, predictors of readmission to the ICU were not consistent (Rosenberg et al., 2000). Bardell et al. (2003) analyzed reasons for readmission in cardiac patients who underwent Coronary Artery Bypass Surgery (CABG), using logistic regression. The authors found that renal failure and prolonged mechanical ventilation were the only predictors of ICU readmission after CABG. Litmathe et al. (2009) also analyzed predictors of readmission to ICU after cardiac surgery. According to their multivariate logistic regression analysis, renal failure, mechanical ventilation (>24 hours), re-exploration for bleeding and low cardiac output were found to be predictors of ICU readmission.

Studying patient flow in the ICU while accounting for readmissions presents an opportunity to gain better insight about reasons for readmission. Therefore, it should help in lowering the rate of readmissions and in minimizing the negative outcomes associated with readmissions.

As mentioned earlier, poor management of the ICU can lead to negative patient outcomes, such as morbidity and mortality. In order to successfully address and reduce such negative outcomes, we should try to identify and characterize high-risk patients. Therefore, we use data mining techniques to identify patients who are more likely to die in the ICU. Our patient status data (alive, deceased) is highly imbalanced, so we discuss the topic of imbalanced data in the next section.

1.4 IMBALANCED DATA

Imbalanced data sets, where the class of interest is rare, occur in many fields. In accounting data, fraudulent cases are rare, but their identification is critical (Chawla et al., 2004). In medicine, the early detection of rare but serious diseases and conditions such as cancer is very important (Chawla et al., 2004; Qiao et al., 2009). Other examples include risk management and text classification (Chawla et al., 2004, Van Hulse et al., 2009).

Standard classification techniques do not account for imbalanced data very well (Chawla et al., 2004; Van Hulse et al., 2007). In fact, those techniques assume that training data sets have evenly distributed classes, which often leads to under-representation or under-sampling of the rare class (Chawla et al., 2004; Yen et al., 2009). A number of solutions for the imbalanced data problem exist at the data level and the algorithmic level (Chawla et al., 2004). Data level strategies modify the information supplied to a general-purpose algorithm so that an approach that does not explicitly accommodate different class sizes or differential misclassification costs will yield useful results. Algorithmic level approaches explicitly modify the classification method but not the data set.

1.5 OBJECTIVES

This dissertation has several objectives. Given the resemblance between patient flow networks and queueing networks, using queueing network models to study and better understand patient flow is possible (Cote, 2000; Terwiesch et al., 2011). Hence, <u>the first objective</u> is to build a mathematical model for the steady state analysis of patient flow, while considering feedback flows (i.e. readmissions). As mentioned earlier, the percentage of readmitted ICU patients is significant, so it is important to account for such patients by modeling both instantaneous and delayed feedback. We build a mathematical model for an arbitrary open queueing network consisting of two M/M/1 nodes, while accounting for instantaneous and delayed feedbacks, and find the steady state solution for the network.

Because mathematical models cannot reflect real world complexities, as they become intractable, we decided to use discrete event simulation. Therefore, the <u>second objective</u> is to build a simulation model of cardiac patient flow in a network of ICUs, using discrete event simulation, while accounting for blocking and readmissions. Simulation models have been historically well established as helpful tools for ICU managers (Kim et al., 2002). According to Ferreira et al. (2008), discrete-event simulation is a "computer modeling strategy in which events are assumed to take place one at a time, with subsequent events happening exclusively after the end of the predecessor." It has two main components: a simulation clock and a future event list (Au-Yeung, 2007). The simulation clock gives the current value of time. The future event list gives the list of times of occurrence of pre-determined future events (Au-Yeung, 2007).

Discrete-event simulation has a number of advantages. It allows managers and administrators to evaluate efficiency and to examine different 'what if' scenarios (Jun et al., 1999). Moreover, it can be used to forecast the effects of changes in patient flow on resource needs (Jun et al., 1999). In addition, unlike mathematical models, discrete-event simulation can model complex patient flow in various healthcare systems (Jun et al., 1999).

Discrete event simulation has been used in various studies to model patient flow. A number of studies modeled patient flow in an accident and emergency department using discrete event simulation (Coats et al., 2001; Connelly et al., 2004). However, most of those studies lacked sufficient data (Au-Yeung, 2007). As a result, the studies were either not validated with real data, or were validated, but their results did not agree with the actual data (Au-Yeung, 2007). Our third objective is to validate our simulation model, using real patient flow data that has been collected over four years. Validation means that our simulation model could be applied in other settings such as the operating room (OR) and the emergency room (ER). Ultimately, modeling patient flow in the entire hospital becomes possible.

While the first three objectives study patient flow mainly from a managerial perspective, patient outcomes should be considered in order to better understand patient flow. Hence, the fourth objective is to compare several popular statistical and data mining techniques in classifying the discharge status (deceased, alive) of patients from the ICU. The fact that the discharge status data is highly imbalanced (2.5% deceased, 97.5% alive) makes it extremely difficult to correctly predict deceased patients.

We compare the performance of logistic regression, discriminant analysis, Classification and Regression Tree (CART) models, C5, and Support Vector Machines (SVM) in predicting the discharge status (alive or deceased, with “deceased” being the class of interest) of patients from the ICU. In order to compare the five methods, we use a variety of misclassification cost ratio (MCR) values of classifying someone in the minority class of interest (deceased) as being in the majority class (alive), to classifying someone who is actually in the majority class as being in the minority group. We use specificity, recall, precision, the F-measure, and confusion entropy (CEN) as criteria for evaluating each method’s classification performance. We also use Hand’s measure to compare the five methods. Being able to correctly classify patients should help in identifying high-risk patients who require extra care and resources, and, therefore, reduce their risk of morbidity and/or mortality.

1.6 RESEARCH CONTRIBUTIONS

This work has a number of contributions. From a methodological perspective, it builds a queueing network model of patient flow, while accounting for both instantaneous and delayed feedback. This is a major contribution, as no healthcare patient flow study has considered feedback flows. It also builds a queueing network model of a network of ICUs, using discrete event simulation. It validates the findings from the proposed simulation model with real cardiac patient flow data that has been collected over four years. Therefore, given that the findings from the model are similar or close to the findings from the real data, our model is valid and applicable in other settings. Compared to existing studies, this is considered a significant contribution as most studies done so far have been limited to small data sets (Au-Yeung, 2007). As a result, those studies are either not validated or are validated using very small datasets (Au-Yeung, 2007).

Our work also demonstrates how the use of the MCR for analyzing imbalanced medical data significantly improves the method's classification performance. Moreover, it illustrates the application of Hand's measure in a highly imbalanced medical data set. Unlike other studies that use artificial or small real data sets, we use a large, real, imbalanced data set to compare the performance of several methods using several measures. Therefore, the data imbalance issues and the difficulties in correctly classifying the minority class in our data should be a more accurate reflection of what would be encountered in real problem environments. Moreover, using an algorithmic level approach, which does not modify the data, to deal with the data imbalance issue, means that our findings are not data specific, and, hence, may be applicable to a significant number of real imbalanced data sets.

From a managerial perspective, using simulation-based queueing network models and accounting for feedback flows should help administrators in better managing the units' limited capacity by predicting the expected demand more accurately. This, in turn, will help administrators to optimally schedule patients and staff. Accordingly, hospital efficiency will improve and hospital costs will decline.

Clinically, our study should improve patients' outcome, as patient waiting time is reduced, early discharge is avoided, and risk of morbidity and/or mortality is minimized. Moreover, being able to identify patients who are more likely to die will help in providing the necessary and timely care to those patients.

1.7 CHAPTER SUMMARY

This chapter presented the background on congestion and rising costs in the US healthcare system. In particular, the problem of hospital congestion was discussed. Studying patient flow was suggested as a potential solution to the problem. In particular, the importance of studying patient flow in a network of ICUs, while accounting for feedback flows, was indicated. The problem of classifying highly imbalanced data was also discussed. The objectives and contributions of our work were identified in this chapter as well.

Our work aims to build a mathematical model, while accounting for instantaneous and delayed feedback. In addition, our work builds a discrete event simulation model of cardiac patient flow in a network of ICUs. We also identify data mining techniques that perform well in classifying highly imbalanced data sets. The major contributions of this work are the consideration of feedback flows in the mathematical model, and the validation of the proposed discrete event simulation model using a large real patient flow dataset. Managerially, the proposed model helps administrators in managing the ICUs. Clinically, our study should improve patients' outcomes, and should help in identifying patients who are at an increased risk of death.

2.0 LITERATURE REVIEW

This chapter introduces the concept of queueing networks. Queueing networks are defined and characterized. Methods used to analyze blocking in queueing networks are described. Specifically, the parametric decomposition method is discussed. The concept of feedback flows is introduced as well. The chapter also summarizes some of the earliest studies, and the existing literature on queueing network models. In addition, it reviews healthcare studies using queueing network models to model patient flow. Moreover, some of the existing literature and methods for classifying imbalanced data sets are reviewed. The purpose of this chapter is to survey the existing literature, and to identify gaps or potential areas for further research.

2.1 QUEUEING NETWORKS

According to Chao et al. (1999), a queueing network is a "system consisting of a finite number of stations that provide services to customers." Service stations are usually called nodes (Chao et al., 1999). Airport terminals, highway systems, and hospital emergency departments are all examples of queueing networks, where customers (i.e. planes, cars, patients) arrive to the system and require some form of service (i.e. plane landing/takeoff, car toll payment, patient triage) and leave the system once their service ends (Chao et al., 1999).

Queueing networks are usually characterized by: (i) the type of system (open, closed) and (ii) the node linkage (tandem, arbitrary with or without feedback flow) (Koizumi, 2002). In open queueing networks, customers arrive to and leave the system, whereas in closed queueing networks, there is a constant number of customers in the system (i.e. customers can't enter or leave the system) (Robertazzi, 1990). In tandem networks, customers flow in a single direction with a single entry and a single exit point, whereas in arbitrary networks, customers may skip certain nodes, and can leave from multiple nodes.

Furthermore, each node is characterized according to (i) its inter-arrival time distribution, (ii) its service time distribution, (iii) the number of servers, (iv) the node's maximum capacity and (v) the service discipline. A node's maximum capacity could include waiting space (i.e., a buffer) between consecutive nodes. In terms of service discipline, there are several kinds including First-In-First-Out (FIFO) and Last-In-First-Out (LIFO) (Robertazzi, 1990). In FIFO, customers are served in the order they arrive, whereas in LIFO the most recent arrival is served first (Robertazzi, 1990). Sojourn time, blocking probability, throughput, and other criteria are used as performance measures of queueing networks (Chao et al., 1999).

2.2 ANALYSIS OF QUEUEING NETWORKS

There are two main methods for analyzing queueing networks: exact and approximation (Osorio et al., 2009). Exact methods include closed form expressions and numerical evaluation of the joint stationary distribution (Osorio et al., 2009). Closed form expressions are difficult to obtain, whereas numerical evaluation requires defining all the transition rates within the network (Osorio et al., 2009). This requirement makes numerical evaluation inflexible, as changes in network topology require redefining the transition rates (Osorio et al., 2009).

Approximation methods include simulation and analytical models (Osorio et al., 2009). While large scale simulations have become possible due to major advances in computers, simulations are still associated with large costs in terms of time and resources (Chao et al., 1999). Moreover, simulations tend to be problem specific, and, therefore, their results may not be generalizable to other problems (Chao et al., 1999). However, simulation models are more realistic and more detailed than analytical models (Osorio et al., 2009). On the other hand, analytical models are simpler, more flexible and less data dependent (Osorio et al., 2009). Nevertheless, with approximation, a theoretical basis is needed to ensure that the analytical model is reasonable and close to the real solution (Chao et al. 1999). Furthermore, simplifying assumptions are sometimes needed to maintain tractability, which, in turn, might make the analytical model unrealistic and therefore inapplicable to the real world.

2.3 QUEUEING NETWORK MODELS

Jackson's network (Jackson, 1957; Jackson, 1963) is considered the most significant contribution to the development of queueing network models. Jackson was able to find a product-form steady state solution for open and closed queueing network models with a tandem or feed-forward flow (Koizumi, 2002). He showed that for an open network, the joint equilibrium distribution of the number of customers in the network is the product of the equilibrium distributions of the number of customers at each station in the network (Koizumi, 2002).

Using mathematical terminology, if we let s_i be the number of customers at station S_i, where i=1,2,3,…k; then

$$P(S_1 = s_1, S_2 = s_2, \dots, S_k = s_k) = (1 - \rho_1)\rho_1^{s_1}(1 - \rho_2)\rho_2^{s_2} \dots (1 - \rho_k)\rho_k^{s_k} \tag{2.1}$$

This is true under the following conditions:

i. Arrivals are Poisson distributed.

ii. Exponentially distributed service times.

iii. Infinite capacity queues.

iv. Stable system ($\rho = \frac{\lambda}{\mu} < 1$).

v. Fixed routing probability.

Since then, several studies have attempted to generalize Jackson's model. Melamed (1979) showed that the sum of all departure rates from a network has a Poisson distribution, while many studies considered blocking (Cohen et al., 1980; Hershey et al., 1981; Weiss et al., 1987).

2.4 QUEUEING NETWORKS WITH BLOCKING

Queueing networks could face blocking due to the finiteness of the buffers (Dallery et al., 1993). There are three types of blocking: blocking-after-service (also known as type-1 blocking or transfer blocking); blocking-before-service (also known as type-2 blocking or service blocking); and repetitive blocking (also known as type-3 blocking or rejection blocking) (Dallery et al., 1993). In blocking-after-service, a server is blocked if the destination buffer is full upon service completion. In blocking-before-service, a customer does not start service until there is available space in the destination buffer. In repetitive blocking, a customer repeatedly receives another service until the destination buffer is available (Dallery et al., 1993).

2.4.1 Decomposition method

According to Dallery et al. (1993) and Lee et al. (1998), the most commonly used approximation method for analyzing queueing networks with blocking is the decomposition method, where the network is decomposed into a set of subsystems. The decomposition method involves three steps: 1) characterizing the subsystems, 2) deriving a set of equations to identify the unknown parameters in each subsystem, and 3) deriving an algorithm for solving the sets of equations and determining the unknown parameters of each subsystem. This provides an approximation to the original system (Dallery et al., 1993).

In general, a system of S+1 servers is decomposed into S subsystems. There are two ways to represent each subsystem. One way is to represent the subsystem with a finite queue and a server being fed by an external arrival process. The second way is to represent the subsystem by two servers (upstream and downstream), separated by a finite capacity buffer, where the upstream server is never starved (Dallery et al., 1993). As for characterizing the subsystems (the second way), the service time distribution of each server can be represented by either an exponential distribution or a phase-type (PH) distribution.

Takahashi et al. (1980) proposed an approximation method for the analysis of open restricted queueing networks. Service time was assumed to be exponential and arrivals were assumed to be Poisson. Because of the large number of nodes and in order to obtain approximate node by node decomposition, the authors suggested using “pseudo-arrival rates” and “effective service rates.” Those suggestions are based on the idea that the blocking probability represents inter-dependency among the nodes. The "pseudo-arrival rate" refers to customers arriving during blocked and non-blocked time intervals, with those arriving during blocked intervals being lost. Effective service time includes both the actual service time and the holding time (in the case of blocking). The proposed method was applied to different networks, and its results were compared with those obtained from simulation and exact calculations. Comparisons showed that the proposed methodology provides a good approximation to system performance measures, such as, the blocking probability.

Altiok et al. (1987) proposed an algorithm for approximately analyzing open exponential queueing networks with blocking. The topology of the network considered by the authors did not consider deadlocks (i.e. no directed cycles in the queueing network). The proposed algorithm decomposes the queueing network into individual queues. Each individual queue is then analyzed independently as an M/PH/1/K queue. The results of the approximation algorithm are given in the form of the marginal probability distribution of number of units in each queue. The algorithm was tested in a three-node and four-node queueing network. When compared to exact numerical data, the algorithm's results had an acceptable level of error.

Perros et al. (1989) proposed a computationally more efficient version of the algorithm proposed by Altiok et al. (1987). The authors considered an open queueing network with blocking-after-service and a feed-forward configuration. Arrivals were assumed to be Poisson and service times were assumed to be exponential. Only one class of customers was considered. Customers were served according to FIFO discipline. For simplicity, the authors only considered the case where external arrivals occur to one particular queue. The earlier proposed algorithm by Altiok et al. (1987) had two main shortcomings. The first shortcoming is the large amount of time required for the accurate construction of phase-type distribution. The second shortcoming is the large amount of CPU time required. In order to address those two shortcomings, Perros et al. (1989) proposed using a two-phase Coxian distribution. That distribution has a very simple structure and can be easily applied to large networks. The algorithm starts at the last queue, because it is not blocked. Once the equilibrium probability distribution of the last queue is obtained, the blocking delay experienced by earlier queues can be represented using a two-phase Coxian distribution. When compared with the earlier proposed algorithm, the new algorithm was found to have comparable accuracy.

Lee et al. (1990) analyzed a queueing network similar to the one considered by Perros et al. (1989). However, external arrivals could occur at any server. In order to avoid feedback flows, the authors restricted the analyses to acyclic networks. Using only information from its nearest neighbor, the proposed algorithm analyzes each queue separately. As a result, the algorithm can provide marginal steady-state occupancy probability for each queue. Two parameters were considered: clearance time (actual service time and any delay) and effective inter-arrival rate. The algorithm makes several assumptions including Poisson arrivals, exponentially distributed clearance time, and no-arrivals to blocked queues. The algorithm also assumes that, upon service completion, units see destination queues in steady state. According to the authors, the proposed algorithm has several advantages. It is simple yet it can solve large networks with general topologies, and it yields accurate results whether the queues with external arrivals have finite or infinite buffers, and whether service rates are high or low. In all of the above studies (Takahashi et al., 1980; Altiok et al., 1987; Perros et al., 1989; and Lee et al., 1990), each subsystem was represented by a finite buffer and a server being fed by an external arrival process (Lee et al., 1998).

In their paper, Dallery et al. (1993) considered a tandem queueing network with blocking-after-service, and utilized the decomposition method to analyze the system. The authors characterized each subsystem according to the second way discussed earlier, where the service times of both servers was approximated by an exponential distribution. After identifying the unknown parameters (service rate), the authors considered three algorithms. Two of the proposed algorithms use equations related to the service time of either the downstream or upstream server along with the conservation of flow equation. The third algorithm uses equations for both the downstream and upstream servers. The third algorithm has the advantage of offering a symmetrical view of the decomposition, which makes it faster than the other two algorithms. The authors showed that all three algorithms are equivalent and provide a unique solution.

2.5 QUEUEING NETWORKS WITH FEEDBACK FLOWS

Feedback flows refer to returning to an earlier visited node in the network. In healthcare, feedback flows refer to the return of patients to a unit they already visited. In other words, feedback flows represent readmission to the same unit either immediately (i.e. instantaneously) or after a certain period of time (i.e. delayed).

The concept of feedback flows was first introduced by Finch (1959) and Takacs (1963). In his paper, Finch (1959) considered two cases of feedback in a network of m servers in series. In the first case, once a customer completes service at a server, it is possible to feedback to an earlier server. Finch referred to such feedback as “terminal feedback.” In the second case, once a customer completes service at a server it is possible to return to the same server with a certain probability. This type of feedback is referred to as "single service feedback." In both cases, customers were assumed to arrive randomly to the first server only, and service time was assumed to be distributed according to the negative exponential distribution. In addition, an upper limit on the number of customers in the system was set. Finch also made the assumption that the probability of feeding back is independent of the state of the system at the time of return and of the customer who just completed service. Moreover, at service completion, there was a non-zero probability of leaving the system. Finch was able to obtain the joint probability distribution of the number of customers at each service stage under equilibrium conditions.

Takacs (1963) considered a single server with Poisson arrivals and a general service time distribution. Service times were assumed to be mutually independent and identically distributed. After service completion, a customer can immediately return with a certain probability to the server for more service, or the customer may depart the system. The probability of feeding back was assumed to be independent of any other event. Takacs was able to obtain the stationary distribution of queue size, and the stationary distribution of total time spent by a customer in the system.

Since then, several studies attempted to consider more general cases of feedback flows. D'Avignon et al. (1976) considered dependent feedback flows, where feeding back may depend on the state of the system or previous feedbacks. On the other hand, Foley et al. (1983) considered delayed feedback, where a customer visits another server before returning to the earlier visited server. In their paper, Foley et al. (1983) described a network consisting of two nodes. The first node has a general service time distribution, while the second node has an exponential service time distribution. The arrival process to the first node was assumed to be Poisson, and upon service completion, units may leave the system or enter the second node. All units in the second node have to go back to the first node. The authors were able to derive properties of the time-dependent queue length process.

2.6 QUEUEING NETWORKS WITH BLOCKING AND FEEDBACK FLOWS

While there are numerous studies of queueing networks with blocking, very few of them considered feedback flows. Takahashi et al. (1980), Altiok et al. (1987), Jun et al. (1989), Perros et al. (1989) and Lee et al. (1990) all studied open queueing network models with arbitrary configurations, but only Jun et al. (1989) took feedback flows into account (Lee et al., 1998). This could be attributed to several reasons. First, according to Disney (1981), it is not theoretically appropriate to assume Poisson arrival rates whenever feedback flows exist (Koizumi, 2002). Second, deadlocks could occur, due to feedback flows and, therefore, the First-Come-First-Serve (FCFS) queue discipline could be violated when trying to resolve deadlocks (Koizumi, 2002).

In their work, Jun et al. (1989) accounted for blocking and feedback flows by assuming that blocked customers are exchanged simultaneously, so that deadlocks are resolved instantaneously. Arrivals were assumed to be Poisson, while service time was characterized by a two-phase Coxian distribution. According to their algorithm, in order for a two-phase Coxian distribution to reflect all the possible deadlocks and delays due to blocking, a very complicated phase-type distribution should be constructed first. Then, using a three-moment approximation, the phase-type distribution is simplified to the two-phase Coxian distribution. Although accurate, the algorithm is restricted to networks consisting of nodes with no more than two directly-linked upstream severs.

Lee et al. (1998) extended their earlier work (Lee et al., 1990) to account for feedback flows, and extended Dallery et al.'s (1993) symmetrical approach to open queueing networks with arbitrary configurations. Deadlocks were assumed to be resolved instantaneously by simultaneously transferring all the blocked customers. Using the decomposition method, the authors decomposed the network into a set of subsystems, and assumed that service times and inter-arrival times are characterized by a generalized exponential distribution. Each subsystem was characterized by one or many upstream servers (never starved) and one downstream server (never blocked), separated by a finite buffer. The authors identified the service rate as the unknown parameter. However, the proposed algorithm did not consider deadlocks, because, according to the authors, doing so would make the model too complicated to solve. Despite that, numerical results showed that the algorithm yields accurate results with short execution time, even when deadlocks exist, as long as they are not too frequent.

2.7 QUEUEING NETWORK MODELS IN HEALTHCARE

Queueing network models have been applied in a variety of healthcare settings to model patient flow. Albin et al. (1990) used a queueing network model to identify causes of delay in a health center appointment clinic. The authors considered an open queueing network with single server infinite capacity nodes. They used QNA, which is a software tool, for analyzing queueing networks. The analysis showed that delays were the result of scheduling problems.

Koizumi et al. (2005) analyzed patient flow in a mental health system, using a queueing network with blocking. All patients were assumed to be treated equally, and a First-Come-First-Service (FCFS) queue discipline was considered. A single-node decomposition method was utilized, and waiting space was assumed to be infinite. The authors defined total patient time, which they refer to as effective time, as the sum of both treatment time and blocking time. The number of patients waiting to enter and waiting time were used for performance evaluation. The study showed that congestion was not a cumulative effect of shortages in all facilities, but rather it was the result of shortage of beds in a particular facility. The authors suggested relying more on simulation models as they are less restrictive about the distribution of arrivals and service time.

Chaussalet et al. (2006) used a closed queueing network to model patient flow in a geriatric department. The authors assumed that the system is always full, and a bed capacity constraint was introduced. Patient service time was assumed to be exponentially distributed. According to the authors, the proposed model should help managers in estimating the long-term effect of changes in the current policies.

Au-Yeung et al. (2007) developed an Approximate Generating Function Analysis (AGFA) technique. Using this technique in queueing networks with class-based priorities, the Laplace transform of the probability density function of customer response time was approximated. The first two moments of customer response time were derived from the approximated Laplace transform. The technique was applied to an Accident and Emergency department, and its results were compared to those obtained using discrete event simulation. The technique was found to perform well under different priority schemes for mean response time. However, some discrepancies appeared when the system was saturated with high workloads. Moreover, the technique did not work as well in closed queueing networks.

Creemers et al. (2007) modeled patient flow in an orthopedic department using parametric decomposition and Brownian motion approaches. Using the Arena software package, discrete-event simulation was used as a validation tool. Arrivals were assumed to be Poisson and service times were assumed to be exponential. The authors studied the impact of outages (between and during job interruptions) on patient flow and on utilization of resources. The results showed that approximations based on decomposition approaches are better than those based on Brownian motion.

Jiang et al. (2008) analyzed the impact of care parallelization on total time spent by a patient in an urgent care center. The authors incorporated fork/join queues into a multi-class open queueing network model. A two-moment parametric decomposition approximation method was used. Unlike other studies, each node was modeled as a GI/G/1 or GI/G/m queue rather than as a M/M/1 or a M/M/m queue. The analysis showed that parallelization did not result in a significant reduction of total patient time in the system.

Litvak et al. (2008) presented a mathematical model for managing the overflow of ICU patients. The model is based on the Equivalent Random Method (ERM), developed for analyzing overflow capacity in circuit-switched telephone systems. The basic idea of ERM is to use a single Equivalent Random unit, which generates the same first two moments of overflow in the original system, to replace several multi-server units. Once established, the Erlang loss formula can be applied. The model was applied to several hospitals in the Rijnmond Region in the Netherlands. The authors considered the effect of coordination of several ICUs within this region on the fraction of regional emergency patients who were not admitted to one of the regional ICUs (i.e., blocked or rejected). Results showed that coordination helped in reducing both the fraction of rejected (blocked) regional emergency patients and the fraction of cancelled operations.

Osorio et al. (2009) studied patient flow in a network of operative and post-operative units at a university hospital. The authors used a finite capacity open queueing network model with bufferless queues. In order to capture between-queue correlation, structural parameters were utilized. The network topology and the queue capacity were considered to be exogenous. Blocking was assumed to occur after service, while service time and time between successive arrivals were assumed to be exponentially distributed. The model was solved using a parametric decomposition method. For validation purposes, the proposed method was compared to existing methods. Results were found to be comparable. The authors applied the proposed model to a network of operative and post-operative units in a university hospital. They were able to identify several sources of congestion and to quantify their impact.

It is interesting to note that only one study (Osorio et al., 2009) modeled patient flow in the ICU. However, the data set was limited to only one year. It is also worth noting that none of the above studies considered feedback flows. These findings reaffirm the need for building validated models of patient flow in the ICU while accounting for feedback flows.

In chapter 1, we introduced the topic of "imbalanced data" as our patient status data is highly imbalanced (2.5% deceased, 97.5% alive). In the next section, we review some of the existing literature on strategies for classifying imbalanced data sets.

2.8 CLASSIFYING IMBALANCED DATA SETS

As mentioned earlier, there are two main strategies to deal with imbalanced data: data level strategies and algorithmic level strategies (Chawla et al., 2004). At the data level, different forms of sampling are used, such as over-sampling and random under-sampling (Chawla et al., 2004). Chawla et al. (2002) combined under-sampling of the majority class with over-sampling of the minority class to account for imbalanced data. Unlike other over-sampling methods, their method creates synthetic minority class examples. The AUC and the Receiver Operating Characteristic (ROC) convex hull strategy were used to evaluate the performance of the proposed methodology, which was tested with C4.5, Ripper, and Naïve Bayes on nine different data sets. The results showed that the proposed methodology performs better than only under-sampling, or varying the loss ratios in Ripper, or varying the class priors in Naïve Bayes.

Padmaja et al. (2007) used a hybrid sampling-based model to detect fraud. Their hybrid sampling technique combines under-sampling with over-sampling, using their Synthetic Minority Over-sampling Technique (SMOTE). The new method was tested on an insurance dataset using C4.5, Naïve Bayes, Radial Basis Function networks, and k-nearest neighbors (k-NN). The optimal classifier was identified by the rates of true positives and true negatives. The results showed that the new method was efficient in detecting fraud.

Thongkam et al. (2009) proposed a hybrid approach for generating higher quality data sets for creating improved breast cancer survival models. Their approach consists of two steps. In the first step, an outlier filtering approach based on C-Support Vector Classification (C-SVC) is used to identify and eliminate outlier instances. In the second step, over-sampling with replacement is used to increase the number of instances in the minority class. Accuracy, sensitivity, specificity, the AUC, and the F-measure were used to evaluate the performance of the proposed approach and to compare it to other approaches. The results showed that the proposed approach improved the performance of breast cancer survival models significantly.

Van Hulse et al. (2009) conducted a comprehensive experimental analysis of 35 real-world imbalanced data sets using 11 different algorithms. The study used seven different sampling techniques, including random under-sampling, random over-sampling, one-sided selection, cluster-based oversampling, Wilson's editing, SMOTE, and border-line SMOTE. To measure the classification performance of each algorithm, the AUC, the Kolmogorov-Smirnov statistic, the geometric mean, the F-measure, accuracy, and the true positive rate were used. Analysis of Variance (ANOVA) was used to assess the statistical significance of the results. Their results showed that random under-sampling performed the best, followed by random over-sampling, SMOTE, border-line SMOTE, one sided selection, and cluster-based oversampling.

Yen et al. (2009) proposed a new under-sampling clustering-based method to increase the accuracy of predicting the minority class. The authors tested their proposed method on real and synthetic datasets. Precision, recall rate and the F measure were used to evaluate the classification accuracy. The authors showed that their new method outperforms existing methods.

At the algorithmic level, cost adjustment, among other solutions, is used (Chawla et al., 2004). Tan (2005) proposed a neighbor-weighted k-NN algorithm for classifying and categorizing text in imbalanced corpora. In Tan's study, the proposed algorithm was used to assign large weights for the neighbors of the rare class and small weights for the neighbors of the large class. The results showed a significant improvement in the classification performance.

Burez et al. (2009) studied customer churn in the service industry. The authors used cost-sensitive learning, which, for a two-class problem, assigns a higher misclassification cost for false negatives than for false positives. Basic and advanced sampling methods and boosting were used as well. The authors used six real-life customer churn data sets to evaluate the performance of the four methods. The AUC was used as an evaluation metric. The results showed that under-sampling when evaluated with the AUC leads to better prediction accuracy.

Liu et al. (2009) also tackled the problem of imbalanced text data. The authors used a simple probability based term weighting scheme that utilizes relevance indicators. The new method was tested on two data sets and compared with other weighting schemes using Support Vector Machines (SVM) and Naïve Bayes. Using this new approach showed improvement for the minority class, while the performance of the majority class was not affected.

Qiao and Liu (2009) suggested three different weighted learning procedures to account for imbalanced data. One weighting scheme is based on class proportions, while another uses class proportion information and within-group misclassification rates information. The third scheme uses adaptive weighted learning. Using multi-category SVM on simulated and real datasets, the results showed that the proposed schemes handled imbalanced data effectively.

The above literature review shows that a variety of approaches has been suggested to handle imbalanced data. However, it is not clear whether the suggested approaches are problem-specific or whether they perform well on other imbalanced data sets and, more specifically, on highly imbalanced medical data.

2.9 CHAPTER SUMMARY

This chapter introduced queueing networks and described the decomposition method that is used for the analysis of blocking in queueing networks. The concept of feedback flows was introduced. Some of the literature on queueing network models and application of queueing network models in the healthcare field was highlighted and reviewed. In addition, some of the literature on the classification of imbalanced data sets was reviewed.

From the queueing literature review, it was noted that only one study looked at patient flow in the ICU. Furthermore, the study used only one year of patient flow data. In addition, none of the healthcare studies accounted for feedback flows. These findings indicate the need for studies of patient flow in the ICU, as well as the need for validation of those models and for the consideration of feedback flows.

From the data mining literature review, it was not clear whether the proposed methods are applicable to highly imbalanced medical data sets. Therefore, it is necessary to evaluate the performance of several statistical and data mining techniques in terms of classifying highly-imbalanced medical data sets.

3.0 DATA AND ANALYSES

This chapter summarizes cardiac patients' characteristics. The chapter also provides details about the queueing network considered in terms of number of units in the network, the best distributional fit for service time at each unit and time between external arrivals to each unit, as well as the transition probability matrix. The mathematical and simulation models are described. We also describe the classification methods and measures used to evaluate the methods' performance in classifying patients' status at discharge.

3.1 DATA

Our data includes 4232 cardiac patients from a university hospital with 12,468 transfers over a period of four years (June, 2006 thru May, 2010). The average age of patients at admission was 68.18 (standard deviation = 14.98) with 60% males. In terms of complications, 4.96% had an infection, 7.61% had pneumonia and 10.11% had renal failure. Table 1 provides a summary of the descriptive statistics of the data.

Table 1. Data descriptive statistics

Variable	N = 4232
Age at admission (years)	68.18 ± 14.98
Gender	n (%)
Male	2539 (60%)
Female	1692 (40%)
Complications	
Infection	207 (4.90%)
Pneumonia	322 (7.61%)
Renal failure	428 (10.11%)

The queueing network we are considering includes seven units: a Coronary Care Unit (CCU), a Cardiac Intermediate Care Unit (CICU), a Cardiothoracic Surgical Intensive Care Unit (CTSICU), a Neurotrauma Surgical Intensive Care Unit (NTSICU), a Post Anesthesia Care Unit (PACU), the Wards and a Cardiac Catheterization Unit (CATH). In our network, the four intensive care units (the CCU, the CICU, the CTSICU and the NTSICU) are considered the internal units. The PACU, Wards, and CATH are considered the external units. Therefore, arrivals from the PACU, the Wards and the CATH are considered external arrivals. Arrivals from other ICUs are considered internal arrivals. Figure 1 shows a typical hospital flow diagram, where patients can enter the system through the emergency department (ER), or they might be referred directly for a specific service at a particular unit in the hospital. The dashed ellipse indicates the network, shown in more detail in Figure 2, which we are interested in modeling within the hospital.

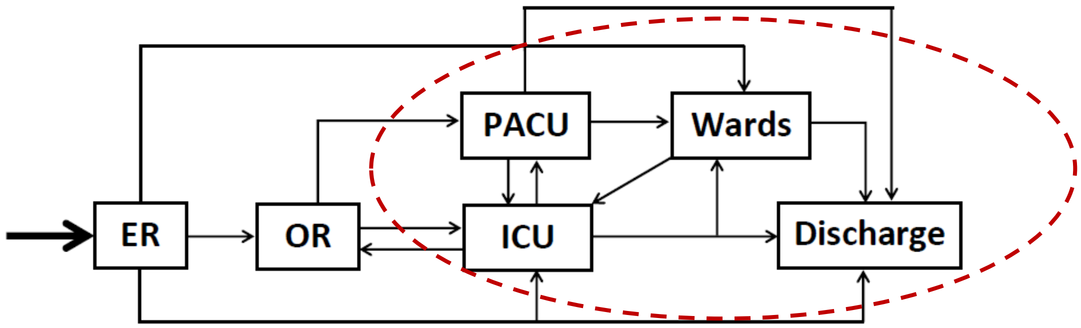


Figure 1. General diagram of patient flow in a hospital (ER='Emergency Department', OR='Operating Room', ICU='Intensive Care Unit', PACU='Post Anesthesia Care Unit')

Figure 2 shows the flow diagram of the queueing network we are modeling. The dashed rectangle indicates the internal units that we are interested in modeling. A two-headed arrow indicates that patient flow is possible in either direction between the two units, and a curved arrow indicates instantaneous feedback. Appendix A shows a diagram of sample patient flow data.

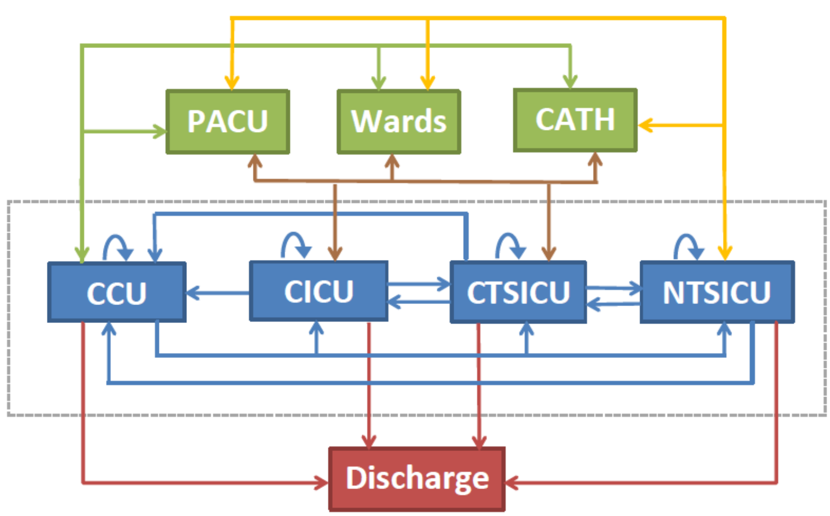


Figure 2. Flow diagram of cardiac patients in the queueing network (CCU='Coronary Care Unit'; CICU='Cardiac Intermediate Care Unit'; CTSICU='Cardiothoracic Surgical Intensive Care Unit'; NTSICU='Neurotrauma Surgical Intensive Care Unit'; PACU='Post Anesthesia Care Unit'; Wards; CATH='Cardiac Catheterization Unit')

Table 2 summarizes the transition probability matrix of cardiac patients in the queueing network. For example, the probability of a patient going from the CICU to the PACU is 36.5%, and the probability of a patient going from the NTSICU to the CTSICU is 7.7%.

Table 2. Transition probability matrix of cardiac patients

From/To	CCU	CICU	CTSICU	NTSICU	PACU	Wards	CATH
CCU	2.9%	32.5%	4.4%	0.2%	36.5%	2.2%	0.9%
CICU	9.6%	16.2%	2.9%	0%	36.5%	0.9%	8.9%
CTSICU	0.6%	0.9%	8.1%	1.7%	1.0%	65.1%	0.2%
NTSICU	0.6%	0%	7.7%	9.8%	3.7%	61.9%	0.4%
PACU	0.3%	0.3%	94.2%	5.2%	-	-	-
Wards	4.0%	4.4%	82.6%	9.0%	-	-	-
CATH	57.4%	37.4%	5.2%	0%	-	-	-

In order to check the distribution of service time in each of the four internal units, we measured the mean, standard deviation and squared coefficient of variation (CV^2) of service time in each unit. Table 3 provides the mean, standard deviation and squared coefficient of variation of service time in the CCU, the CICU, the CTSICU, and the NTSICU. It is interesting that the service time distribution in all the units could be approximated using the exponential distribution (CV^2 very close to 1).

Table 3. Mean, standard deviation, squared coefficient of variation (CV^2) of service time in Coronary Care Unit (CCU), Cardiac Intermediate Care Unit (CICU), Cardiothoracic Surgical Intensive Care Unit (CTSICU) and Neurotrauma Surgical Intensive Care Unit (NTSICU)

Unit	Service Time (hours)		
	Mean (μ)	Standard deviation (σ)	Squared coefficient of Variation (CV^2)
CCU	118.63	123.95	1.09
CICU	134.20	140.0	1.09
CTSICU	89.01	97.19	1.19
NTSICU	86.04	89.80	1.09

We also plotted the measured distribution of service time of each internal unit. Figures 3, 4, 5 and 6 show measured distribution of service time in the CCU, the CICU, the CTSICU, and the NTSICU, respectively.

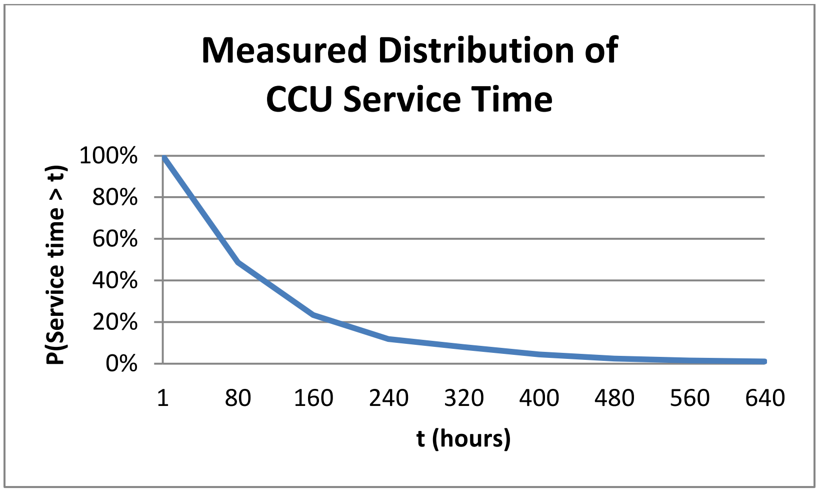


Figure 3. Measured distribution of Coronary Care Unit (CCU) service time

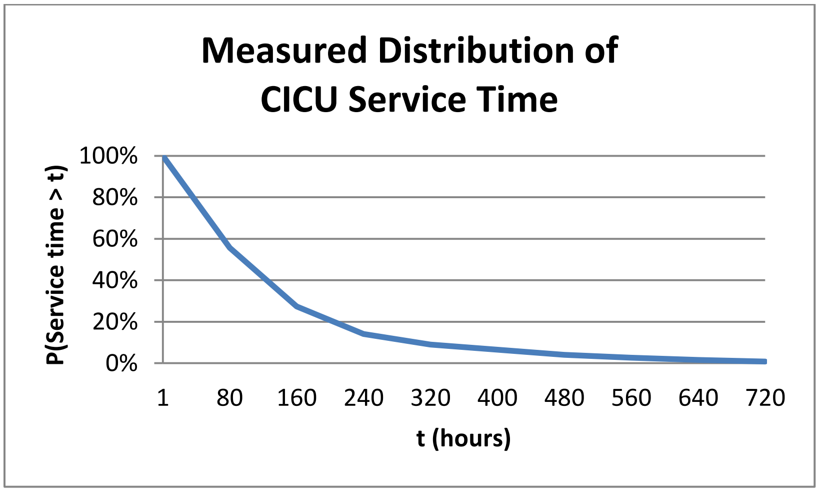


Figure 4. Measured distribution of Cardiac Intermediate Care Unit (CICU) service time

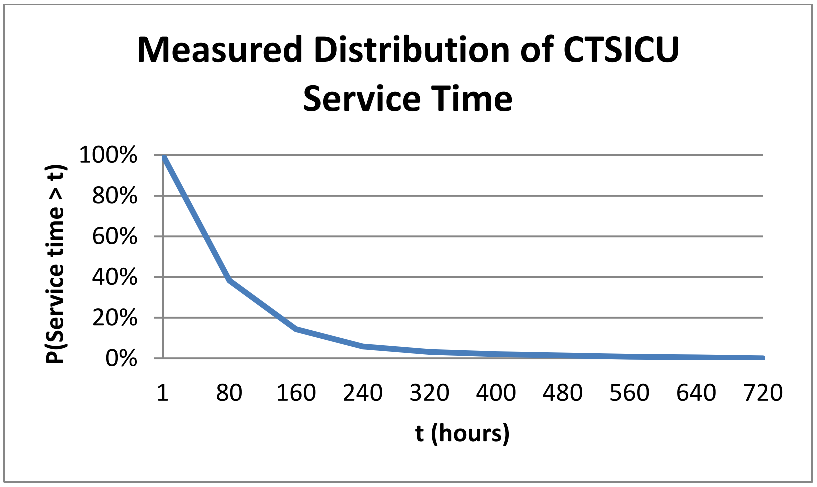


Figure 5. Measured distribution of Cardiothoracic Surgical Intensive Care Unit (CTSICU) service time

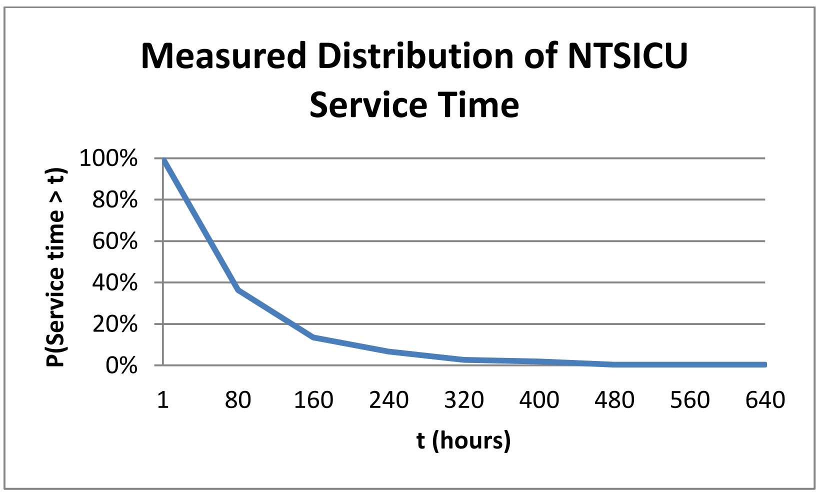


Figure 6. Measured distribution of Neurotrauma Surgical Intensive Care Unit (NTSICU) service time

We also measured the mean, standard deviation and squared coefficient of variation (CV^2) of time between external arrivals to each of the four internal units. Table 4 summarizes the mean, standard deviation and squared coefficient of variation of time between external arrivals to the CCU, the CICU, the CTSICU, and the NTSICU. Unlike the service time distributions, only the time between external arrivals distributions to the CCU could be approximated by the exponential distribution (CV^2 close to 1). Therefore, for the other three internal units (the CICU, the CTSICU, and the NTSICU), we used "Easyfit" Excel add-in to obtain the best distribution fit. Accordingly, the best distribution fit for the CICU, the CTSICU, and the NTSICU was the Gamma distribution.

Table 4. Mean, standard deviation, squared coefficient of variation (CV^2) of time between external arrivals to Coronary Care Unit (CCU), Cardiac Intermediate Care Unit (CICU), Cardiothoracic Surgical Intensive Care Unit (CTSICU) and Neurotrauma Surgical Intensive Care Unit (NTSICU)

Unit	Time between external arrivals (hours)		
	Mean (μ)	Standard Deviation (σ)	Squared Coefficient of Variation (CV^2)
CCU	189.85	201.55	1.13
CICU	271.43	366.73	1.83
CTSICU	8.11	13.63	2.83
NTSICU	128.13	468.23	13.35

The measured distribution of time between external arrivals to each of the four internal units was plotted as well. Figures 7, 8, 9 and 10 show the measured distribution of time between external arrivals to the CCU, the CICU, the CTSICU, and the NTSICU respectively.

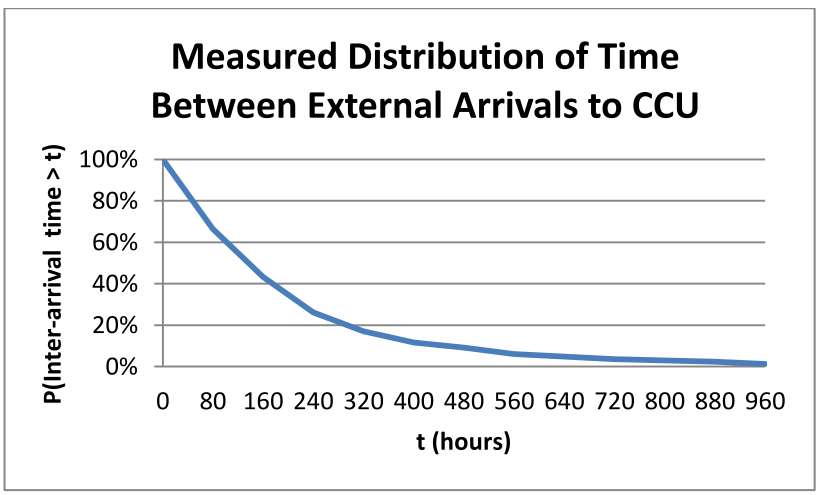


Figure 7. Measured distribution of time between external arrivals to Coronary Care Unit (CCU)

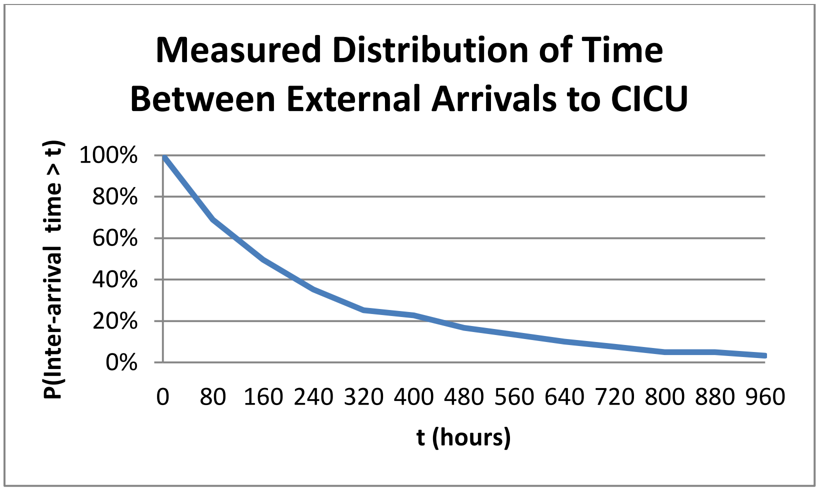


Figure 8. Measured distribution of time between external arrivals to Cardiac Intermediate Care Unit (CICU)

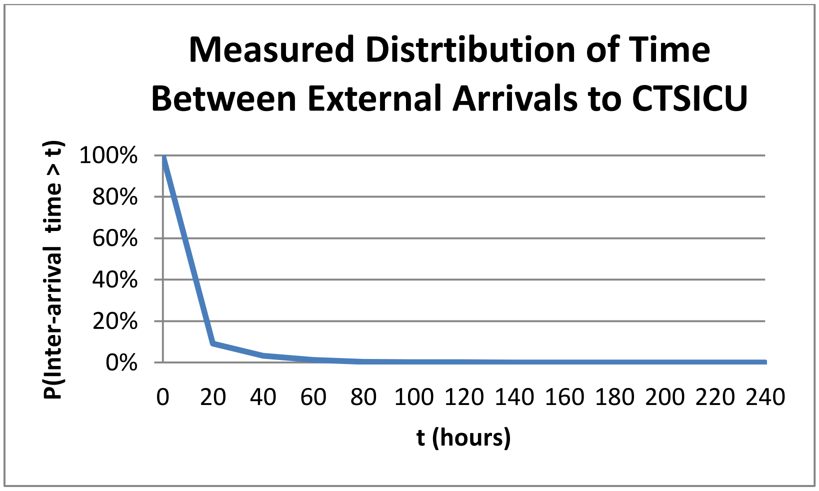


Figure 9. Measured distribution of time between external arrivals to Cardiothoracic Surgical Intensive Care Unit (CTSICU)

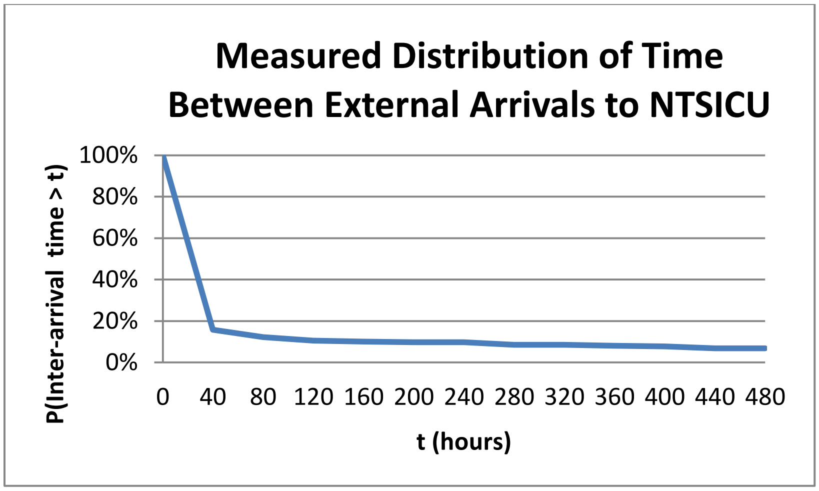


Figure 10. Measured distribution of time between external arrivals to Neurotrauma Surgical Intensive Care Unit (NTSICU)

3.2 QUEUEING NETWORK WITH FEEDBACK

For building a mathematical model of a queueing network while accounting for instantaneous and delayed feedback, we first started with the case of a single-server node with Markovian arrivals and service times (M/M/1) and instantaneous feedback. That type of network has already been studied in the literature so we compared our steady state solution to the one already published. Then, we extended the network to an arbitrary open queueing network of two M/M/1 nodes with both instantaneous and delayed feedback.

3.2.1 Open queueing network (one M/M/1 node) with instantaneous feedback

We consider an open queueing network consisting of one single server node (Figure 11) with external Poisson arrivals (r) and exponential service time (μ). We also assume that there is an infinite waiting space, a single class of patients, and that patients are served on First-Come-First-

Serve (FCFS) basis. After service completion, a patient may come back immediately to the unit with probability p (instantaneous feedback) or leave the system with probability 1-p. Our main objective is to find the steady state probability of having n patients in the system (denoted p_n).

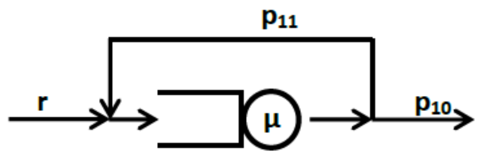


Figure 11. Open queueing network with instantaneous feedback

3.2.1.1 Solving the open queueing network (one M/M/1 node) with instantaneous feedback

In order to find the steady state solution for the queueing network in Figure 11, we first define the difference equations for $p_n(t)$. Then we find the differential-difference equations for $p_n(t)$, and we calculate the steady state solution for p_n.

In order to write the difference equations for $p_n(t)$, we consider all possible ways that the system can get to state E_n (i.e. n patients in the system) at time t + Δt while assuming that inter-arrival times and service times are both independent of each other and of the state at time t.

For $n \geq 1$:

$$p_n(t+\Delta t) = p_n(t) - \mu\,\Delta t\, p_n(t) - r\,\Delta t\, p_n(t) + p\,\mu\,\Delta t\, p_n(t) + (1-p)\mu\,\Delta t\, p_{n+1}(t) + r\,\Delta t\, p_{n-1}(t) + o(\Delta t) \quad (3.1)$$

For n = 0:

$$p_0(t+\Delta t) = p_0(t) - r\,\Delta t\, p_0(t) + (1-p)\mu\,\Delta t\, p_n(t) + o(\Delta t) \quad (3.2)$$

Dividing the difference equations (3.1 and 3.2) by Δt and taking the limit as $\Delta t \to 0$ results in the following differential-difference equations:

$$\frac{\partial p_n(t)}{\partial t} = -(\mathrm{r} + \mu)\, p_n + p\, \mu\, p_n + (1-p)\, \mu\, p_{n+1} + r\, p_{n-1} \qquad (n \geq 1) \tag{3.3}$$

$$\frac{\partial p_0(t)}{\partial t} = -r\, p_0 + (1-p)\, \mu\, p_1 \qquad (n = 0) \tag{3.4}$$

In order to get the steady state solution for $p_n(t)$, we take the limit as $t \to \infty$ which yields:

$$-(\mathrm{r} + \mu)\, p_n + p\, \mu\, p_n + (1-p)\, \mu\, p_{n+1} + r\, p_{n-1} = 0 \qquad (n \geq 1) \tag{3.5}$$

and

$$-r\, p_0 + (1-p)\, \mu\, p_1 = 0 \tag{3.6}$$

The steady state solution is shown in Chapter 4, section 4.2.

3.2.2 Open queueing network (two M/M/1 nodes) with instantaneous and delayed feedback

Next, we consider an arbitrary open queueing network consisting of two single server nodes (node 1 and node 2, respectively) (Figure 12). Each node has external Poisson arrivals (r_1 and r_2, respectively) and exponential service time (μ_1 and μ_2, respectively). We also assume that there is an infinite waiting space capacity for each node, a single class of patients, and that patients are served on a First-Come-First Serve (FCFS) basis. Each node has instantaneous feedback with probabilities p_{11} and p_{22}, respectively. P_{12} is the transfer probability from node 1 to node 2. P_{10} is the probability of leaving the system from node 1. P_{20} is the probability of leaving the system from node 2 and p_{21} is the transfer probability from node 2 to node 1.

There is also delayed feedback where a patient goes from node 1 to node 2 with probability p_{12} and then feeds back to node 1 with probability p_{21}, or a patient may externally arrive to node 2, go to node 1 with probability p_{21}, and then go from node 1 to node 2 with probability p_{12}. Both types of feedback are assumed to be independent of the state of the system and of any previous feedback. Our main objective is to find the steady state solution of having n_1 patients in node 1 and n_2 patients in node 2 (denoted $p_{n1,n2}$).

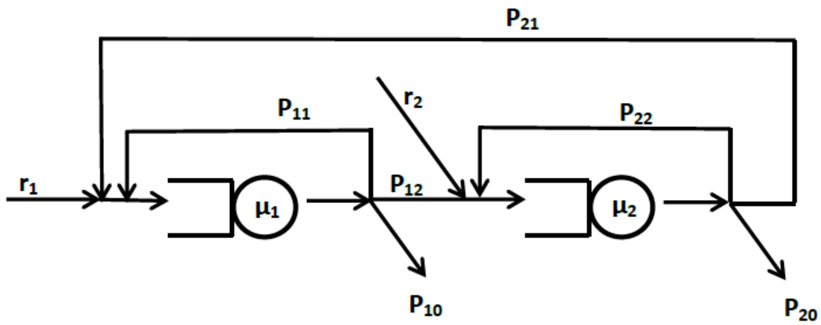


Figure 12. Open queueing network with instantaneous and delayed feedback

3.2.2.1 Solving the open queueing network (two M/M/1 nodes) with instantaneous and delayed feedback

In order to find the steady state solution for the queueing network in Figure 12, we first define the difference equations for $p_{n1,n2}(t)$. Then we find the differential-difference equations for $p_{n1,n2}(t)$ and calculate the steady state solution for $p_{n1,n2}$.

In order to write the difference equations for $p_{n1,n2}(t)$, we consider all possible ways that the system can get to state $E_{n1,n2}$ (i.e. n_1 patients in node 1 and n_2 patient in node 2) at time $t + \Delta t$, assuming that inter-arrival times and service times are both independent of each other and of the state at time t.

For $n_1 \geq 1$, $n_2 \geq 1$:

$$p_{n1,n2}(t+\Delta t) = p_{n1,n2}(t) - r_1\,\Delta t\,p_{n1,n2}(t) - \mu_1\,\Delta t\,p_{n1,n2}(t) - r_2\,\Delta t\,p_{n1,n2}(t) - \mu_2\,\Delta t\,p_{n1,n2}(t) + p_{11}\,\mu_1\,\Delta t\,p_{n1,n2}(t) + p_{22}\,\mu_2\,\Delta t\,p_{n1,n2}(t) + p_{20}\,\mu_2\,\Delta t\,p_{n1,n2+1}(t) + \mu_2\,\Delta t\,p_{n1,n2-1}(t) - r_1\,\Delta t\,p_{n1-1,n2}(t) + p_{21}\,\mu_2\,\Delta t\,p_{n1-1,n2+1}(t) + p_{12}\,\mu_1\,\Delta t\,p_{n1+1,n2-1}(t) + p_{10}\,\mu_1\,\Delta t\,p_{n1+1,n2}(t) \tag{3.7}$$

For $n_1 \geq 1$, $n_2 = 0$:

$$p_{n1,0}(t+\Delta t) = p_{n1,0}(t) - (r_1 + \mu_1 + r_2 - p_{11}\,\mu_1)\Delta t\,p_{n1,0}(t) + p_{20}\,\mu_2\,\Delta t\,p_{n1,1}(t) + p_{12}\,\mu_1\,\Delta t\,p_{n1+1,0}(t) + r_1\,\Delta t\,p_{n1-1,0}(t) + p_{21}\,\mu_2\,\Delta t\,p_{n1-1,1}(t) \tag{3.8}$$

For $n_1 = 0$, $n_2 \geq 1$:

$$p_{0,n2}(t+\Delta t) = p_{0,n2}(t) - (r_1 + \mu_2 + r_2 - p_{22}\,\mu_2)\Delta t\,p_{0,n2}(t) + p_{10}\,\mu_1\,\Delta t\,p_{1,n2}(t) + p_{20}\,\mu_2\,\Delta t\,p_{0,n2+1}(t) + r_2\,\Delta t\,p_{0,n2-1}(t) + p_{12}\,\mu_1\,\Delta t\,p_{1,n2-1}(t) \tag{3.9}$$

For $n_1 = 0$, $n_2 = 0$:

$$p_{0,0}(t+\Delta t) = p_{0,0}(t) - r_1\,\Delta t\,p_{0,0}(t) - r_2\,\Delta t\,p_{0,0}(t) + p_{10}\,\mu_1\,\Delta t\,p_{1,0}(t) + p_{22}\,\mu_2\,\Delta t\,p_{n1,n2}(t) + p_{20}\,\mu_2\,\Delta t\,p_{0,1}(t) \tag{3.10}$$

Dividing the difference equations (3.7, 3.9, 3.10 and 3.11) by Δt and taking the limit as Δt → 0 results in the following differential-difference equations:

$$\frac{\partial p_{n1,n2}(t)}{\partial t} = -(r_1 + r_2 + \mu_1 p_{12} + \mu_1 p_{10} + \mu_2 p_{20} + \mu_2 p_{21})p_{\mathrm{n1,n2}} + \mu_2 p_{20} p_{\mathrm{n1,n2+1}} + r_2 p_{\mathrm{n1,n2-1}}$$

$$+ r_1 p_{\mathrm{n1-1,n2}} + \mu_2 p_{21} p_{\mathrm{n1-1,n2+1}} + \mu_1 p_{12} p_{\mathrm{n1+1,n2-1}} + \mu_1 p_{10} p_{\mathrm{1+n1,n2}}$$

($n_1 \geq 1$, $n_2 \geq 1$)

(3.11)

$$\frac{\partial p_{n1,0}(t)}{\partial t} = -(r_1 + r_2 + \mu_1 p_1 + \mu_1 p_{10})p_{\mathrm{n1,0}} + \mu_2 p_{20} p_{\mathrm{n1,1}} + \mu_1 p_{10} p_{\mathrm{n1+1,0}} + r_1 p_{\mathrm{n1-1,0}}$$

$$+ \mu_2 p_{21} p_{\mathrm{n1-1,1}}$$

($n_1 \geq 1$, $n_2 = 0$)

(3.12)

$$\frac{\partial p_{0,n2}(t)}{\partial t} = -(r_1 + r_2 + \mu_2 p_{20} + \mu_2 p_{21})p_{\mathrm{0,n2}} + \mu_1 p_{10} p_{\mathrm{1,n2}} + \mu_2 p_{20} p_{\mathrm{0,n2+1}} + r_2 p_{\mathrm{0,n2-1}}$$

$$+ \mu_1 p_{12} p_{\mathrm{1,n2-1}}$$

($n_1 = 0$, $n_2 \geq 1$)

(3.13)

$$\frac{\partial p_{0,0}(t)}{\partial t} = -r_1 p_{0,0} - r_2 p_{0,0} + \mu_1 p_{10} p_{1,0} + \mu_2 p_{20} p_{0,1}$$

($n_1 = 0$, $n_2 = 0$)

(3.14)

In order to get the steady state solution for $p_{n1,n2}(t)$, we take the limit as $t \rightarrow \infty$, which yields:

$$-(r_1 + r_2 + \mu_1 p_{12} + \mu_1 p_{10} + \mu_2 p_{20} + \mu_2 p_{21}) p_{\mathrm{n1,n2}} + \mu_2 p_{20} p_{\mathrm{n1,n2+1}} + r_2 p_{\mathrm{n1,n2-1}} + r_1 p_{\mathrm{n1-1,n2}} + \mu_2 p_{21} p_{\mathrm{n1-1,n2+1}} + \mu_1 p_{12} p_{\mathrm{n1+1,n2-1}} + \mu_1 p_{10} p_{\mathrm{1+n1,n2}} = 0$$

($n_1 \geq 1$, $n_2 \geq 1$)

(3.15)

$$-(r_1 + r_2 + \mu_1 p_1 + \mu_1 p_{10}) p_{\mathrm{n1,0}} + \mu_2 p_{20} p_{\mathrm{n1,1}} + \mu_1 p_{10} p_{\mathrm{n1+1,0}} + r_1 p_{\mathrm{n1-1,0}} + \mu_2 p_{21} p_{\mathrm{n1-1,1}} = 0$$

($n_1 \geq 1$, $n_2 = 0$)

(3.16)

$$-(r_1 + r_2 + \mu_2 p_{20} + \mu_2 p_{21}) p_{0,\mathrm{n2}} + \mu_1 p_{10} p_{1,\mathrm{n2}} + \mu_2 p_{20} p_{0,\mathrm{n2+1}} + r_2 p_{0,\mathrm{n2-1}} + \mu_1 p_{12} p_{1,\mathrm{n2-1}} = 0$$

($n_1 = 0$, $n_2 \geq 1$)

(3.17)

$$-r_1 p_{0,0} - r_2 p_{0,0} + \mu_1 p_{10} p_{1,0} + \mu_2 p_{20} p_{0,1} = 0$$

(3.18)

The steady state solution is shown in Chapter 4, section 4.3.

Because mathematical models cannot account for real-world complexities, as they become too complicated and intractable, we built a discrete event simulation model that considers both blocking and feedback flows (section 3.3).

3.3 SIMULATION MODEL

For the second and third objectives, we built a discrete event simulation model representing the queueing network shown in Figure 2 using Arena 13.90. The model has seven units: a CCU, a CICU, a CTSICU, a NTSICU, a PACU, Wards, and a CATH. The capacity of each unit was set according to the number beds in each unit. The transfer probability from one unit to another was set equal to the transition probabilities in Table 2. Blocking-after-service was considered, where a patient does not leave his/her current unit until a bed becomes available in the next unit. No blocking was assumed to occur if a patient is discharged.

The simulation model was run for 100,000 days with a warm-up period of 50,000 days. The warm-up period is necessary to ensure that the system reaches steady state before any statistics are recorded. We ran the model twice. In the first run we assumed Markovian arrivals and service times (M/M/s) for each unit. In the second run, we assumed generally distributed inter-arrival times and exponential service times (G/M/s) for each unit. In order to validate the simulation model, we compared the results from each run to the actual patient flow data.

3.3.1 M/M/s model

3.3.1.1 Coronary Care Unit (CCU)

In the first simulation run, the CCU was assumed to have Markovian arrivals and service times. The average time between external arrivals (from all three external units combined) to the CCU was 189.85 hours. The average service time in CCU was 118.63 hours (Exp(0.005), Exp(0.008) respectively). The capacity in the CCU was set equal to 12 beds. Patients leaving the CCU to the CICU, the CTSICU, the NTSICU, the PACU, the Wards or the CATH were assumed to face blocking-after-service, where a patient remains in his/her bed until a bed becomes available at the next destination. Discharged patients were assumed to face no blocking, as bed capacity was set to infinity.

3.3.1.2 Cardiac Intermediate Care Unit (CICU)

For the CICU, we assumed an exponential distribution with an average of 271.43 hours (Exp(0.004)) for time between external arrivals, and an exponential distribution with an average of 134.20 hours (Exp(0.007)) for service time in the CICU. Bed capacity was set at 30. Blocking-after-service was assumed when patients leave the CICU to the CCU, the CTSICU, the PACU, the Wards, or the CATH. No blocking was assumed if a patient is discharged.

3.3.1.3 Cardiothoracic Surgical Intensive Care Unit (CTSICU)

For the CTSICU, the time between external arrivals was assumed to be exponentially distributed with an average of 8.11 hours (Exp(0.123)). The CTSICU service time was assumed to be exponentially distributed with an average of 89.01 hours (Exp(0.011)). We set bed capacity equal to 18 with patients facing blocking-after-service when leaving to the CCU, the CICU, the NTSICU, the PACU, the Wards, or the CATH.

3.3.1.4 Neurotrauma Surgical Intensive Care Unit (NTSICU)

For the NTSICU, the time between external arrivals was assumed to be exponential with an average of 128.13 hours (Exp(0.008)). Service time was assumed to be exponential with an average of 86.04 hours (Exp(0.012)). Bed capacity was set at 14 and after-service-blocking was assumed to occur when patients leave to the CCU, the CTSICU, the PACU, the Wards, or the CATH.

3.3.2 G/M/s model

3.3.2.1 Coronary Care Unit (CCU)

For the overall time between external arrivals (from the PACU, the Wards and the CATH combined) to the CCU, using “Easyfit”, we assumed a Gamma distribution with parameters $\alpha = 0.89$ and $\beta = 213.97$ (Gamma(213.97, 0.89)). For the service time distribution in the CCU we assumed an exponential distribution with an average service time of 118.63 hours (Exp(0.008)). The capacity in the CCU was set at 12 beds. Patients leaving the the CCU may face blocking when going to the CICU, the CTSICU, the NTSICU, the PACU, the Wards, or the CATH.

3.3.2.2 Cardiac Intermediate Care Unit (CICU)

We assumed an exponential distribution with an average of 134.20 hours (Exp(0.007)) for service time in the CICU. For combined external arrivals from all three external units to the CICU, we assumed a Gamma distribution with parameters $\alpha = 0.55$ and $\beta = 495.52$ (Gamma(495.52, 0.55)). Bed capacity was set at 30. Blocking-after-service was assumed when patients leave the CICU to the CCU, the CTSICU, the PACU, the Wards, or the CATH. No blocking was assumed to occur if a patient is discharged.

3.3.2.3 Cardiothoracic Surgical Intensive Care Unit (CTSICU)

Using “Easyfit”, we assumed a Gamma distribution for the combined time between external arrivals to CTSICU (Gamma(22.92, 0.35)). Service time in CTSICU was assumed to be exponentially distributed with an average of 89.01 hours (Exp(0.011)). We set bed capacity equal to 18 with patients facing blocking-after-service when leaving to the CCU, the CICU, the NTSICU, the PACU, the Wards, or the CATH.

3.3.2.4 Neurotrauma Surgical Intensive Care Unit (NTSICU)

As with the other units, the NTSICU time between external arrivals distribution was assumed to be Gamma with parameters $\alpha = 0.07$ and $\beta = 1711$ (Gamma(1711, 0.07)). The NTSICU’ service time was assumed to be exponential with average 86.04 hours (Exp(86.04)). Bed capacity was set at 14. After-service-blocking was assumed to occur when patients leave to the CCU, the CTSICU, the PACU, the Wards, or the CATH.

For the first three objectives, we modeled and studied patient flow mainly from a managerial perspective. However, in order to better understand patient flow, patients’ clinical outcomes such as mortality should be considered.

3.4 CLASSIFYING PATIENTS’ STATUS AT DISCHARGE

For the fourth objective, we compared several statistical and data mining techniques (logistic regression, discriminant analysis, Classification and Regression Tree (CART) models, C5, and

Support Vector Machines (SVM)) in terms of classifying ICU patients' status at discharge as either alive or deceased. For the purposes of our study, patients who died in the ICU, the minority class, constitute the group of interest, and are the "positives." Patients who were alive at the time of discharge, the majority class, are the "negatives".

We used specificity, recall, precision, the F-measure, and confusion entropy (CEN) as criteria for evaluating each method's performance. We used a variety of misclassification cost ratios (MCRs) for classifying someone in the minority class of interest (deceased) as being in the majority class (alive), to classifying someone who is actually in the majority class as being in the minority group. We used cross-validation to assess the classification performance of each method for each MCR. Using a high MCR should force a classification technique to identify the rare cases correctly more often, by making it more costly to misclassify them. We also used Hand's measure to compare the five methods.

3.4.1 Classification methods

Logistic regression is a mathematical modeling technique used to describe the probability of occurrence of one of two possible outcomes of the dependent variable and its relation to a set of predictor variables (Kleinbaum et al., 1998). Discriminant analysis is a technique used to find linear combinations of features that separate two or more groups of events (Johnson et al., 2007). CART and C5 are tree-based methods, in which the feature space is partitioned into a set of regions bounded by hyper-planes parallel to the axes and a simple model is then fit into each one. They differ in the way they grow and prune the trees (Hastie et al., 2001; Linoff et al., 2004). SVM constructs linear boundaries in a transformed version of the feature space (Linoff et al., 2004).

3.4.2 Classification measures

3.4.2.1 Specificity, recall, precision and F-measure

Specificity measures the proportion of actual negatives that are correctly specified. Recall (also known as sensitivity) measures the proportion of actual positives that are correctly specified. Precision is the proportion of identified positives that are correctly classified. The F-measure is a combination of recall and precision calculated as $\mathrm{F} = 2 * \frac{\mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$. Table 5 shows the calculation of specificity, recall and precision.

Table 5. Summary table for calculation of specificity, recall and precision

		Test Result		
Actual		Alive (-)	Deceased (+)	
	Alive (-)	True Negative (TN)	False Positive (FP)	Specificity = TN/(TN+FP)
	Deceased (+)	False Negative (FN)	True Positive (TP)	Recall = TP/(TP+FN)
			Precision = TP/(TP+FP)	

3.4.2.2 Confusion entropy (CEN)

CEN only considers the misclassified samples. The smaller the number of incorrectly classified samples, the smaller the CEN value (Wei et al., 2010). The CEN of a dataset is obtained as follows. First, the misclassification probability of each class is calculated: $\mathrm{P_j} = \frac{\sum_i \mathrm{C_{i,j}} + \sum_i \mathrm{C_{j,i}}}{2\sum_{i,j} \mathrm{C_{i,j}}}$ where i=1,2,…, j=1,2,…, $C_{i,j}$ is the number of samples from class i classified to class j (Wei et al., 2010). Second, the CEN of each class is calculated, so the CEN of class j is

$\mathrm{CEN_j} = -\sum_{i=1,i\neq j}^{N+1} \mathrm{P^j_{j,i}} \log_{2N} \mathrm{P^j_{j,i}} - \sum_{i=1}^{N+1} \mathrm{P^j_{i,j}} \log_{2N} \mathrm{P^j_{i,j}}$ where N is the number of classes,

$P_{j,i}^{j} = \frac{C_{ij}}{\sum_{i=1}^{N+1}(C_{j,i}+C_{i,j})}$ is the misclassification probability of classifying the samples of class j to other classes i subject to class j and $P_{i,j}^{j}$ is the misclassification probability of classifying other classes i to class j subject to class j (Wei et al., 2010). Then the overall CEN is calculated as the sum of the products of misclassification probabilities and the CEN of each class in the dataset (Wei et al., 2010). Appendix B shows step-by-step calculation of CEN for a data set that includes 208 patients (203 alive, 5 deceased) that was classified using SVM at MCR 1:1 and 100:1.

Even though a lower CEN value is better, because our data is highly imbalanced, CEN is expected to increase as MCR increases because even though increasing the MCR increases the number of correctly classified deceased patients, the number of misclassified alive patients also increases. There are many more alive patients than deceased ones. As a result, the total number of misclassified samples increases, hence the value of the CEN measure increases. Thus, we state the following hypothesis:

> **Hypothesis 1:** For logistic regression, discriminant analysis, CART, C5 and SVM, as MCR increases, recall, precision, the F-measure, and CEN are expected to increase, while specificity is expected to decrease.

Table 6 summarizes the expected performance of each measure as the MCR is increased. Increasing the MCR increases the penalty for misclassifying a positive, relative to the cost of misclassifying a negative, so that recall, precision, the F-measure, and CEN are expected to increase as MCR is increased, while specificity is expected to decrease as MCR is increased.

Table 6. Expected performance of specificity, recall, precision, the F-measure and confusion entropy (CEN) as the misclassification cost ratio (MCR) is increased

Measure	Expected Performance
Specificity	Decrease
Recall	Increase
F measure	Increase
Confusion entropy (CEN)	Increase

3.4.2.3 Hand's measure

A major disadvantage of the Area under the Receiver Operating Characteristic curve (AUC) is that it uses different misclassification cost distributions for different classifiers (Hand, 2009). In other words, using the AUC is equivalent to saying that the severity of misclassifying a class is different using different classifiers (Hand, 2010). Hand's measure was developed as an alternative to the AUC (Hand, 2009; Hand, 2010). Instead of specifying a particular value for the relative severity of misclassifications, Hand's measure specifies a distribution (Hand, 2009; Hand, 2010). Hand's measure values range from zero to one, where higher values represent better performance (Hand, 2009; Hand, 2010).

Hand proposed using the Beta distribution (Beta(α,β)) for the relative severity of misclassifications (Hand, 2009). Using the Beta distribution ensures that different researchers would get the same results for the same data set, while accounting for the different severities of misclassifying one class compared to misclassifying another, by varying the values of its parameters α and β (Hand, 2009). The default α=β=2 indicates that the severity is the same for misclassifying either class (Hand, 2009).

For Hand's measure, we used Beta (2, 40), Beta (40, 2) and Beta (2, 2). The choice of the Beta parameters is based on the number of deceased patients with respect to overall sample size (52/2080=0.025) which corresponds to a Beta mode of 1/40 when α = 2 and β = 40; the mode of a Beta distribution is (α-1)/(α+β-2). For comparison purposes, we also chose Beta (40, 2) and Beta (2, 2). Choosing a Beta distribution with α=40 and β=2 resembles an MCR of 1:100. For those values, the cost of misclassifying a minority class is low, so that Hand's measure value is likely to be close to zero, because there is little penalty for predicting all cases as being in the majority class. That situation is also true when α=2 and β=2, which is comparable to an MCR of 1:1. When α=2 and β=40, comparable to an MCR of 100:1, the penalty for misclassifying a minority class example is high, so that actual positives are likely to be predicted to be positives, and we expect Hand's measure values to be higher, and close to 0.5.

For the expected performance of Hand's measure as the α and β values of the Beta distribution are varied, we state the following hypothesis:

> **Hypothesis 2:** For logistic regression, discriminant analysis, CART, C5 and SVM, as α decreases and β increases, Hand's measure value is expected to increase

Table 7 summarizes the expected performance of Hand's measure for three different Beta distribution parameter choices.

Table 7. Expected performance of Hand's measure for different choices of Beta distribution parameters

Beta (α, β)	Expected Performance
(40, 2)	~ 0
(2, 2)	~ 0
(2, 40)	~ 0.5

3.5 CHAPTER SUMMARY

This chapter provides a summary of cardiac patient flow data. The network being studied was described in detail in terms of its transition probability matrix and the best distributional fits for the time between external arrivals and for the service times for each unit. The mathematical model was introduced, and the steps for finding the steady state solution were shown. The simulation model was described, as well. Moreover, the classification methods and measures used to classify patients' status at discharge were explained. The expected performance of each method under each measure was stated as a hypothesis, and summarized in tabular form.

4.0 RESULTS

This chapter presents the results of the two queueing networks we introduced in Chapter 3. The steady state solution for the first queueing network is compared with the published results, and a numerical example of the steady state solution of the second queueing network is provided. The results of the two discrete event simulation runs are presented and compared with the actual flow data. This allows for the validation of the simulation model. We also present the classification results for the discharge status of ICU patients.

4.1 QUEUEING NETWORK WITH FEEDBACK

The two queueing networks we introduced in Chapter 3 are considered "special cases" of a Jackson network. We have external Poisson arrivals, exponential service times, infinite capacity queues, an FCFS discipline, and fixed routing probabilities between servers. However, the presence of feedback loops makes the total arrival process into the server non-Poisson. We cannot make the assumption that each server is an independent M/M/1, as we would with a Jackson network. Poisson arrival processes (external and internal) are not independent Poisson processes, so their merge is not a Poisson process (Harchol-Balter, 2012). However, Harchol-Balter (2012) showed in Theorem 17.1 that a product form solution still exists for cyclic queueing networks that fit the Jackson network and that such networks can be analyzed using the balance equation approach.

The balance equation for each state is "the rate of jobs leaving the state equals the rate of job entering the state" (Harchol-Balter, 2012). For example (Harchol-Balter, 2012), consider a system in state (n_1, n_2, n_3 ,...,n_k), where n_1, n_2, … n_k represent the number of jobs at server j (j=1,2,…,k). The system leaves its current state (n_1, n_2, n_3 ,...,n_k) if there is an external arrival (r_i) or if there is a service completion in one of the servers without returning to the same server (μ_i(1-p_{ii})). The rate of jobs leaving the state (n_1, n_2, n_3 ,...,n_k) is:

$$\pi_{n1,n2,\dots,nk} \cdot \left[\sum_{i=1}^{k} r_i + \sum_{i=1}^{k} u_i(1-p_{ii})\right]$$

On the other hand, the system can enter the state (n_1, n_2, n_3 ,...,n_k) if there is an external arrival (r_i), a departure to the outside ($\mu_i p_{i,out}$) or an internal transition ($\mu_j p_{ij}$). So the rate of jobs entering the state (n_1, n_2, n_3 ,...,n_k) is:

$$\sum_{i=1}^{k} \pi_{n_1,\dots,n_{i-1},\dots,n_k} \cdot r_i + \sum_{i=1}^{k} \pi_{n_1,\dots,n_{i+1},\dots,n_k} \cdot \mu_i p_{i,out} + \sum_{i} \sum_{j \neq i} \pi_{n_1,\dots,n_{i-1},\dots,n_{j+1},\dots,n_k} \cdot \mu_j p_{j,i}$$

Therefore, for (n_1, n_2, n_3,...,n_k) state, the balance equation is:

$$\pi_{n1,n2,\dots,nk} \cdot \left[\sum_{i=1}^{k} r_i + \sum_{i=1}^{k} u_i(1-p_{ii})\right]$$

$$= \sum_{i=1}^{k} \pi_{n_1,\dots,n_{i-1},\dots,n_k} \cdot r_i + \sum_{i=1}^{k} \pi_{n_1,\dots,n_{i+1},\dots,n_k} \cdot \mu_i p_{i,out}$$

$$+ \sum_{i} \sum_{j \neq i} \pi_{n_1,\dots,n_{i-1},\dots,n_{j+1},\dots,n_k} \cdot \mu_j p_{j,i}$$

The above balance equation is only for one particular state (n_1, n_2, n_3 ,...,n_k), so we still need to consider the balance equation for all other possible states (Harchol-Balter, 2012).

4.1.1 Local balance approach

The local balance approach was suggested to simplify the large number of balance equations. The suggestion is driven by the fact that dealing with the local balance equations is much easier because they are much simpler than the global equation. Therefore, it is easier to check that a solution satisfies the equations (Harchol-Balter, 2012).

The approach works by breaking down the left and right hand sides of the balance equations into k + 1 matching components. Once you find a solution that maintains the equality for each matching component (local balance), it is also a solution for the whole equation (global balance) (Harchol-Balter, 2012).

4.2 STEADY STATE SOLUTION FOR A QUEUEING NETWORK WITH INSTANTANEOUS FEEDBACK

Using the local balance approach, we can define the total arrival rate to each node as:

$$\lambda_i = r_i + \sum_{j=1}^{k} \lambda_j p_{ji} \quad ,$$

where λ_i is the total arrival rate to node i, r_i is the external arrival rate to node i, $\sum_{j=1}^{k} \lambda_j p_{ji}$ is the internal transition to node i from all other nodes j (j=1,…k) and p_{ji} is the probability of going from node j to node i. For the total arrival rate to the node 1 we get:

$$\lambda = r + \lambda p \tag{4.1}$$

Solving for λ we get:

$$\lambda = \frac{r}{1-p} \tag{4.2}$$

According to Jackson's result (2.1), we get the following steady state solution for the queueing network model with instantaneous feedback:

$$\pi_n = \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right)$$

$$\pi_n = \left(\frac{r}{\mu(1-p)}\right)^n \left(1 - \frac{r}{\mu(1-p)}\right) \tag{4.3}$$

These results confirm the results found by Burke (1976).

4.3 STEADY STATE SOLUTION FOR A QUEUEING NETWORK WITH INSTANTANEOUS AND DELAYED FEEDBACK

Again, using the local balance approach, we can define the total arrival rate to each node as follows:

For node 1:
$$\lambda_1 = r_1 + \lambda_1 p_{11} + \lambda_2 p_{21} \tag{4.4}$$

and for node 2:
$$\lambda_2 = r_2 + \lambda_2 p_{22} + \lambda_1 p_{12} \tag{4.5}$$

Solving (4.4) for λ_1 we get:
$$\lambda_1 = \frac{r_1 + \lambda_2 p_{21}}{1 - p_{11}} \tag{4.6}$$

Solving (4.5) for λ_2 we get:
$$\lambda_2 = \frac{r_2 + \lambda_1 p_{12}}{1 - p_{22}} \tag{4.7}$$

Plugging (4.7) in (4.6) we get:
$$\lambda_1 = \frac{r_1(1-p_{22}) + r_2 p_{21}}{(1-p_{11})(1-p_{22}) - p_{12}p_{21}} \tag{4.8}$$

Plugging (4.8) in (4.7) we get:

$$\lambda_2 = \frac{r_2[(1-p_{11})(1-p_{22}) - p_{12}p_{21}] + p_{12}[r_1(1-p_{22}) + r_2 p_{21}]}{[(1-p_{11})(1-p_{22}) - p_{12}p_{21}](1-p_{22})} \tag{4.9}$$

According to Jackson's result (2.1), we get the following steady state solution for the queueing network model with instantaneous and delayed feedback:

$$\pi_{n1,n2} = \left(\frac{\lambda_1}{\mu_1}\right)^{n1}\left(1-\frac{\lambda_1}{\mu_1}\right)\left(\frac{\lambda_2}{\mu_2}\right)^{n2}\left(1-\frac{\lambda_2}{\mu_2}\right)$$

$$\pi_{n1,n2} = \left(\frac{r_1(1-p_{22})+r_2p_{21}}{\mu_1(1-p_{11})(1-p_{22})-p_{12}p_{21}}\right)^{n1}\left(1-\frac{r_1(1-p_{22})+r_2p_{21}}{\mu_1(1-p_{11})(1-p_{22})-p_{12}p_{21}}\right)$$
$$\left(\frac{r_2[(1-p_{11})(1-p_{22})-p_{12}p_{21}]+p_{12}[r_1(1-p_{22})+r_2p_{21}]}{\mu_2[(1-p_{11})(1-p_{22})-p_{12}p_{21}](1-p_{22})}\right)^{n2}$$
$$\left(1-\frac{r_2[(1-p_{11})(1-p_{22})-p_{12}p_{21}]+p_{12}[r_1(1-p_{22})+r_2p_{21}]}{\mu_2[(1-p_{11})(1-p_{22})-p_{12}p_{21}](1-p_{22})}\right)$$

(4.10)

4.3.1 Numerical example of the steady state solution

We assigned numerical values for the parameters in 4.8 and 4.9 and solved for λ_1 and λ_2. Table 8 shows the numerical values assigned for each parameter in equations 4.8 and 4.9.

Table 8. Assigned numerical values for parameters in equations 4.8 and 4.9

Parameter	Assigned Value
r_1	0.2
r_2	0.25
μ_1	0.6
μ_2	0.5
p_{11}	0.25
p_{12}	0.15
p_{21}	0.35
p_{22}	0.1

Solving for λ_1 and λ_2, we get 0.43 and 0.35 respectively. Then, using discrete event simulation we built a model similar to the network we considered in Figure 12. We assumed infinite waiting capacity at each node, an FCFS discipline, Poisson external arrivals, exponential service time at each node, and fixed routing probabilities between the two nodes. We used the same parameter values that as were assigned in equations 4.8 and 4.9. The model was run for 100,000 simulated days with a warm-up period of 50,000 days to ensure that the system reached steady state before any statistics were collected. The simulation model run results were compared with the numerical solution. They were identical. Appendix C shows a diagram of the simulation model, built in the Arena software package.

4.4 SIMULATION MODEL RESULTS

We ran two simulation models. In the first model, we assumed Markovian arrivals and service times (M/M/s) for each unit. In the second model, we assumed generally distributed inter-arrival times and exponential service times (G/M/s) for each unit.

4.4.1 Comparing mean service time

Table 9 compares the results from both simulation runs with the actual data in terms of mean service time (hours). The results of the simulation runs were compared with actual data and they were found not to be statistically significantly different (p-value < 0.05).

Table 9. Comparison of mean service time in Coronary Care Unit (CCU), Cardiac Intermediate Care Unit (CICU), Cardiothoracic Surgical Intensive Care Unit (CTSICU) and Neurotrauma Surgical Intensive Care Unit (NTSICU) between actual data and simulation runs

Unit	Actual Data Mean service time (hours)	M/M/s Mean service time (hours)	G/M/s Mean service time (hours)
CCU	118.63	118.17	116.44
CICU	134.20	133.65	134.15
CTSICU	89.01	89.02	88.99
NTSICU	86.04	84.79	85.51

4.4.2 Comparing mean time between external arrivals

Table 10 compares the results from both simulation runs with the actual data in terms of mean time between external arrivals (hours). Again, no statistically significant differences were found between each of simulation run results and actual data (p-value < 0.05).

Table 10. Comparison of mean time between external arrivals to Coronary Care Unit (CCU), Cardiac Intermediate Care Unit (CICU), Cardiothoracic Surgical Intensive Care Unit (CTSICU) and Neurotrauma Surgical Intensive Care Unit (NTSICU) between actual data and simulation runs

Unit	Actual Data Mean time between external arrivals (hours)	M/M/s Mean time between external arrivals (hours)	G/M/s Mean time between external arrivals (hours)
CCU	189.95	195.97	194.36
CICU	271.41	272.53	270.65
CTSICU	8.11	8.07	8.09
NTSICU	128.13	126.96	125.73

Given the above results, our simulation model is considered to be validated, and we can use it to check different "what if" scenarios. For example, we could test how changing the transition probabilities or arrival rates influence the mean service time or other metrics, such as utilization.

4.5 CLASSIFICATION RESULTS

Each of the five methods compared was adjusted for the following variables: age at admission to the ICU, gender, ICD-9 (International Classification of Diseases-9th edition/revision) code, surgeon code, hours of ventilation in the ICU, total time for each visit to the ICU (a patient might make multiple visits to the ICU during a single hospital stay), time between successive visits to the ICU, and reported complications while in the ICU (renal, pulmonary, vascular, infections, neurological, surgical, other). The outcome variable is a binary variable indicating the patient's status at ICU discharge (deceased=1, alive=0).

We chose the following set of MCRs for classifying a minority class member as being in the majority class to classifying someone in the majority class as being in the minority: 1:1, 5:1, 10:1, 15:1, 25:1, 40:1, 50:1, 75:1, and 100:1. We were primarily interested in determining if the misclassification of the minority (deceased) group decreased as the MCR was increased, and which method had the best improvement in correctly classifying the deceased group as a function of the MCR.

Specificity, recall, precision, the F-measure, and the CEN were calculated for each method for each MCR. The five measures were compared for each method as the MCR was increased. Figures 13 through 17 show the performance of each method for all measures as the MCR was increased.

4.5.1 Logistic regression

For logistic regression, specificity was inversely related to MCR, while recall was directly related to it. Both precision and the F-measure initially increased with the MCR, but then leveled off. CEN steadily increased as MCR increased, which is sensible because even though increasing the MCR increases the number of correctly classified deceased patients, the number of misclassified alive patients also increases. As a result, the total number of misclassified samples increases and hence the value of the CEN measure increases. Overall, using a high MCR appears to improve logistic regression's ability to correctly classify the minority (deceased) group.

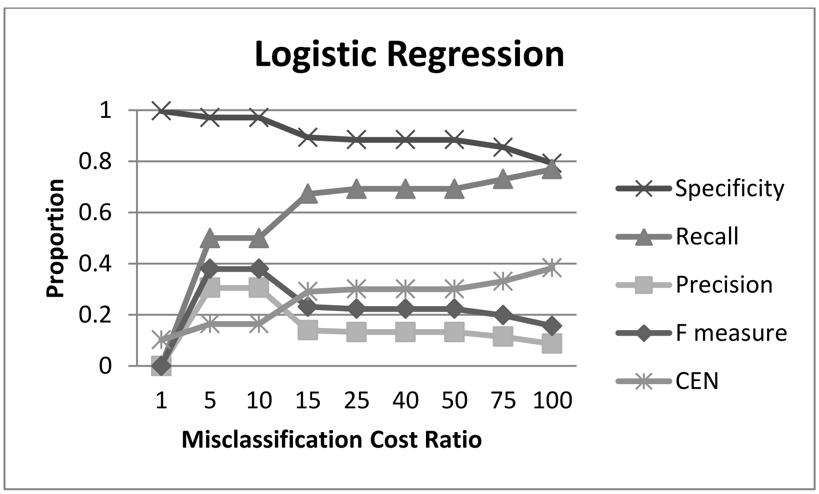


Figure 13. Logistic regression's specificity, recall, precision, the F-measure and confusion entropy (CEN) as the misclassification cost ratio (MCR) is increased

4.5.2 Discriminant analysis

Discriminant analysis had similar trends to those in logistic regression. Specificity decreased as MCR increased, while recall increased. An MCR of 50 results in a high specificity and recall (~0.78). Both precision and the F-measure increased initially with increasing MCR, and then decreased. CEN increased steadily as MCR increased. A high value for MCR appears to improve discriminant analysis' ability to identify the minority group correctly more often.

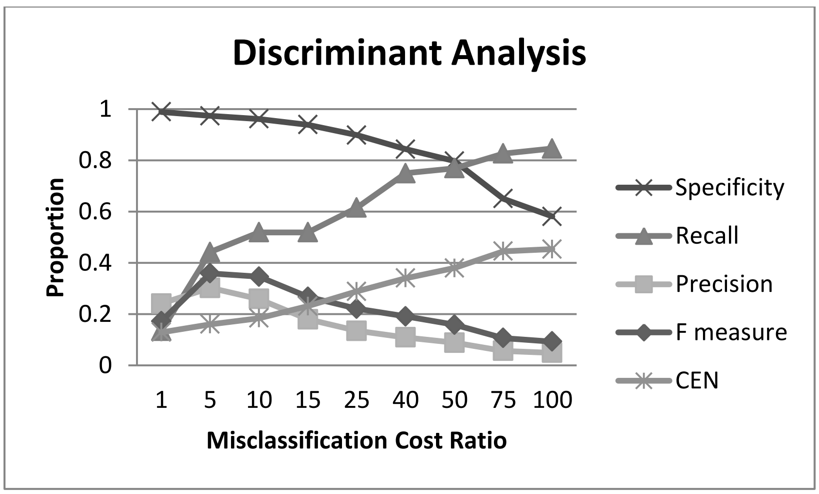


Figure 14. Discriminant analysis' specificity, recall, precision, the F-measure and confusion entropy (CEN) as the misclassification cost ratio (MCR) is increased

4.5.3 Classification and Regression Tree (CART) models

For CART, the results are not as clear-cut. Specificity remained high as MCR was increased. Recall increased initially, leveled off, and then decreased. The same was true for precision and the F-measure. The CEN increased initially and then decreased. Unlike logistic regression and discriminant analysis, increasing MCR does not appear to increase CART's ability to correctly classify cases in the minority class.

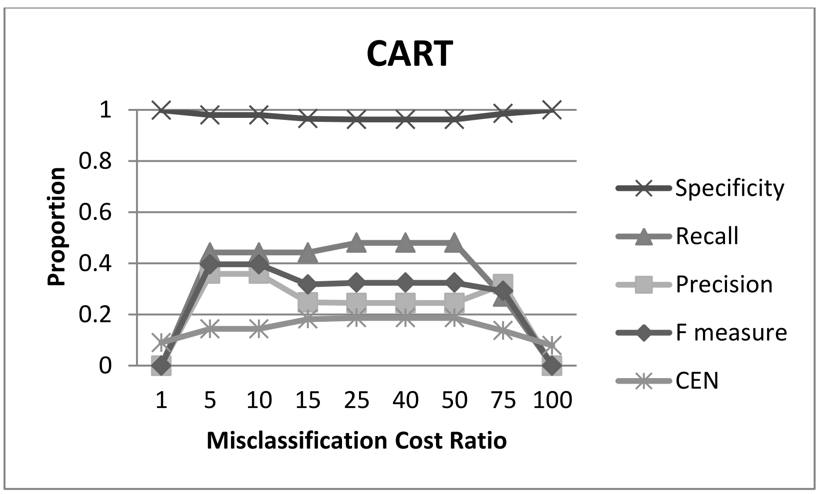


Figure 15. Classification and Regression Tree (CART) models' specificity, recall, precision, the F-measure and confusion entropy (CEN) as the misclassification cost ratio (MCR) is increased

4.5.4 C5

C5's recall increased and its specificity decreased as the MCR increased, consistent with expectations. An MCR of 40 results in a relatively high specificity and recall (~0.75). Both precision and the F-measure increased initially with the MCR, but then leveled off. CEN steadily increased as MCR increased. The behavior of C5 as shown in Figure 16 supports the contention that setting a sufficiently high value for the MCR substantially improves a method's classification of the minority group.

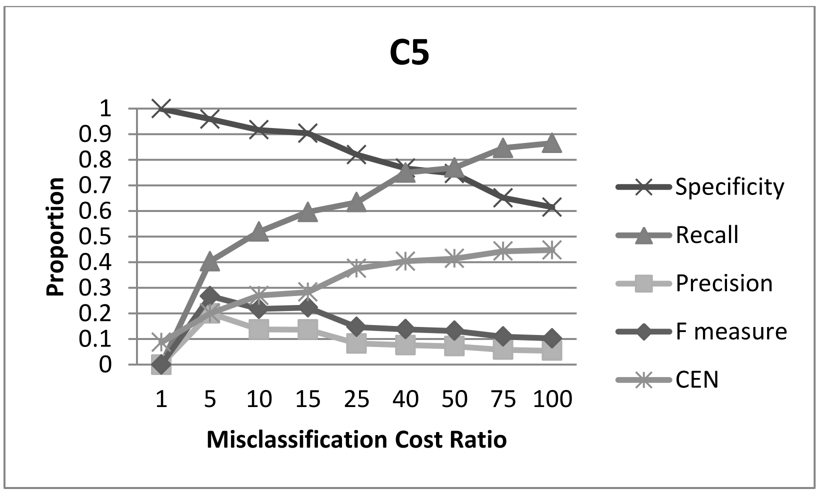


Figure 16. C5's specificity, recall, precision, the F-measure and confusion entropy (CEN) as the misclassification cost ratio (MCR) is increased

4.5.5 Support Vector Machine (SVM)

SVM's behavior as the MCR increases is similar to that of logistic regression, discriminant analysis, and C5.

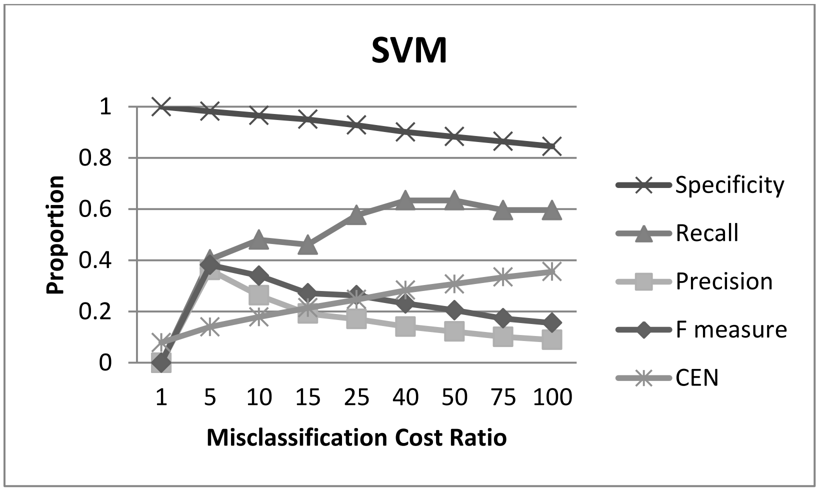


Figure 17. Support Vector Machine's (SVM) specificity, recall, precision, the F-measure and confusion entropy (CEN) as the misclassification cost ratio (MCR) is increased

Figures 13, 14, 15, 16, 17 show that for most of the algorithms, using a large value of the MCR results in the desired improvement in the classification of the minority cateogry. That finding is consistent across four methods (logistic regression, discriminant analysis, C5, and SVM) as indicated by the five measures (specificity, recall, precision, the F measure and CEN). Therefore, misclassification costs help standard classification techniques to better account for imbalanced data, by identifying the minority group correctly more often.

Figures 13 through 17 display the results for all measures for each method, permitting an analysis of the behavior of an individual algorithm. Figures 18 though 22 display the results of all methods for each measure, permitting an analysis of the behavior of each measure.

4.5.6 Specificity

Using specificity, the positive effect of MCR on the correct classification of the deceased group is consistent across the various methods (except for CART). Specificity decreased as the MCR was increased.

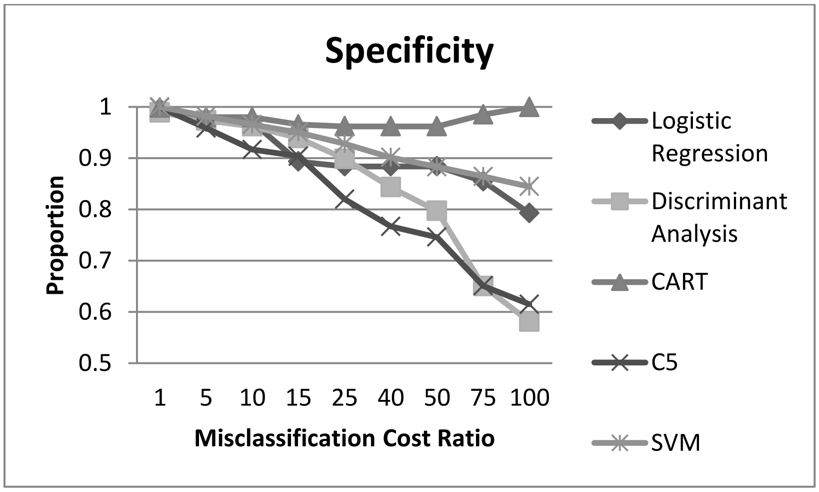


Figure 18. Specificity of logistic regression, discriminant analysis, Classification and Regression Tree (CART) models, C5 and Support Vector Machine (SVM) as the misclassification cost ratio (MCR) is increased

4.5.7 Recall

The effectiveness of the MCR in driving a method to correctly predict the minority category is evident as well in Figure 19. Recall values improved as the MCR was increased. At an MCR value of 100, recall was highest for C5. For CART, the trend was not the same – recall increased until MCR reached 50, after which it decreased.

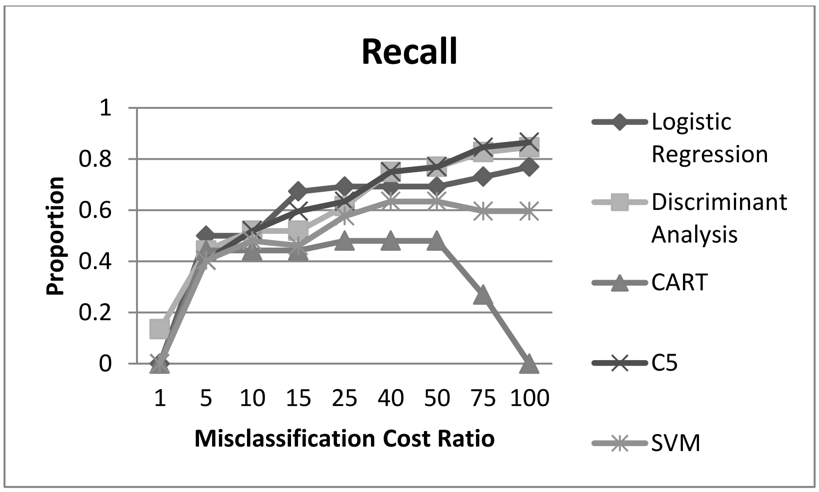


Figure 19. Recall of logistic regression, discriminant analysis, Classification and Regression Tree (CART) models, C5 and Support Vector Machine (SVM) as the misclassification cost ratio (MCR) is increased

4.5.8 Precision

All methods have generally similar trends for precision as the MCR increases. Precision initially increases, levels off, and then decreases. Discriminant analysis had the highest initial precision; CART and SVM had the highest precision overall. It appears that increasing the MCR beyond five or ten does not improve any method's precision. If precision is the measure of interest, then the effectiveness of increasing the MCR in improving minority group classification is not clear.

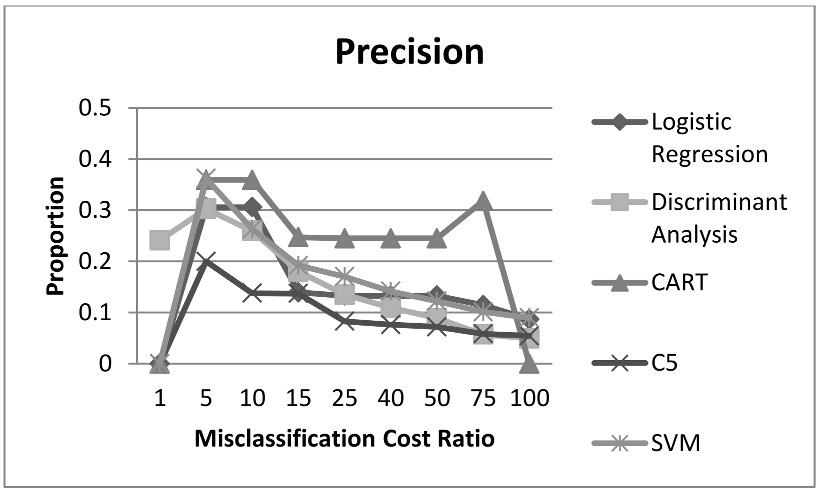


Figure 20. Precision of logistic regression, discriminant analysis, Classification and Regression Tree (CART) models, C5 and Support Vector Machine (SVM) as the misclassification cost ratio (MCR) is increased

4.5.9 F-measure

All the methods appear to have similar trends (initially increasing, leveling off, then decreasing) in their F-measures as the MCR increases. An MCR of five or ten results in the highest F-measure values across all methods. The results here are similar to those using precision.

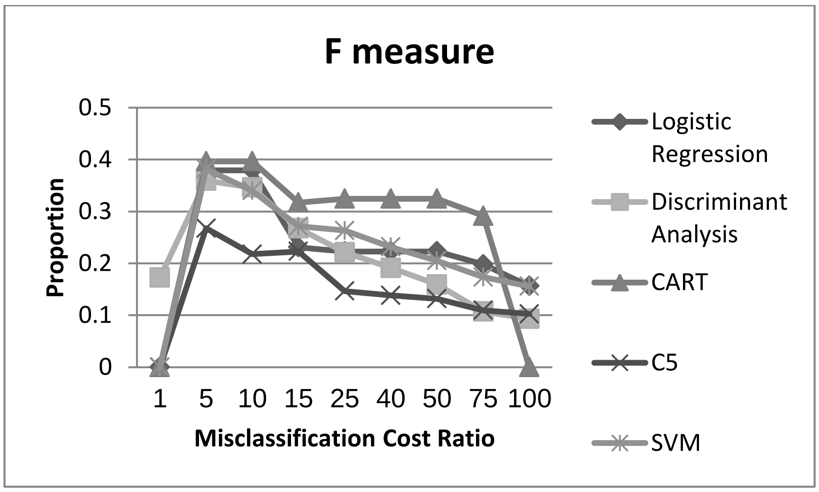


Figure 21. F-measure of logistic regression, discriminant analysis, Classification and Regression Tree (CART) models, C5 and Support Vector Machine (SVM) as the misclassification cost ratio (MCR) is increased

4.5.10 Confusion entropy (CEN)

Because a lower CEN is better, Figure 22 shows that CART provides the best CEN value for all values of the MCR. All methods except for CART show a steady deterioration in CEN with increasing MCR with the gap between CART and the best of the other methods increasing with increasing MCR.

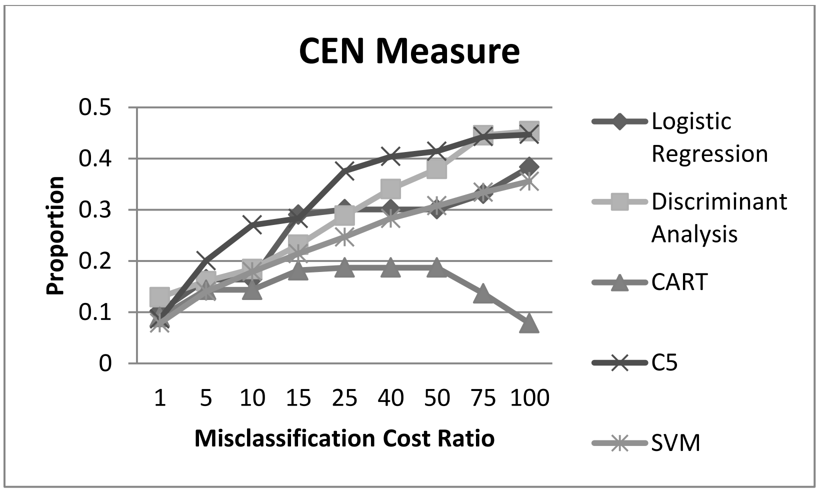


Figure 22. Confusion Entropy (CEN) of logistic regression, discriminant analysis, Classification and Regression Tree (CART) models, C5 and Support Vector Machine (SVM) as the misclassification cost ratio (MCR) is increased

Figures 18 through 22 support a conclusion that specificity, recall, and CEN are sensitive to changes in the MCR, and that increasing the MCR results in improved values for those measures. Precision and the F-measure do not appear to be sensitive to changes in the MCR. We do not have any theoretical basis for that insensitivity, but further work might produce an explanation for it. As long as the criteria of interest include specificity, recall, or CEN, there appears to be support for a strategy of using a large MCR value in order to have a minority category adequately classified by a conventional algorithm.

4.5.11 Hand's measure

Larger values of Hand's measure are preferred to smaller ones. Figure 23 shows that logistic regression and discriminant analysis are the better performing methods when Hand's measure is the criterion of interest, and that their performance is virtually the same.

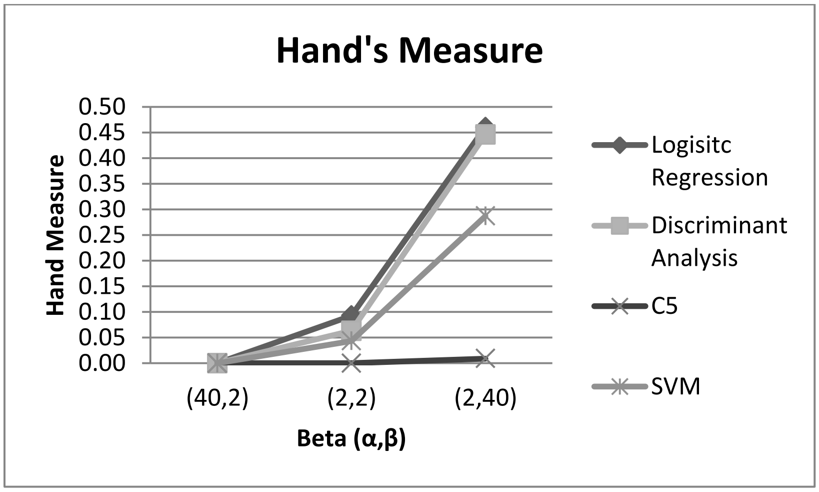


Figure 23. Hand's measure values of logistic regression, discriminant analysis, Classification and Regression Tree (CART) models, C5 and Support Vector Machine (SVM) as Beta distribution parameters are changed

Table 11 shows the Hand measure values for each method using Beta (40,2); (2,2) and (2,40).

Table 11. Hand measure values for logistic regression, discriminant analysis, Classification and Regression Tree (CART) models, C5 and Support Vector Machine (SVM) using different Beta distribution parameters

Method	Beta (α,β)		
	(40, 2)	(2, 2)	(2, 40)
Logistic regression	0	0	0.4606
Discriminant analysis	0	0.0631	0.4463
Classification and Regression Tree (CART) models	0	0	0
C5	0	0.001	0.0089
Support Vector Machine (SVM)	0	0.0435	0.2872

4.6 SUMMARY OF CLASSIFICATION RESULTS

Because there did not seem to be a significant relationship between the MCR and either precision or the F-measure, those two measures were not used to determine the best method. CART did not appear to show any clear trends in its performance measures as the MCR changed. Using specificity, recall, and CEN as criteria for comparing model performance, C5 and SVM appear to be superior to the others. At an MCR value of 100, C5 had the highest recall, while SVM had the highest specificity and lowest CEN. Using Hand's measure as a criterion for comparing each method's performance, logistic regression performed the best, with a Hand's measure value of 0.4606 using a Beta distribution with $\alpha = 2$ and $\beta = 40$.

4.7 CHAPTER SUMMARY

The steady state solutions for the two queueing networks discussed in Chapter 3 are presented. For the first queueing network with instantaneous feedback, our results match those already published in the literature. For the second queueing network with instantaneous and delayed feedback, we used the local balance approach to find the steady state solution of having n_1 and n_2 patients in nodes 1 and 2 respectively. We provided a numerical example of the steady state solution for the second queueing network.

As far as the discrete event simulation model, we compared the results, in terms of time between external arrivals to each node and service times at each node, from both simulation runs (M/M/s and G/M/s) to actual patient flow data. The simulation results were very similar and were found not to be statistically significantly different from the actual data.

We also presented the classification results of patient status at discharge from the ICU. Using MCR and specificity, recall, and CEN as criteria, C5 and SVM performed the best. When using Hand's measure as a criterion, logistic regression was found to be the best method.

5.0 CONCLUSIONS AND FUTURE RESEARCH

In this chapter, we conclude our work, and provide an overall summary of the dissertation. We reiterate the objectives that we set in Chapter 1, and point out what we found and accomplished. We also highlight the contributions and implications (research and practical) of our work. In addition, we provide ideas and directions for extending our work, and for future research.

5.1 CONCLUSIONS

In Chapter 1, we set out to achieve four main objectives. One, building a mathematical model of a queueing network with both instantaneous and delayed feedback flows, and finding the steady state solution. Two, building a discrete event simulation model of cardiac patient flow in a network of ICUs. Three, validating the discrete event simulation model using real patient flow data. Four, comparing several statistical and data mining techniques in classifying patients' discharge status from the ICU.

As far as the first objective, we were able to successfully find a steady state solution for the number of patients in nodes 1 and 2, respectively, in a two M/M/1 node open queueing network with instantaneous and delayed feedback. The solution was validated numerically and with discrete event simulation.

For the second and third objectives, we built two discrete event simulation models (M/M/s and G/M/s) and validated our results, the time between external arrivals to each unit and the service time at each unit, by comparing them to the actual patient flow data. The network we considered includes four internal units (a CCU, a CICU, a CTSICU, and a NTSICU) and three external units (a PACU, Wards, and a CATH). Comparisons showed that there were no statistically significant differences between our simulation model results and the actual data.

For the fourth and final objective, we were able to improve the classification performance (patient status at discharge) of several statistical and data mining techniques by using different MCR values, and we were able to identify the methods that performed the best according to several measures.

5.2 CONTRIBUTIONS AND IMPLICATIONS

Building a mathematical model and finding a steady state solution for a queueing network while accounting for instantaneous and delayed feedback is considered a major contribution, as we are unaware of any patient flow study that has accounted for both instantaneous and delayed feedback. Our data indicate the existence of both instantaneous and delayed feedback. Therefore, the problem we are studying is realistic which makes our results relevant and applicable to the real world.

Moreover, building and validating a discrete event simulation model for patient flow in a network of ICUs is also considered a major contribution, as most patient flow models have not been validated or have been validated, with small data sets. This validation provides the opportunity for managers and administrators to analyze the capacity and throughput of their units and to identify bottlenecks. In addition, having a validated simulation model allows the managers/administrators to test the effects of changes in patients' arrival rate on the overall system performance, and to be better able to meet demand by optimally scheduling patients, staff, and resources. This, in turn, should improve patients' status by reducing wait time, risk of morbidity and/or mortality and probability of readmission, by ensuring that patients are receiving appropriate and timely care.

Furthermore, the ability to correctly classify patients who are more likely to die will help administrators and clinicians in identifying high risk patients who require extra care and attention, and therefore reduce the likelihood of mortality in the ICU.

5.3 FUTURE WORK

Our work could be extended along several directions (methodologically and practically). Considering the non-Markovian case of queueing networks while accounting for instantaneous and delayed feedback is one possibility. Assuming different patient classes is another possibility. We could also try to classify readmitted patients, as we did with patients' discharge status. Another extension that we discuss in the next few sections is using the Analytic Hierarchy Process (AHP) to come up with optimal admission and/or discharge policies from the ICU.

5.3.1 Non-Markovian queueing network

The mathematical model we built was Markovian in terms of both time between external arrivals and service time. Upon checking our data, we found out that service times are indeed exponentially distributed in the four internal units we considered. Time between external arrivals was non-Markovian. Given that non-Markovian networks are more common, we should consider building a non-Markovian network, accounting for both instantaneous and delayed feedback. We could compare the results from the Markovian and non-Markovian queueing networks to measure the effect of considering the more realistic case of non-Markovian inter-arrival times and service times.

5.3.2 Patient classes

In both the mathematical model and discrete event simulation models we built, we assumed that all patients belong to a single class. It would be interesting to consider or assume different classes of patients.

According to our data, we can identify two main groups of patients based on the type of procedure they are undergoing. We have two main classes of patients based on the International Classification of Diseases 9th revision (ICD-9) code. The first class of patients (class 1) are those who underwent operations on the valve and septa of the heart (ICD-9 codes 35.0 thru 35.9) and the second class of patients (class 2) are those who underwent all other types of operations. The first class of patients includes 2294 patients (54.2%) and the second class of patients includes 1938 patients (45.8%). Table 12 shows some descriptive statistics for the two patient classes. It is evident that the two classes are statistically significantly different in terms of mean age, mean service time and status at discharge.

Table 12. Descriptive statistics of patient classes

Variable	Class 1 (N=2294)	Class 2 (N=1938)	p-value
Age at admission	69.85 $\pm$ 14.90	65.92 $\pm$ 14.81	< 0.0001
Gender (Male)	1402 (61.1%)	1134 (58.5%)	0.089
Status at discharge (Alive)	2162 (94.2%)	1757 (90.7%)	< 0.0001
Service time (hour)	26.23 $\pm$ 61.41	22.41 $\pm$ 46.32	0.025

We also checked the routing probabilities within the network we considered in Figure 2. We were able to find differences in the routing probabilities between the two classes. For example, the probabilities of going from the NTSICU to the PACU are 6% and 2% for patient classes 1 and 2, respectively. The probabilities of going from the NTSICU to the CTSICU are 12% and 4% in classes 1 and 2, respectively. Tables 13 and 14 show the transition probability matrices for patient classes 1 and 2, respectively.

Table 13. Transition probability matrix of class 1 patients

From/To	CCU	CICU	CTSICU	NTSICU	PACU	Wards	CATH
CCU	3.0%	32.0%	4.0%	0%	36.0%	2.0%	1.0%
CICU	10.0%	17.0%	2.0%	0%	37.0%	1.0%	8.0%
CTSICU	1.0%	1.0%	9.0%	2.0%	1.0%	65.0%	0%
NTSICU	0%	0%	12.0%	10.0%	6.0%	60.0%	0%
PACU	0%	0%	95.0%	5.0%	-	-	-
Wards	4.0%	4.0%	84.0%	8.0%	-	-	-
CATH	57.0%	37.0%	6.0%	0%	-	-	-

Table 14. Transition probability matrix of class 2 patients

From/To	CCU	CICU	CTSICU	NTSICU	PACU	Wards	CATH
CCU	2.0%	33.0%	5.0%	0%	37.0%	2.0%	1.0%
CICU	9.0%	15.0%	4.0%	0%	36.0%	1.0%	10.0%
CTSICU	1.0%	1.0%	7.0%	2.0%	1.0%	65.0%	0%
NTSICU	1.0%	0%	4.0%	10.0%	2.0%	64.0%	1.0%
PACU	0%	0%	93.0%	0%	-	-	-
Wards	4.0%	5.0%	81.0%	0%	-	-	-
CATH	58.0%	38.0%	4.0%	0%	-	-	-

Building both a mathematical model and a discrete event simulation model with both instantaneous and delayed feedback while considering the two patient classes that we just identified could be worthwhile. It would be interesting to compare our current results to those obtained while assuming different patient classes to see how decisions regarding capacity allocations or scheduling change. We might be able to identify one particular class of patients that is causing the bottleneck so instead of adding more beds to a specific unit, we can assign a specific number of beds to these patients without the need for extra beds or staff.

5.3.3 Classifying readmitted patients

Readmission rates have become a gold standard for assessing hospital quality improvement. Moreover, hospital readmissions have been getting a lot of attention from the HHS, as demonstrated by requiring hospitals to publish their readmission rates starting in 2013.

It is worth noting that all the readmission studies we reviewed in section 1.3.1 used logistic regression to predict readmissions to the ICU. Given that the results were not found to be consistent across the different studies, it would be interesting to use other methodologies, such as data mining techniques, to try to identify patients who are more likely to be readmitted. It is possible that data mining techniques will be better able to identify common causes of readmission. Because we are dealing with imbalanced readmission data (about 7% of ICU patients are readmitted), we can apply our results from the classification of patients' discharge status to the readmission data, and check whether the same identified methods would perform the best in this setting.

We believe that data mining has the potential to make a significant impact in the healthcare field. Data mining can be used to discover factors relevant to mortalities, readmissions, or other adverse patient outcomes that are likely to have an impact on hospital quality processes.

5.3.4 Admission/discharge policies

There have been numerous studies trying to address the various difficulties often faced when managing an ICU. Some studies are from a purely clinical perspective. Such studies usually try to characterize patients who are in the ICU, with the goal of better serving and eventually curing them (Henning et al. 1987, Ranucci et al. 2007). Other studies are from a managerial or operations perspective. Some researchers are interested in better understanding the ICU system as a whole in order to optimally manage the ICU by minimizing costs and ensuring the efficient use of resources (Kim et al. 1999, Kim et al. 2000, Ridge et al. 1998). Other researchers are interested in finding the optimal discharge policy, in order to accommodate the maximum number of patients while ensuring the well-being of those discharged patients (Chan et al. 2010, KC et al. 2012).

In their work "Maximizing throughput of hospital ICU with patient readmissions" Chan et al. (2010) attempted to come up with an optimal policy for deciding which patients to discharge from ICU due to capacity limitations, in order to accommodate more critical patients.

The goal of such a policy would be to maximize the number of patients entering, benefiting from, and exiting the ICU. The article had several contributions. The authors calibrated the suggested model empirically, which supports its applicability. The authors also demonstrated that the suggested policy is robust. Moreover, the authors showed that the suggested policy incurs a readmission load that is close to the readmission load incurred under an optimal policy. Furthermore, the article had a number of practical benefits. The suggested policy made effective and efficient use of ICU resources, and could be used to determine staffing levels.

The authors proposed a greedy heuristic to solve a complicated dynamic programming problem. The discharge policy that they came up with is based on discharging patients who have the lowest expected readmission load, in order to accommodate more critical patients. The authors theoretically showed that under this greedy heuristic, the readmission load incurred is close to the readmission load that would have been incurred had an optimal discharge policy been used. The authors then analyzed real empirical data from seven different private hospitals. They compared the performance of the suggested policy to other existing policies and showed that it consistently performs better than existing policies under different assumptions.

In "An econometric analysis of patient flows in the cardiac ICU" KC and Terwiesch (2012) developed an econometric model, based on data from a large US teaching hospital. The model considered patient recovery, discharge from, and potential readmission to, the ICU. The authors considered the tradeoffs often faced by decision makers in the ICU when the ICU is full. That is, whether to discharge a patient early or to cancel surgeries.

The paper had several contributions, including the estimation of the impact of ICU occupancy on patient's ICU length of stay. In particular, the authors found that, on average, a patient who is discharged from a busy ICU will have a length of stay 18% shorter than that of a patient with similar medical conditions who is discharged from a less busy ICU. The authors also found that being discharged early increases the likelihood of readmission to the ICU within the same hospital stay, and leads to longer lengths of stay on subsequent visits. Moreover, the authors demonstrated that an aggressive discharge policy frees up capacity only for low-risk patients. Finally, the authors established that an aggressive discharge policy increases revenue per bed-day for low-risk patients who are discharged from a busy ICU, but decreases it for high-risk patients who are discharged from a busy ICU.

From these two articles, it is interesting to see the different approaches that the authors utilized in order to decide on an optimal discharge policy. Unlike all the other articles, the first article (Chan et al. 2010) used empirical data from seven different hospitals to calibrate the suggested model. This is a major strength for the study as it increases the validity and generalizability of its results. However, while the proposed discharge policies proved to be effective from a managerial perspective, it is unclear whether the policies would be acceptable by the decision makers at the ICU.

Several issues should be taken into account when discharging patients such as patient conditions, hospital regulations, and hospital policies. While the proposed discharge policies take patient's conditions into account, we are not sure whether the policies can be justified. Therefore, we suggest accounting for other factors when proposing a discharge policy, in order to make such a policy easily implementable as well as acceptable. One tool that might prove effective is the Analytic Hierarchy Process (AHP).

5.4.4.1 Analytic Hierarchy Process (AHP)

Using the Analytic Hierarchy Process (AHP) for group decision making might prove extremely valuable when studying the management of the ICU. Such methods allow you to consider the opinions of the various parties involved and affected by these decisions at the same time. Therefore, reaching a decision that is acceptable by all sides becomes easier.

The application of AHP in healthcare is not new. It has been used extensively in a variety of healthcare related issues, including medical diagnosis and organ transplantation eligibility (Liberatore et al. 2008). AHP has also been used to measure the operational performance of ICUs (Dey et al., 2006). However, AHP has not been used for coming up with an optimal discharge policy for patients in the ICU. AHP could be used to address the limited capacity issues of the ICU. The AHP model can consider the guidelines for ICU discharge, as well as patients' characteristics and the opinions of all the parties involved. Such a model should help the administrators in making well-informed decisions that ensure the efficient use of the limited ICU resources and the well-being of all patients.

5.4 CHAPTER SUMMARY

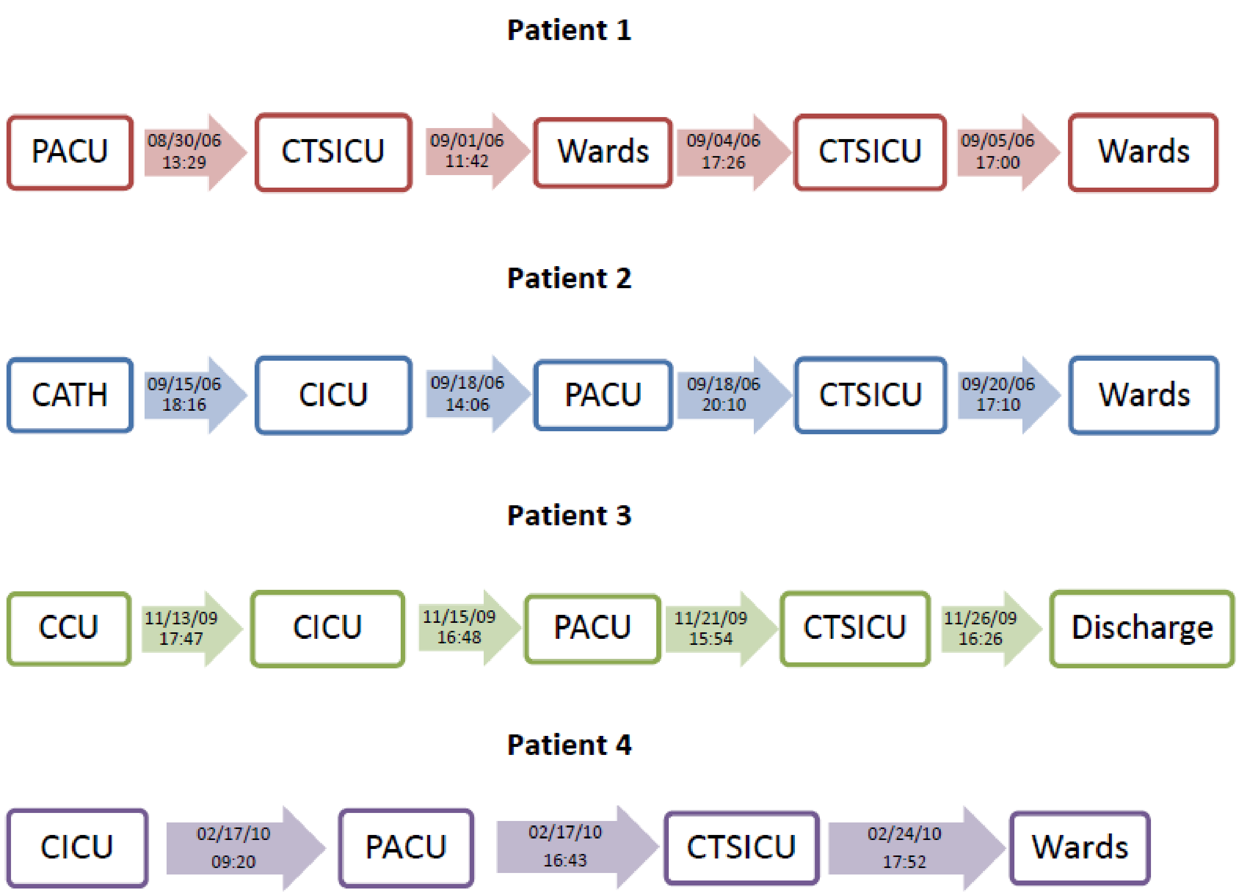
In this chapter we summarized the objectives, accomplishments and contributions of this dissertation. We were able to successfully find the steady state solution of an open queueing network, accounting for instantaneous and delayed feedback. The steady state solution was validated numerically and using simulation. We were also able to build a discrete event simulation model of patient flow in a network of ICUs, and to validate the model using real patient flow data that was collected over four years. This validation allows for the application of our model in the real world. In addition, we were able to improve the classification performance of several statistical and data mining techniques in terms of correctly identifying patients who are more likely to die in the ICU. Again, this should help clinicians and administrators in providing the necessary, appropriate and timely care to high-risk patients and reduce the risk of mortality.

This dissertation is an ongoing process as highlighted by the several directions and ideas we have for future research. There are still a lot of opportunities to extend this work and contribute on the research level and in practice. On the research level, finding a steady state solution for a non-Markovian queueing network while accounting for instantaneous and delayed feedback, and/or assuming different classes of patients, both represent significant opportunities and contributions.

On the practical level, incorporating different patient classes, identifying patients who are more likely to be readmitted, and establishing guidelines for admitting and/or discharging patients represent opportunities that should have a significant and positive impact in the real world.

APPENDIX A

SAMPLE PATIENT FLOW DATA



APPENDIX B

CONFUSION ENTROPY (CEN) CALCULATION

Support Vector Machine (SVM) classification results
for Misclassification Cost Ratio (MCR) 1:1

		Predicted		
Actual		Alive (-)	Deceased (+)	Total
	Alive (-)	203	0	203
	Deceased (+)	5	0	5
	Total	208	0	208

Support Vector Machine (SVM) classification results
for Misclassification Cost Ratio (MCR) 100:1

		Predicted		
Actual		Alive (-)	Deceased (+)	Total
	Alive (-)	171	32	203
	Deceased (+)	3	2	5
	Total	174	34	208

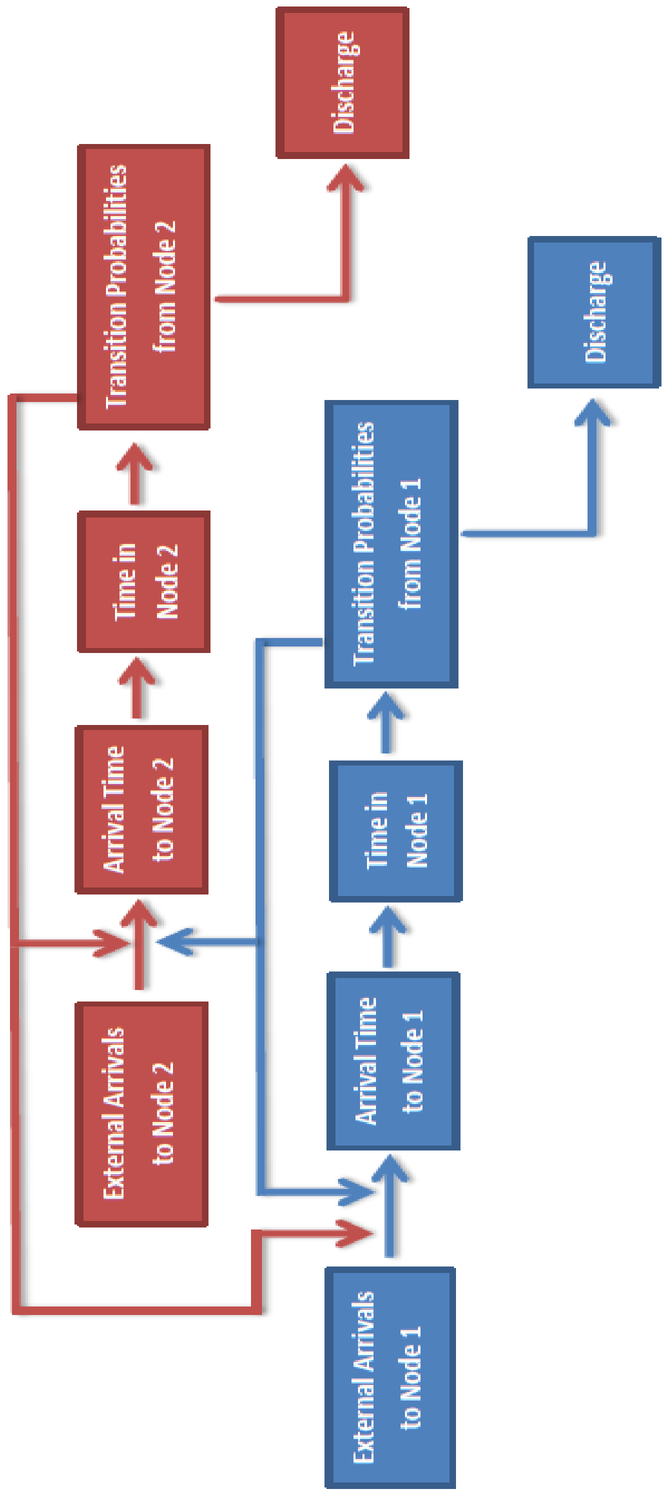
Confusion Entropy (CEN) calculation

	Misclassification Cost Ratio (MCR) 1:1	Misclassification Cost Ratio (MCR) 100:1
Misclassification probability for each class (alive = 1, deceased = 2)	$P_1 = \frac{2(203)+5}{2(208)} = 0.98$ $P_2 = \frac{5}{2(208)} = 0.01$	$P_1 = \frac{2(171)+32+3}{2(208)} = 0.91$ $P_2 = \frac{2(2)+3+32}{2(208)} = 0.09$
CEN of each class	$P_{1,2}^{1} = \frac{0}{203+208} = 0$ $P_{1,2}^{2} = \frac{5}{203+208} = 0.01$ $P_{2,1}^{1} = \frac{0}{203+208} = 0$ $P_{2,1}^{2} = \frac{5}{5} = 1$ $\mathrm{CEN}_1 = -\big((0.01)(\log_4 0.01) + (0)(\log_4 0)\big) = 0.08$ $\mathrm{CEN}_2 = \big((0)(\log_4 0) + (1)(\log_4 1)\big) = 0$	$P_{1,2}^{1} = \frac{32}{174+203} = 0.08$ $P_{1,2}^{2} = \frac{32}{34+5} = 0.82$ $P_{2,1}^{1} = \frac{3}{174+203} = 0.008$ $P_{2,1}^{2} = \frac{3}{34+5} = 0.08$ $\mathrm{CEN}_1 = -\big((0.008)(\log_4 0.008) + (0.08)(\log_4 0.08)\big) = 0.36$ $\mathrm{CEN}_2 = -\big((0.82)(\log_4 0.82) + (0.08)(\log_4 0.08)\big) = 0.51$
Overall CEN	$\mathrm{CEN} = (0.98)(0.08) + (0.01)(0) = 0.08$	$\mathrm{CEN} = (0.91)(0.36) + (0.09)(0.51) = 0.37$

APPENDIX C

DISCRETE EVENT SIMULATION MODEL

A diagram of a simple discrete event simulation model is shown on the next page


External Arrivals to Node 1
Arrival Time to Node 1
Time in Node 1
Transition Probabilities from Node 1
Discharge
External Arrivals to Node 2
Arrival Time to Node 2
Time in Node 2
Transition Probabilities from Node 2
Discharge

BIBLIOGRAPHY

Alban, R. F., Nisim, A. A., Ho, J., Nishi, G. K.,& Shabot, M. M. (2006). Readmission to surgical intensive care increases severity-adjusted patient mortality. The Journal of Trauma, 60(5), 1027-1031.

Albin, S. L., Barrett, J., Ito, D., & Mueller, J. E. (1990). A queueing network analysis of a health center. Queueing Systems, 7(1), 51-61.

Altiok, T., & Perros, H. G. (1987). Approximate analysis of arbitrary configurations of open queueing networks with blocking. Annals of Operations Research, 9(1), 481-509.

Au-Yeung, S. W. M. (2007). Response times in healthcare systems. Doctoral dissertation. Imperial College London.

Au-Yeung, S. W. M., Harrison, P. G., & Knottenbelt, W. J. (2007). Approximate queueing network analysis of patient treatment times. Proceedings of the 2nd international conference on Performance evaluation methodologies and tools.

Bardell, T., Legare, J. F., Buth, K. J., Hirsch, G. M., & Ali, I. S. (2003). ICU readmission after cardiac surgery. European Journal of Cardio-Thoracic Surgery, 23(3), 354-359.

Burez, J., & den Poel, V. (2009). Handling class imbalance in customer churn prediction. Expert Systems with Applications, 36(3), 4626-4636.

Burke, P. J. (1976). Proof of a conjecture on the interarrival-time distribution in an M/M/1 queue with feedback. IEEE Trans. Comm., 24, 575-576.

Carter, M. (2002). Diagnosis: Mismanagement of resources. OR/MS Today, 29(2),26-32.

Centers for Medicare and Medicaid Services, Office of the Actuary, National Health Statistics Group and US Department of Commerce, Bureau of Economic Analysis and Bureau of the Census.

Chao, X., Miyazawa, M., & Pinedo, M. (1999). Queueing Networks. John Wiley & Sons.

Chan, C. W., Farias, V. F., Bambos, N., & Escobar, G. J. (2010). Maximizing throughput of hospital intensive care with patient readmissions. Working Paper, Stanford University.

Chaussalet, T. J., Xie, H., & Millard, P. (2006). A closed queueing network approach to the analysis of patient flow in healthcare systems. Methods of Information in Medicine, 45(5), 492-497.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 341-378.

Chawla, N. V., Japkowicz, N., & Kolcz, A. (2004). Editorial: Special issue on learning from imbalanced data sets. SIGKDD Explorations, 6(1), 1–6.

Chrusch, C. A., Olafson, K. P., Mcmillan, P. M., Roberts, D. E., & Gray, P. R. (2009). High occupancy increases risk of early death or readmission after transfer from intensive care. Critical Care Medicine, 37(10), 2753-2758.

Coats, T. J., & Michalis, S. (2001). Mathematical modeling of patient flow through an accident and emergency department. Emergency Medicine Journal, 18,190-192.

Cochran, J. K., & Bharti, A. (2006). Stochastic bed balancing of an obstetrics hospital. Health Care Management Science, 9(1), 31-45.

Cohen, M. A., Hershey, J. C., & Weiss, E. N. (1980). Analysis of capacity decisions for progressive patient care hospital facilities. Health Services Research, 15, 145-160.

Conlon, N., O'Brien, B., Herbison, G. P., & Marsh, B. (2008). Long-term functional outcome and performance status after intensive care unit re-admission: a prospective survey. British Journal of Anesthesia, 100(2): 219-223.

Connelly, L. G., & Bair, A. E. (2004). Discrete event simulation of emergency department activity: A platform of system-level operations research. Academic Emergency Medicine, 11(11), 1177-1185.

Costa, A. X., Ridley, S. A., Shahani, A. K., Harper, P. R., De Senna, V., & Nielsen, M. S. (2003). Mathematical modeling and simulation for planning critical care capacity. Anesthesia, 58, 320-327.

Cote, M. J. (2000). Understanding patient flow. Decision Line, 31, 8-10.

Creemers, S., & Lambrecht, M. R. (2007). Modeling a healthcare system as a queueing network: The case of a Belgian hospital.

Dallery, Y., & Frein, Y. (1993). On decomposition methods for tandem queueing networks with blocking. Operations Research, 41(2), 386-399.

D'Avignon, G. R., & Disney, R. L. (1976). Single server queues with state dependent feedback, INFORS, 14, 168-180.

Dey, P. K., Hariharan, S., & Clegg, B. T. (2006). Measuring the operational performance of intensive care units using the analytic hierarchy approach. International Journal of Operations & Production Management, 26(8), 849-865.

Disney, R. L. (1981). Queuing networks. American Mathematics Society, Proceedings of Symposia in Applied Mathematics, 25, 53-83.

Dobson, G., Lee, H-H., & Pinker, E. (2010). A model of ICU bumping. Operations Research, 58(6), 1564-1576.

Ferreira, R. B., Coelli, F. C., Pereira, W. C. A., & Almeida, R. M. V. R. (2008). Optimizing patient flow in a large hospital surgical centre by means of discrete-event computer simulation models. Journal of Evaluation in Clinical Practice, 14(6), 1031-1037.

Finch, P. D. (1959). Cyclic queues with feedback. Journal of Royal Statistical Society, B, 21(1), 1-6.

Foley, R. D., & Disney, R. L. (1983). Queues with delayed feedback. Advanced Applied Probability, 15, 162-182.

Green, L. V. (2002). How many hospital beds? Inquiry, 39(4), 400–412.

Gruenberg, D. A., Shelton, W., Rose, S. L., Rutter, A. E., Socaris, S., & McGee, G. (2006). Factors influencing length of stay in the intensive care unit. American Journal of Critical Care 15(5), 502-509.

Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. Machine Learning, 77(1): 103-123.

Hand, D. J. (2010). Evaluating diagnostic tests: The area under the ROC curve and the balance of the errors. Statistics in Medicine, 29(14): 1502-1510.

Harchol-Balter, M. (2012). Performance modeling and design of computer systems.

Henning, R. J., McClish, D., Daly, B., Nearman, H., Franklin, C., & Jackson, D. (1987). Clinical characteristics and resource utilization of ICU patients: Implications for organization of intensive care. Critical Care Medicine, 15(3), 264-269.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). The elements of statistical learning. New York: Springer.

Hershey, J. C., Weiss, E. N., & Cohen, M. A. (1981). A stochastic service network model with applications to hospital facilities. Operations Research, 29, 1-22.

Jackson, J. R. (1963). Jobshop-like queueing systems. Management Science, 10, 131-142.

Jackson, J. R. (1957). Networks of waiting lines. Operations Research, 5(4), 518-521.

Jiang, L., & Giachetti, R. E. (2008). A queueing network model to analyze the impact of parallelization of care on patient cycle time. Healthcare Management Science, 11(3), 248-261.

Johnson, R. A., Wichern, & D. W. (2007). Applied multivariate statistical analysis. Upper Saddle River, New Jersey: Pearson Prentice Hall

Joint Commission Resources. (2004). Managing Patient Flow.

Jun, J. B., Jacobson, S. H., & Swisher, J. R. (1999). Application of discrete-event simulation in healthcare clinics: A survey. The Journal of Operational Research Society, 50(2), 109-123.

Jun, K. P., & Perros, H. G. (1989). Approximate analysis of arbitrary configuration of queueing networks with blocking. Proceedings of 1st International Workshop on Queueing Networks with Blockings. North-Holland, Amsterdam, 259-280.

KC, S. D., & Terwiesch C. (2012). An econometric analysis of patient flows in the cardiac intensive care unit. Manufacturing & Service Operations Management, 14(1), 50-65.

Kim, S. C., Horowitz, I., Young, K. K., & Buckley, T. A. (1999). Analysis of capacity management of the intensive care unit in a hospital. European Journal of Operational Research 115 (1), 36-46.

Kim, S. C., Horowitz, I., Young, K. K., & Buckley T. A. (2000). Flexible bed allocation and performance in the intensive care unit. Journal of Operations Management 18, 427-443.

Koizumi, N. (2002). A queueing network model with blocking: Analysis of congested patients flows in mental health systems. Doctoral dissertation. University of Pennsylvania.

Koizumi, N., Kuno, E., & Smith, T. E. (2005). Modeling patient flows using a queueing network with blocking. Healthcare Management Science, 8(1), 49-60.

Koo, P-H., Nielsen, K. B., Jang, J., & Kolker, A. (2010). Simulation-based patient flow analysis in an endoscopy unit. IEEE Workshop on Healthcare Management, Feb. 18-20, 2010. Venice, Italy.

Kleinbaum, D. G., Kupper, L. L., Muller, K. E., & Nizam, A (1998). Applied regression analysis and other multivariable methods. Pacific Grove, California: Brooks/Cole Publishing Company

Lee, H. S., Bouhchouch, A., Dallery, Y., & Frein, Y. (1998). Performance evaluation of open queueing networks arbitrary configuration and finite buffers. Annals of Operations Research, 79(0), 181-206.

Lee, H. S., & Pollock, S. M. (1990). Approximate analysis of open acyclic exponential queueing networks with blocking. Operations Research, 38(6), 1123-1134.

Liberatore, M. J., & Nydick, R. L. (2008). The analytic hierarchy process in medical and health care decision making: A literature review. European Journal of Operational Research 189(1), 194-207.

Linoff, G. S., & Berry, MJA (2004). Data Mining Techniques. Indianapolis, Indiana: Wiley Publishing, Inc.

Litmathe, J., Kurt, M., Feindt, P., Gams, E., & Boeken, U. (2009). Predictors and outcome of ICU readmission after cardiac surgery. The Thoracic and Cardiovascular Surgery Journal, 57(7), 391-394.

Litvak, N., van Rijsbergen, M., Boucherie, R. J., & van Houdenhoven, M. (2008). Managing the overflow of intensive care patients. European Journal of Operational Research, 185(3), 998-1010.

Liu, Y., Loh, H. T., & Sun, A. (2009). Imbalanced text classification: A term weighting approach. Expert Systems with Applications 36(1), 690-701.

Melamed, B. (1979). Characterization of Poisson traffic system in Jackson queueing networks. Advances in Applied Probability, 11, 422-438.

Osorio, C., & Bierlaire, M. (2009). An analytic finite capacity queueing network model capturing the propagation of congestion and blocking. European Journal of Operational Research, 196(3), 996-1007.

Padmaja, T. M., Dhulipalla, N., Krishan, P. R., Bapi, R. S., & Laha, A. (2007). An imbalanced data classification model using hybrid sampling technique for fraud detection. Lecture Notes in Computer Science, doi: 10.1007/978-3-540-77046-6.

Perros, H. G., & Snyder, P. M. (1989). A computationally efficient approximation algorithm for feed-forward open queueing networks with blocking. Performance Evaluation, 9(3), 217-224.

Pronovost, P. J., Needham, D. M., Waters, H., Birkmeyer, C. M., Calinawan, J. R., Birkmeyer, J.D., & Dorman, T. (2004). Intensive care unit physician staffing: Financial modeling of the leapfrog standard. Critical Care Medicine, 32 (6), 1247-1253.

Qiao, X., & Liu, Y. (2009). Adaptive weighted learning for imbalanced multicategory classification. Biometrics, 65(1), 159-168.

Ranucci, M., Bellucci, C., Conti, D., Cazzaniga A., & Maugeri, B. (2007). Determinants of early discharge from the intensive care unit after cardiac operations. Annals of Thoracic Surgery, 83, 1089-1095.

Ridge, J. C., Jones, S. K., Nielsen, M. S., & Shahani, A. K. (1998). Capacity planning for intensive care units. European Journal of Operational Research, 105 (2), 346-355.

Robertazzi, T. G. (1990). Computer networks and systems: Queueing theory and performance evaluation. Spring-Verlag.

Rosenberg, A. L., & Watts, C. (2000). Patients readmitted to ICUs: A systematic review of risk factors and outcomes. Chest, 118(2), 492-502.

Smith, C., Cowan, C., Sensenig, A., & Catlin, A. (2005). Health spending growth slows in 2003. Health Affairs, 24 (1), 185-194.

Takacs, L. (1963). A single-server queue with feedback. The Bell System Technical Journal, 42, 505-519.

Takahashi, Y., Miyahara, H., & Hasegawa, T. (1980). An approximation method for open restricted queueing networks. Operations Research, 28(3), 594-602.

Tan, S. (2005). Neighbor-weighted K-nearest neighbor for imbalanced text corpus. Expert Systems with Applications, 28(4), 667-671.

Terwiesch, C., KC, D., & Khan, J. M. (2011). Working with capacity limitations: Operations management in critical care. Critical Care, 15(4), 308-314.

Thongkam, J., Xu, G., Zhang, Y., & Huang, F. (2009). Toward breast cancer survivability prediction models through improving training space. Expert Systems with Applications, 36(10), 12200-12209

US Department of Health and Human Services (HHS). Fiscal year 2010 budget in brief. Retrieved June 25, 2011 from http://dhhs.gov/asfr/ob/docbudget/2010budgetinbriefl.html

Vanberkel, P. T., Boucherie, R. J., Hans, E. W., Hurink, J. L., & Litvak, N. (2010). A survey of healthcare models that encompass multiple departments. International Journal of Health Management and Information, 1(1), 37-69.

Van Hulse, J., & Khoshgoftaar, T. M. (2009). Knowledge discovery from imbalanced and noisy data. Data & Knowledge Engineering, 68(12), 1513-1542.

Van Hulse, J., Khoshgoftaar, T. M., & Napolitano, A. (2007). Experimental perspective on learning from imbalanced data. In Proceedings of the 24^{th} International Conference on Machine Learning, Corvalis, Oregon, 935-942.

Wei, J. M., Yuan, X. J., Qing, H. H., & Wang, S. Q. (2010). A novel measure for evaluating classifiers. Expert Systems with Applications 37(5): 3799-3809.

Weiss, E. N., & McClain, J. O. (1987). Administrative days in acute care facilities: A queueing-analytic approach. Operations Research, 35(1), 35-44.

Yen, S., & Lee, Y. (2009). Cluster-based under-sampling approaches for imbalanced data distributions. Expert Systems with Applications, 36(3), 5718-5727.